EDITORES:

Manuel A. Serrano - Eduardo Fernández-Medina

Cristina Alcaraz - Noemí de Castro - Guillermo Calvo

Actas de las VI Jornadas Nacionales
(JNIC2021 LIVE)

INVESTIGACIÓN EN ciber segu ridad

Ediciones de la Universidad
de Castilla-La Mancha

Investigación en Ciberseguridad

Actas de las VI Jornadas Nacionales (JNIC2021 LIVE)

Online 9-10 de junio de 2021
Universidad de Castilla-La Mancha

Investigación en Ciberseguridad

Actas de las VI Jornadas Nacionales (JNIC2021 LIVE)

Online 9-10 de junio de 2021
Universidad de Castilla-La Mancha

Editores:
Manuel A. Serrano,
Eduardo Fernández-Medina,
Cristina Alcaraz
Noemí de Castro
Guillermo Calvo

Ediciones de la Universidad
de Castilla-La Mancha

Cuenca, 2021




Edita: Ediciones de la Universidad de Castilla-La Mancha

Colección JORNADAS Y CONGRESOS n.º 34

Esta editorial es miembro de la UNE, lo que garantiza la difusión y comercialización de sus publicaciones a nivel nacional e internacional.

I.S.B.N.: 978-84-9044-463-4
D.O.I.: http://doi.org/10.18239/jornadas_2021.34.00

Hecho en España (U.E.) – *Made in Spain (E.U.)*

Bienvenida del Comité Organizador

Tras la parada provocada por la pandemia en 2020, las VI Jornadas Nacionales de Investigación en Ciberseguridad (JNIC) vuelven el 9 y 10 de Junio del 2021 con energías renovadas, y por primera vez en su historia, en un formato 100% online. Esta edición de las JNIC es organizada por los grupos GSyA y Alarcos de la Universidad de Castilla-La Mancha en Ciudad Real, y con la activa colaboración del comité ejecutivo, de los presidentes de los distintos comités de programa y del Instituto Nacional de Ciberseguridad (INCIBE). Continúa de este modo la senda de consolidación de unas jornadas que se celebraron por primera vez en León en 2015 y le siguieron Granada, Madrid, San Sebastián y Cáceres, consecutivamente hasta 2019, y que, en condiciones normales se habrían celebrado en Ciudad Real en 2020.

Estas jornadas se han convertido en un foro de encuentro de los actores más relevantes en el ámbito de la ciberseguridad en España. En ellas, no sólo se presentan algunos de los trabajos científicos punteros en las diversas áreas de ciberseguridad, sino que se presta especial atención a la formación e innovación educativa en materia de ciberseguridad, y también a la conexión con la industria, a través de propuestas de transferencia de tecnología. Tanto es así que, este año se presentan en el Programa de Transferencia algunas modificaciones sobre su funcionamiento y desarrollo que han sido diseñadas con la intención de mejorarlo y hacerlo más valioso para toda la comunidad investigadora en ciberseguridad.

Además de lo anterior, en las JNIC estarán presentes excepcionales ponentes (Soledad Antelada, del Lawrence Berkeley National Laboratory, Ramsés Gallego, de Micro Focus y Mónica Mateos, del Mando Conjunto de Ciberdefensa) mediante tres charlas invitadas y se desarrollarán dos mesas redondas. Éstas contarán con la participación de las organizaciones más relevantes en el panorama industrial, social y de emprendimiento en relación con la ciberseguridad, analizando y debatiendo el papel que está tomando la ciberseguridad en distintos ámbitos relevantes.

En esta edición de JNIC se han establecido tres modalidades de contribuciones de investigación, los clásicos artículos largos de investigación original, los artículos cortos con investigación en un estado más preliminar, y resúmenes extendidos de publicaciones muy relevantes y de alto impacto en materia de ciberseguridad publicados entre los años 2019 y 2021. En el caso de contribuciones de formación e innovación educativa, y también de transferencias se han considerado solamente artículos largos. Se han recibido para su valoración un total de 86

contribuciones organizadas en 26, 27 y 33 artículos largos, cortos y resúmenes ya publicados, de los que los respectivos comités de programa han aceptado 21, 19 y 27, respectivamente. En total se ha contado con una ratio de aceptación del 77%. Estas cifras indican una participación en las jornadas que continúa creciendo, y una madurez del sector español de la ciberseguridad que ya cuenta con un volumen importante de publicaciones de alto impacto.

El formato online de esta edición de las jornadas nos ha motivado a organizar las jornadas de modo más compacto, distinguiendo por primera vez entre actividades plenarias (charlas invitadas, mesas redondas, sesión de formación e innovación educativa, sesión de transferencia de tecnología, junto a inauguración y clausura) y sesiones paralelas de presentación de artículos científicos. En concreto, se han organizado 10 sesiones de presentación de artículos científicos en dos líneas paralelas, sobre las siguientes temáticas: detección de intrusos y gestión de anomalías (I y II), ciberataques e inteligencia de amenazas, análisis forense y cibercrimen, ciberseguridad industrial, inteligencia artificial y ciberseguridad, gobierno y riesgo, tecnologías emergentes y entrenamiento, criptografía, y finalmente privacidad.

En esta edición de las jornadas se han organizado dos números especiales de revistas con elevado factor de impacto para que los artículos científicos mejor valorados por el comité de programa científico puedan enviar versiones extendidas de dichos artículos. Adicionalmente, se han otorgado premios al mejor artículo en cada una de las categorías. En el marco de las JNIC también hemos contado con la participación de la Red de Excelencia Nacional de Investigación en Ciberseguridad (RENIC), impulsando la ciberseguridad a través de la entrega de los premios al *Mejor Trabajo Fin de Máster en Ciberseguridad* y a la *Mejor Tesis Doctoral en Ciberseguridad*. También se ha querido acercar a los jóvenes talentos en ciberseguridad a las JNIC, a través de un CTF (Capture The Flag) organizado por la Universidad de Extremadura y patrocinado por Viewnext.

Desde el equipo que hemos organizado las JNIC2021 queremos agradecer a todas aquellas personas y entidades que han hecho posible su celebración, comenzando por los autores de los distintos trabajos enviados y los asistentes a las jornadas, los tres ponentes invitados, las personas y organizaciones que han participado en las dos mesas redondas, los integrantes de los distintos comités de programa por sus interesantes comentarios en los procesos de revisión y por su colaboración durante las fases de discusión y debate interno, los presidentes de las sesiones, la Universidad de Extremadura por organizar el CTF y la empresa Viewnext por patrocinarlo, los técnicos del área TIC de la UCLM por el apoyo con la plataforma de comunicación, los voluntarios de la UCLM y al resto de organizaciones y entidades patrocinadoras, entre las que se encuentra la Escuela Superior de Informática, el Departamento de Tecnologías y Sistemas de Información y el Instituto de Tecnologías y Sistemas de Información, todos ellos de la Universidad de Castilla-La Mancha, la red RENIC, las cátedras (Telefónica e Indra) y aulas (Avanttic y Alpinia) de la Escuela Superior de Informática, la empresa Cojali, y muy especialmente por su apoyo y contribución al propio INCIBE.

Manuel A. Serrano, Eduardo Fernández-Medina
Presidentes del Comité Organizador
Cristina Alcaraz
Presidenta del Comité de Programa Científico
Noemí de Castro
Presidenta del Comité de Programa de Formación e Innovación Educativa
Guillermo Calvo Flores
Presidente del Comité de Transferencia Tecnológica

Índice General

Comité Ejecutivo

Juan Díez González	INCIBE
Luis Javier García Villalba	Universidad de Complutense de Madrid
Eduardo Fernández-Medina Patón	Universidad de Castilla-La Mancha
Guillermo Suárez-Tangil	IMDEA Networks Institute
Andrés Caro Lindo	Universidad de Extremadura
Pedro García Teodoro	Universidad de Granada. Representante de red RENIC
Noemí de Castro García	Universidad de León
Rafael María Estepa Alonso	Universidad de Sevilla
Pedro Peris López	Universidad Carlos III de Madrid

Comité Organizador

Presidentes del Comité Organizador

Eduardo Fernández-Medina Patón	Universidad de Castilla-la Mancha
Manuel Ángel Serrano Martín	Universidad de Castilla-la Mancha

Finanzas

David García Rosado	Universidad de Castilla-la Mancha
Luis Enrique Sánchez Crespo	Universidad de Castilla-la Mancha

Actas

Antonio Santos-Olmo Parra	Universidad de Castilla-la Mancha

Difusión

Julio Moreno García-Nieto	Universidad de Castilla-la Mancha
José Antonio Cruz Lemus	Universidad de Castilla-la Mancha
María A Moraga de la Rubia	Universidad de Castilla-la Mancha

Webmaster

Aurelio José Horneros Cano	Universidad de Castilla-la Mancha

Logística y Organización

Ignacio García-Rodriguez de Guzmán	Universidad de Castilla-la Mancha
Ismael Caballero Muñoz-Reja	Universidad de Castilla-la Mancha
Gregoria Romero Grande	Universidad de Castilla-la Mancha
Natalia Sanchez Pinilla	Universidad de Castilla-la Mancha

Comité de Programa Científico

Presidenta

Cristina Alcaraz Tello	Universidad de Málaga

Miembros

Aitana Alonso Nogueira	INCIBE
Marcos Arjona Fernández	ElevenPaths
Ana Ayerbe Fernández-Cuesta	Tecnalia
Marta Beltrán Pardo	Universidad Rey Juan Carlos
Carlos Blanco Bueno	Universidad de Cantabria
Jorge Blasco Alís	Royal Holloway, University of London
Pino Caballero-Gil	Universidad de La Laguna
Andrés Caro Lindo	Universidad de Extremadura
Jordi Castellà Roca	Universitat Rovira i Virgili
José M. de Fuentes García-Romero de Tejada	Universidad Carlos III de Madrid
Jesús Esteban Díaz Verdejo	Universidad de Granada
Josep Lluis Ferrer Gomila	Universitat de les Illes Balears
Dario Fiore	IMDEA Software Institute
David García Rosado	Universidad de Castilla-La Mancha
Pedro García Teodoro	Universidad de Granada
Luis Javier García Villalba	Universidad Complutense de Madrid
Iñaki Garitano Garitano	Mondragon Unibertsitatea
Félix Gómez Mármol	Universidad de Murcia
Lorena González Manzano	Universidad Carlos III de Madrid
María Isabel González Vasco	Universidad Rey Juan Carlos I
Julio César Hernández Castro	University of Kent
Luis Hernández Encinas	CSIC
Jorge López Hernández-Ardieta	Banco Santander
Javier López Muñoz	Universidad de Málaga
Rafael Martínez Gasca	Universidad de Sevilla
Gregorio Martínez Pérez	Universidad de Murcia

David Megías Jiménez	Universitat Oberta de Cataluña
Luis Panizo Alonso	Universidad de León
Fernando Pérez González	Universidad de Vigo
Aljosa Pasic	ATOS
Ricardo J. Rodríguez	Universidad de Zaragoza
Fernando Román Muñoz	Universidad Complutense de Madrid
Luis Enrique Sánchez Crespo	Universidad de Castilla-La Mancha
José Soler	Technical University of Denmark-DTU
Miguel Soriano Ibáñez	Universidad Politécnica de Cataluña
Victor A. Villagrá González	Universidad Politécnica de Madrid
Urko Zurutuza Ortega	Mondragon Unibertsitatea
Lilian Adkinson Orellana	Gradiant
Juan Hernández Serrano	Universitat Politécnica de Cataluña

Comité de Programa de Formación e Innovación Educativa

Presidenta

Noemí De Castro García	Universidad de León

Miembros

Adriana Suárez Corona	Universidad de León
Raquel Poy Castro	Universidad de León
José Carlos Sancho Núñez	Universidad de Extremadura
Isaac Agudo Ruiz	Universidad de Málaga
Ana Isabel González-Tablas Ferreres	Universidad Carlos III de Madrid
Xavier Larriva	Universidad Politécnica de Madrid
Ana Lucila Sandoval Orozco	Universidad Complutense de Madrid
Lorena González Manzano	Universidad Carlos III de Madrid
María Isabel González Vasco	Universidad Rey Juan Carlos
David García Rosado	Universidad de Castilla - La Mancha
Sara García Bécares	INCIBE

Comité de Transferencia Tecnológica

Presidente

Guillermo Calvo Flores	INCIBE

Miembros

José Luis González Sánchez	COMPUTAEX
Marcos Arjona Fernández	ElevenPaths
Victor Villagrá González	Universidad Politécnica de Madrid
Luis Enrique Sánchez Crespo	Universidad de Castilla – La Mancha

COMUNICACIONES

Sesión de Investigación A1:
Detección de intrusiones y gestión de anomalías I

http://doi.org/10.18239/jornadas_2021.34.01

Distributed Architecture for Intrusion Detection in IoT Networks using Smart Contracts

Rafael Z. A. da Mata*, Francisco L. de Caldas Filho†, Lucas M. C. e Martins‡,
Fábio L. L. de Mendonça§, and Rafael T. de Sousa Jr.¶
Cyber Security INCT Unit 6, Electrical Engineering Department, University of Brasilia, Brasilia, Brazil
e-mails: {rafael.zerbini} @uiot.org, {francisco.lopes, lucas.martins, fabio.mendonca}@redes.unb.br, {desousa} @unb.br
ORCID: * 0000-0002-5246-8858, † 0000-0001-5419-2712, ‡ 0000-0001-6436-7408,
§ 0000-0001-7100-7304, and ¶ 0000-0003-1101-3029

***Abstract*—DDoS attacks against distributed networks of IoT devices and applications are increasing. In this scenario, countermeasures must also use distributed security means. This paper proposes a distributed architecture of an intrusion detection system for IoT networks using smart contracts in blockchain. The proposed architecture supports interaction mechanisms between the local controllers using smart contracts to distribute security rules among the different devices on the IoT network. To validate the proposal, this paper presents the assessment of parameters required to ensure an acceptable level of reliability in sharing security rules.**

***Index Terms*—IoT security, intrusion detection, smart contracts, blockchain**

Tipo de contribución: *Investigación original*

I. Introduction

The number of Internet of Things (IoT) devices has grown exponentially over the years. According to a survey presented by [1], in 2015 there were 4.9 billion connected IoT devices and the forecast is that the number of connected devices in 2020 would be 25 billion, an increase of over 500% in just five years. To facilitate the configuration of these devices for end-users, manufacturers often use standard access passwords to configure these elements, thus making them vulnerable to unauthorized access by third-parties [2] thus creating the need for security mechanisms that work on resource-constrained devices [3].

This loophole allows these devices to be used for distributed denial of service (DDoS) attacks as described in [4], turning them part of botnets networks.

There are several security mechanisms to identify whether IoT devices are part of a botnet and to carry out malicious traffic mitigation actions [5], [6], [7]. The mitigation of malicious traffic can occur directly on the device, which identifies abnormal situations in its operation and takes actions to perform corrections, or on the network layer by using firewalls and Intrusion Prevention Systems (IPS) capable of recognizing abnormal situations and blocking traffic. Mitigation directly on the device involves the application of Host-Based Intrusion Detection System (HIDS), where an application running on the device identifies abnormal situations in its processes, on network ports, or in system files and acts to correct the fault.

This method stops the attack at the source, avoiding overloading the network infrastructure. However, it has its limitations:

- IoT devices have low computational power, hampering the execution of verification programs;
- The number of IoT devices grows exponentially making it hard and impracticable to apply security rules to each of them.

The model proposed by the previous work develops a HIDS in which the fault checking rules are configured on a controller in the cloud. The final IoT devices carry out periodic verification to check the need to include new rules in their own table and to take actions to solve problems included by the administrator in these rules. The system was adapted to be lightweight and run on devices with low computational power, being successfully tested on single-board computing (SBCs), such as Raspberry Pi.

However, this model has a dependence on the controller installed in cloud computing resources to be able to serve a large number of devices. Local networks or environments with a large volume of data should always check for new rules on the Internet, making the use of link more expensive and not allowing rules to be updated when there is no Internet connection available. The rules should also be entered into the cloud controller.

The model proposed in this work aims to solve some deficiencies presented in the previous work by creating a defense system associated with local controllers, in which IoT devices search for updated information directly from these local ones. The update of the rules that will be replicated to the devices does not occur only in the device-cloud model, but point-to-point, allowing that rules created and successfully validated in other environments by their local controller can be replicated, creating an always up-to-date federated environment.

To ensure that communication between the controllers occurs safely and only between controllers that are part of the requested federation, an environment was developed in this work where controllers exchange rules with each other using private smart contracts running in a public blockchain network, such as that described in [8].

This paper proposes a HIDS solution distributed for IoT networks. In order for organizations (local controllers of different networks) to share specific rules with other organizations, a solution based on smart contracts in a blockchain environment is proposed. The main advantages of this architecture are consensus on adding rules to the system, inviolability of the rules in the system, authentication in the

exchange of information, and complete audit.

In addition to this introduction, Section II presents the basis of the concepts discussed in this article and Section III point correlated works out and highlights their differences. Section IV describes the proposal of this paper, including its architecture and its description, while Section V presents and discusses the security analysis of shared rules and, finally, in Section VI, the conclusion and possible future work are presented.

II. Theoretical Foundation

Host-Based Intrusion Detection System, or HIDS, is a software traditionally used to perform monitoring routines aimed at detecting intrusion in storage systems [7]. Monitoring, in traditional HIDS's such as those studied by [9], focuses mainly on metric analyzes such as writing to disk and memory, repeated access to certain ports, entry, and exit of processes, in addition to others considered anomalous, as identified by [10]. In the context of IoT networks, there is a need for HIDS's that maintain the analysis and detection profile, but that also keep reduced computational cost, avoiding an overload in the system as shown by [11].

The existence of a single point of failure in IoT architectures is a negative aspect widely discussed, as highlighted by [12] and [13]. As an alternative to this problem, several studies have already made efforts to decentralize the distribution of information, including concerning security rules and for this reason have suggested the use of blockchain with its blocks joined by encryption [14], where the later block has the cryptographic hash of the previous block. The main role of each block is to store records of various transactions, such as changing security rules or a cryptocurrency transfer.

This approach using blockchain is considered advantageous, as seen in [15], because it offers mechanisms that solve problems such as the single point of failure, given the use of distributed architecture typical of blockchain architectures [16], [17]. Another relevant point, also highlighted by [14], is the consensus that is defined by the author as the judgment that each node belonging to the blockchain has the same block sequence. The majority vote defines the veracity of the information contained in the block under analysis.

Regarding the stored information in blockchain, like the aforementioned security rules, one of the alternatives is smart contracts that is defined by [18] as software with deterministic behavior executed in distributed networks whose main function is to mediate between two interested nodes in the exchange of information contained in blockchain. The use of smart contracts has gained much recognition due to its ability to transfer assets in the form, mainly, of cryptocurrencies. Another relevant definition is given by [19] which conceptualizes smart contracts as programs that execute correctly without the intervention of a trusted authority and cites Ethereum as responsible for implementing the smart model contracts currently adopted and also the consensus model applied by it, which is very robust due to it's verification scheme that only allows registration in the blocks for transactions considered valid [20].

III. Related Works

Given the expansion of IoT networks and the continuing need to provide security for these networks and their systems, several solutions using HIDS and blockchain have already been proposed. The review by [21] addresses the main challenges involved in intrusion detection, in particular:

- The problem of data sharing due to distrust between involved parties and also the possibility of reducing the amount of information shared by the same parties due to concerns about privacy, thus reducing the training optimization of detection algorithms;
- The use of a central server that provides trust management, becoming increasingly complex as the organization grows, thus increasing the possibility of internal attacks where nodes belonging to the organization authorize network access for external attackers.

Aiming at solving problems presented by [21], the project developed in this article makes use of blockchain to intensify mutual trust between the parties involved in sharing information and also to favor auditing and inviolability of them, in addition to being implemented in a distributed network to provide more robust trust management, from the point of view of the controllers, since it does not have a single central server as a single point of failure. In the scenario presented in the article, the referred information refers to the exchange of security rules between controllers and nodes.

In the context of traditional IoT infrastructures, [13] presents the distrust between system nodes and the existence of a single point of failure as the two main obstacles to continuous security management. The authors suggest, as a solution, the use of a distributed architecture using blockchain to store user interactions with the IoT environment so that these interactions are persisted in the form of blocks representing the history of the node or user and for this an encrypted token is used to guarantee the legitimacy of the user in question. Unlike the [13] proposal, which uses blockchain to store node history, our proposal uses smart contracts to propagate and store rules on the blockchain network. In addiction to this, our proposal was designed in the form of a distributed network whose validation of the information disseminated by the controllers is done through consensus mechanisms.

SVELTE is an intrusion detection system for IoT networks based on 6LoWPAN protocol proposed in [22], by using a hybrid of signature detection and anomaly detection based approach to perform intrusion detection. However, the system is designed to detect mostly routing attacks such as sinkhole attack, and the selective-forwarding attacks. The work of Surendar and Umamakeswari [23] also proposes a system for the 6LoWPAN protocol but using fewer resources than SVELTE, with a lower packet dropping ration and a more significant number of nodes into the evaluation. Our proposed solution is a novel detection with constrains and improved the shortcoming from SVELTE and INTI.

The work presented in [24] designed a HIDS, wich combines the advantages of Signature Intrusion Detection System (SIDS) and Anomaly-based Intrusion Detection System (AIDS). They also show that SIDS have a low false-

positive rate and show that the hybrid combination of both techniques presents a good accuracy compared to each one of their own. They were also preoccupied with the high energy consumption from the AIDS running all the time, so they used a mechanism only to activate the AIDS when a new attack's signature is expected to occur, which is verified by a SIDS. Their solution is a significant cost benefit between accuracy and energy consumption. We propose a different solution, mainly because our IDS runs in the IoT device.

Khraisat et al. work presented in [25] also proposed a hybrid IDS using signature and anomaly detection, combining a C5 classifier and One-Class Support Vector Machine classifier. Their work aimed to detect known vulnerabilities and zero-days with a high detection accuracy and a low false-alarm rate at the same time. Their model was tested against the Bot-IoT dataset, which includes legitimate IoT network traffic and several types of attacks, their results show a higher detection rate and a lower false-positive rate when compared to SIDS or AIDS techniques individually. We propose only the signature approach and our proposal is executed on the IoT device.

Diro and Chilamkurti's work described in [26] states that many zero-day attacks keep surging due of the many new protocols focused on IoT and most of these attacks are variants of well-known malwares, indicating that even machine learning models don't have a good response against these new attack's variants. Based on that, they proposed a Deep Learning based approach, in which the model would be resilient to small mutation or new attacks since deep learning introduces self-taught and compression capabilities. They showed that their approach was more effective in attack detection. Their approach also was created to work in a distributed environment, and their experiments showed that it was superior than the centralized detection system.

The paper [27] reports the growth of botnet networks related to the increasing number of IoT devices that have security breaches and are included in botnets such as the cited Mirai. However, this work does not address security measures to mitigate such attacks.

In [7], authors show the high frequency of DDoS attacks on IoT networks and cites the need to implement countermeasures that are distributed on the network to offer more effective coordination of IoT nodes. Their presented solution is a HIDS whose main function is to protect the backbone of the IoT network. The proposed HIDS is designed with conventional functionalities such as the authentication system using user and password, search for signatures of known attacks, verification of resource allocation, active processes, open service port and active connections, and as a differential offers the possibility to interact with your HIDS Controller to manage intrusion detection actions distributed over the network in DDoS attack situations. In our proposal presented in this paper, there is more than one provider, different from the scenario proposed by [7]. The main advantages of having more than one rules provider is the speed at which they are broadcast over networks, thus offering a faster response to security incidents such as DDoS and zero day cases.

In [5], a proposal is made for a subscription HIDS system to detect attacks and vulnerabilities in devices that want to connect to IoT networks through HIDS agents for IoT devices that perform the tests determined by the rules. The test result is compared to an expected result and, if it is outside of these parameters, the safety rule is executed and the HIDS controller is notified. The HIDS controller is responsible for managing HIDS agents in the distribution and updating of rules, analysis of alerts, treatment of threats and definition of premises. The scenario presented by [5] has a HIDS controller responsible for accessing the remote database to propagate the rules on the IoT network, different from our proposal that aims to implement a distributed network in which there is no dependency on a single node for spreading the rules to decrease the response time, especially in attack situations such as the aforementioned DDoS.

Finally, the authors suggest the use of blockchain as a way to solve a few challenges such as improving the security and privacy of information, increasing the reliability of the service through reputation systems, negotiating transactions, and dispute management. Our proposal also has a mechanism of reputation system similar to that of the proposal due to the distributed architecture of the rules which also contributes to the aspects of reliability with regard to the degree of confidence of each rule that is disseminated by the network.

IV. Proposed Architecture

The solution proposed in this work is composed of IoT devices that are capable of running the HIDS instance developed in the previous work and remote controllers, in which the security signatures that the final devices need to obey are registered. The final controllers can exchange rules with each other, allowing a federated system in which anomalous behaviors identified in one controller are passed on to the others, allowing the end devices to react to security breaches in a co-ordinated and decentralized manner according to the rules that are propagated by your parent company.

Periodically, the end devices consult remote controllers to find out if there are new rules to be saved in their database. This behavior is similar to anti-virus [28] systems. The possibility of rules that will be registered at the controller and replicated to the final devices are numerous, from the analysis of improper queries to suspicious DNS addresses as in [29], to the verification of any suspicious process as defined in previous work.

The process of exchange rules between controllers is done through a smart contract hosted on the Ethereum [20] blockchain. The rules in the smart contract follow the structure defined below.

- *name*: rule name;
- *id*: unique rule identifier;
- *createdAt*: creation date;
- *premisse*: value considered correct for a given condition or characteristic of the system;
- *testCase*: evaluation code to verify characteristics or conditions on the devices, and returns the result for verification;
- *action*: if the result of the case test is different from the value defined in the premise, the defined action is executed;

- *approvals*: nodes that marked rules as approved, nodes can revoke their votes later, and its added to a subfield called *revoke*;
- *denies*: number of nodes that marked rules as denied, the node also need to provide a reason for denying the rule;

A rule is a block of information that helps the HIDS agent to take predefined actions. This information block has the fields that are shown in Fig. 2 and are described below:

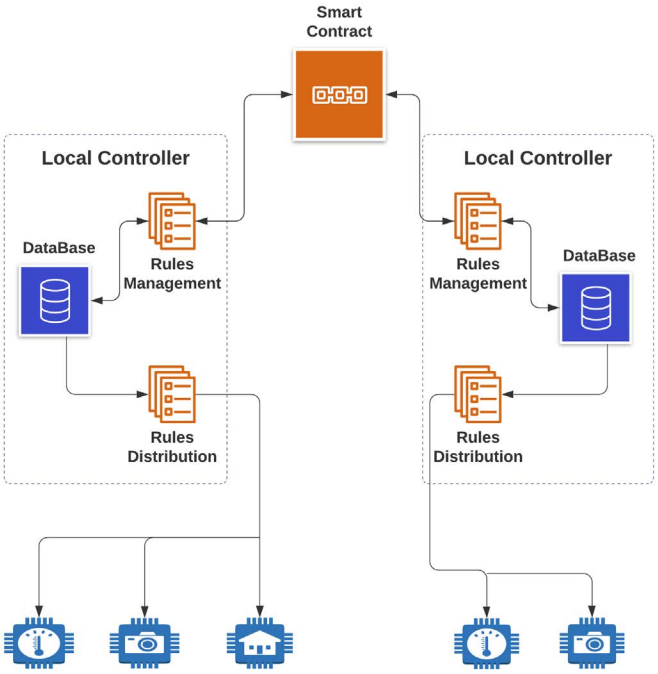


Fig. 1: Distributed Architecture


```
{
  "id": 1,
  "name": "os_name",
  "created_at": "2020-08-10T12:00:00",
  "premise": "Linux",
  "action": "print(\"System is not supported.\")\nraise System exit",
  "test_case": "import platform\noutput = platform.system()",
  "approvalls": [
    {
      "address": "0xc4aEb20798368C48B27280847E187BB332b9bC77"
    },
    {
      "address": "0x5A0b54d5dc17e0AADC383d2DB43b0A0d3e029c4C",
      "revoke": "Wrong premisse output"
    }
  ] ,
  "denies": [
    {
      "address": "0xd4aFb20798742C48B27280847E187BB332b9bC77",
      "reason": "Wrong premisse output"
    }
  ]
}
```


Fig. 2: Example of security rule scheme.

As soon as each node accesses a rule in the smart contract, it has the option to mark the rule as approved or denied, based on its security measures. In situations where more than half of the nodes reject a specific rule, it is marked as invalid and can no longer be accessed by other nodes. As long as the previous condition is not met, nodes can access the rule. Each node is responsible for choosing the number of approvals necessary for the rule to be considered valid for the same.

This architecture brings decentralization in the process of sharing rules between devices on different networks, it also has the advantage of having a form of consensus, in which a node can validate a rule based on other nodes validation. Another advantage is scalability since the largest amount of us offers more rules to distribute and a greater guarantee that rules are safe.

Also, the solution offers flexibility in the implementation of the rule control for each network, since the rules only need to follow a defined pattern when they will be shared for the entire network. Thus, a node that accesses a specific rule on the local network can modify it to meet its needs.

A. HIDS Agent

Each IoT device has a HIDS agent that is responsible for receiving the rules from the controller, saving them to its local database, and executing the rule. After rule execution, the device generates a report that is sent to the controller. In this report, the device informs the results of the case test and, if an action has been performed, it also reports whether it was performed successfully. Reports generated by the agent allow the administrator to keep updated on the status of the network, investigating possible vulnerabilities.

B. Smart Contract

A smart contract is an automatic execution contract whose terms of the contract between different parties are created via code. The agreements contained therein exist in a decentralized and distributed blockchain network. The code controls execution and transactions are traceable and irreversible.

Smart contracts allow reliable transactions and agreements to be carried out between anonymous and disparate parties, without the need for a central authority, legal system, or external enforcement mechanism. The smart contract that stores the rules were developed using the Ethereum blockchain.

The implemented smart contract has three main functions accessed by the nodes:

- *registerApproval*: this function is responsible for registering the approval of each node, the function stores the node address;
- *registerDeny*: this function is responsible for registering the rule deny from each node, the function store the node address and information from the node about the reason for the denial;
- *revogueApproval*: this function is responsible for revoking a previous approval from a node, the node must inform the reason to call a revoke from his vote and all nodes that approved the rule are notified about the revoke, and can also call the function under the same pretense. The revoked votes are marked as revoked in the approval list and added to the deny list.

The smart contract is published by the creator of the rule that is going to be distributed. At any moment any node can consult the numbers of approvals or denies from a rule and

also see the reason for revokes, in case the node wishes to fix the issues itself.

C. Rules Management

The rules that will be executed by HIDS are defined in their entirety at the parent company. The controller can operate within a hierarchical cloud environment, in which the rules registered with the controller are copied to the final devices, with one controller serving all of the final IoT devices. This model has advantages such as simplicity of implementation, however, it presents points of failure that deserve attention, they are:

- The unavailability of communication with the cloud server makes it impossible to send new rules;
- There is a single cloud controller which if is compromised can affect all end devices;
- Depending on the distance between cloud servers and end devices, latency can compromise the exchange of rules and delay reactions to security locks on end devices.

The model proposed in this work, as can be seen in Fig. 1, uses a smart contract system responsible for distributing security rules to controllers in different networks who are responsible for managing, persisting in their own database, and distribute these same security rules to the local devices with which you have a direct connection acting as the default gateway. In this way, there is a considerable reduction in latency, and each end device can operate without a direct connection to the Internet.

The management of the rules is carried out by the network administrator so that he has total autonomy over creating, deleting, updating, and reading the rules. This service is also responsible for informing which rules should be shared with other controllers via smart contract and which have autonomy only locally. In the rules management service, the administrator also checks the rules received from other controllers, using a smart contract, and decides whether they should be applied to their rule base. This service is also responsible for activating the function in the smart contract, validating or not the rule provided by the network.

D. Rules Distribution

With each update or creation of rules by the administrator, the rule distribution service is triggered and aims to distribute the new rule to the devices, according to the definition made by the administrator of which devices should receive the rule.

Confidentiality in the exchange of information is guaranteed by the use of encryption, thus preventing systems that analyze the contents of TCP/IP packets or metadata from obtaining success in obtaining information, as demonstrated in [30]. For this, the exchange of messages between the controller and end devices is performed using SSL so that, being initiated by a TLS handshake [31], the rules are provided to the end device when the session is granted to the device.

V. Security Analysis of Shared Rules

This section assesses the security of the rules shared in the smart contract, considering the research question on how many approvals are needed on a rule to ensure that a rule is secure.

To understand the research question, it is interesting to consider the following scenario: a controller sends a dishonest rule, using the smart contract, and if other controllers cooperate with the dishonest node, the rule can generate vulnerabilities on the devices that receive the rule. That is why it is important to establish a minimum security threshold on the received rule to ensure the security of the networks.

To validate the consensus of the rules, a statistical study was carried out to assess the probability that the number of approvals is greater than the number of denials, given a certain number of confirmations imposed by a node.

For this purpose, tests were carried out following a binomial model to simulate nodes approving or denying transactions. The binomial models used have two probabilities, p, that a node approves a rule, which are $p = 0.6$ and $p = 0.8$, and two probabilities of rejecting a rule, q, $q = 0.4$ and $q = 0.2$ respectively. For the tests, different node values were used, which are called N, and the values used were $N = 10$, $N = 50$ and $N = 100$. The minimum values chosen for approval, k, were $N/8$, $N/4$ and $N/2$. The distribution for honest nodes follows Eq. 1 and for dishonest nodes follows Eq. 2.

$$P(A) = \sum P(\{(e_1, ..., e_N)\}) = \binom{N}{k} \cdot p^k (1-p)^{N-k} \quad (1)$$

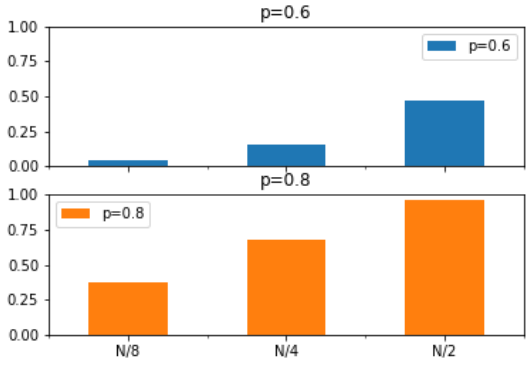
$$P(D) = \sum P(\{(e_1, ..., e_N)\}) = \binom{N}{k} \cdot q^k (1-q)^{N-k} \quad (2)$$

Analyzing Fig. 3, it can be seen that for $p = 0.6$, the probability of the rule being valid is negligible, and may be discarded, which can also be seen in Table I. As for $p = 0.8$, confirmations above $N/4$ prove to be sufficiently reliable, with a probability greater than fifty percent in all cases, which can be seen in Table II. It is interesting to note that with $p = 0.8$ the difference between the probabilities for $N/2$ and $N/4$ are very small, making $N/4$ an acceptable choice for ensuring network security.

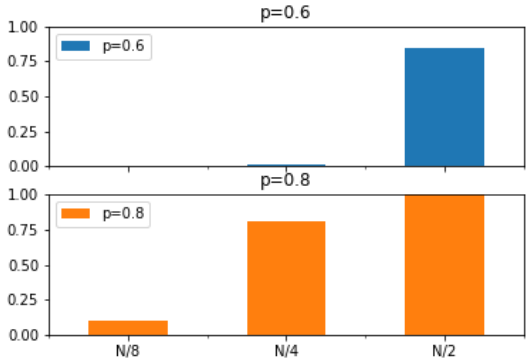
TABLE I: Results for $p = 0.6$

Nº of nodes (N)	Min. accepted (k)	Probab. of $A > k$
10	10/8	0.044
10	10/4	0.154
10	10/2	0.466
50	50/8	0
50	50/4	0.013
50	50/2	0.844
100	100/8	0
100	100/4	0.001
100	100/2	0.956

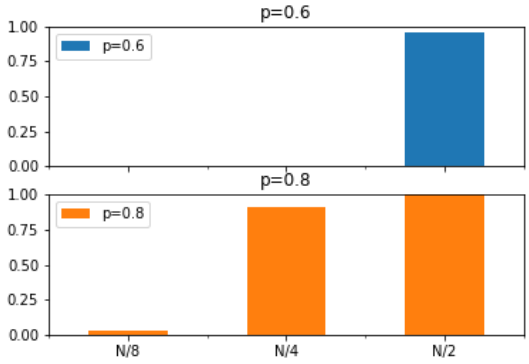
From the analysis presented, it is possible to conclude that $N/2$ is the ideal value to validate the rule. It is also possible to observe that for $p = 0.6$ having few nodes ($N = 10$), the values for confirmation are still no more than 0.5, which shows that to guarantee security in the network with few nodes it is necessary to wait for a little more than half of the nodes to approve the rule. For $p = 0.8$, high-reliability values start from $N/4$, regardless of the number of nodes, which can be advantageous if the node has a high degree of confidence in the other nodes in the network, which makes it interesting to increase the speed of updating the rules on your devices.



(a) $N = 10$



(b) $N = 50$



(c) $N = 100$

Fig. 3: Comparison of results of approval and rejection of rules when using 10, 50, and 100 nodes

TABLE II: Results for $p = 0.8$

Nº of nodes (N)	Min. accepted (k)	Probab. of $A > k$
10	10/8	0.375
10	10/4	0.677
10	10/2	0.960
50	50/8	0.103
50	50/4	0.813
50	50/2	0.999
100	100/8	0.025
100	10/4	0.912
100	10/2	0.999

VI. Conclusion and Future Works

The proposed solution is an evolution of a previous work in which a distributed architecture for intrusion detection in IoT networks using smart contracts in blockchain was developed. A statistical study was carried out to assess the minimum consensus required between nodes to ensure that security checks are legitimate. This study concluded that it is possible to accept the legitimacy of security checks, with a probability greater than fifty percent, waiting for at least half of the nodes to ensure their legitimacy. The study also demonstrated scenarios where the nodes have a high degree of trust with each other, in these scenarios a quarter of confirmations is sufficient to guarantee the legitimacy of security checks.

As future work we have:

- analysis of more complex scenarios with different minimum values of consensus and different probabilities of acceptance for transactions in the Binomial model;
- development of analytical models that consider network and time parameters, for example, using the *mainnet* of the Ethereum network and using a Poisson-based statistical model to assess the minimum time to reach consensus;
- competitive analysis between different smart contract platforms described by [32] to achieve greater speed and/or lower cost between transactions.

Acknowledgment

This work was supported in part by CNPq - Brazilian National Research Council, Grant 312180/2019-5 PQ-2, Grant BRICS 2017-591 LargEWiN, and Grant 465741/2014-2 INCT in Cybersecurity, in part by CAPES - Brazilian Higher Education Personnel Improvement Coordination, Grant 23038.007604/2014-69 FORTE, and Grant 88887.144009/2017-00 PROBRAL, in part by the Brazilian Ministry of the Economy, Grant 005/2016 DIPLA, and Grant 083/2016 ENAP, in part by the Institutional Security Office of the Presidency of Brazil, Grant ABIN 002/2017, in part by the Administrative Council for Economic Defense, Grant CADE 08700.000047/2019-14, and in part by the General Attorney of the Union, Grant AGU 697.935/2019.

References

[1] Gartner, Inc., *Gartner Says 6.4 Billion Connected "Things" Will Be in Use in 2016, Up 30 Percent From 2015*, Nov 2015. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2015-11-10-gartner-says-6-billion-connected-things-will-be-in-use-in-2016-up-3

[2] C. Kolias, G. Kambourakis, A. Stavrou, and J. Voas, "DDoS in the IoT: Mirai and Other Botnets," *Computer*, vol. 50, no. 7, pp. 80–84, 2017.

[3] H. G. C. Ferreira and R. T. de Sousa Junior, "Security analysis of a proposed internet of things middleware," *Cluster Computing*, vol. 20, no. 1, pp. 651–660, 2017.

[4] A. A. Y. R. Fares, F. L. de Caldas Filho, W. F. Giozza, E. D. Canedo, F. L. L. de Mendonça, and G. D. A. Nze, "DoS Attack Prevention on IPS SDN Networks," in *2019 Workshop on Communication Networks and Power Systems (WCNPS)*. IEEE, 2019, pp. 1–7.

[5] B. V. Dutra, J. F. Alencastro, F. L. Caldas Filho, L. M. C. Martins, R. T. de Sousa Jr., and R. O. Albuquerque, "HIDS by signature for embedded devices in IoT networks," in *Actas de las V Jornadas Nacionales de Investigación en Ciberseguridad (JNIC 2019)*. Cáceres, Spain: Universidad de Extremadura, Jun 2019, pp. 53–61.

[6] D. G. V. Gonçalves, G. d. O. Kfouri, B. V. Dutra, J. F. d. Alencastro, F. L. d. Caldas Filho, L. M. C. e. Martins, R. d. O. Albuquerque, and R. T. de Sousa Jr, "IPS architecture for IoT networks overlaid on SDN [Arquitetura de IPS para redes IoT sobrepostas em SDN]," in *XIX Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, São Paulo-SP, 2019.

[7] G. O. Kfouri, D. G. V. Gonçalves, B. V. Dutra, J. F. d. Alencastro, F. L. Caldas Filho, L. M. C. Martins, B. J. G. Praciano, R. d. O. Albuquerque, and R. T. de Sousa Jr, "Design of a Distributed HIDS for IoT Backbone Components," in *FedCSIS (Communication Papers)*, Leipzig, Germany, 2019, pp. 81–88.

[8] R. Yuan, Y.-B. Xia, H.-B. Chen, B.-Y. Zang, and J. Xie, "ShadowEth: Private Smart Contract on Public Blockchain," in *J. Comput. Sci. Technol.*, vol. 33, 2018, pp. 542–556.

[9] F. Sabahi and A. Movaghar, "Intrusion detection: A survey," in *2008 Third International Conference on Systems and Networks Communications*, 2008, pp. 23–26.

[10] D. Wagner and P. Soto, "Mimicry attacks on host-based intrusion detection systems," in *Proceedings of the 9th ACM Conference on Computer and Communications Security*, 2002, pp. 255–264.

[11] M. F. Elrawy, A. I. Awad, and H. Hamed, "Intrusion detection systems for IoT-based smart environments: a survey," *Journal of Cloud Computing*, vol. 7, no. 21, Dec 2018.

[12] S. Xie, Z. Zheng, W. Chen, J. Wu, H.-N. Dai, and M. Imran, "Blockchain for Cloud Exchange: A Survey," *Computers & Electrical Engineering*, vol. 81, p. 106526, Jan 2020.

[13] R. Agrawal, P. Verma, R. Sonanis, U. Goel, A. De, S. A. Kondaveeti, and S. Shekhar, "Continuous Security in IoT Using Blockchain," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 6423–6427.

[14] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, "An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends," *2017 IEEE International Congress on Big Data (BigData Congress)*, pp. 557–564, 2017.

[15] H.-N. Dai, Z. Zheng, and Y. Zhang, "Blockchain for Internet of Things: A Survey," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8076–8094, 2019.

[16] D. Carboni, "Feedback based reputation on top of the bitcoin blockchain," *ArXiv*, vol. abs/1502.01504, 2015.

[17] R. A. Bruce, R. T. de Sousa Júnior, F. L. L. de Mendonça, J. P. Pimentel, M. T. de Holanda, and F. L. de Caldas Filho, "Blockchain for Interbank Operations [Blockchain para Operações Interbancárias]," in *Atas das Conferências Ibero-Americanas WWW/Internet 2019 e Computação Aplicada 2019*, Lisbon, Portugal, 2019.

[18] C. Laneve, C. S. Coen, and A. Veschetti, *On the Prediction of Smart Contracts' Behaviours*. Cham, Switzerland: Springer International Publishing, 2019, pp. 397–415.

[19] M. Bartoletti, "Smart Contracts Contracts," *Frontiers in Blockchain*, vol. 3, p. 27, Jun 2020.

[20] V. Buterin, "A Next-Generation Smart Contract and Decentralized Application Platform," Ethereum Foundation, Tech. Rep., 2015. [Online]. Available: https://ethereum.org/en/whitepaper/

[21] W. Meng, E. W. Tischhauser, Q. Wang, Y. Wang, and J. Han, "When Intrusion Detection Meets Blockchain Technology: A Review," *IEEE Access*, vol. 6, pp. 10 179–10 188, 2018.

[22] S. Raza, L. Wallgren, and T. Voigt, "SVELTE: Real-time intrusion detection in the Internet of Things," *Ad Hoc Networks*, vol. 11, no. 8, pp. 2661–2674, May 2013.

[23] S. Madhawa, P. Balakrishnan, and U. Arumugam, "Data driven intrusion detection system for software defined networking enabled industrial internet of things," *Journal of Intelligent & Fuzzy Systems*, vol. 34, pp. 1289–1300, 2018, 3. [Online]. Available: https://doi.org/10.3233/JIFS-169425

[24] H. Sedjelmaci, S. M. Senouci, and M. Al-Bahri, "A lightweight anomaly detection technique for low-resource IoT devices: A game-theoretic methodology," in *2016 IEEE International Conference on Communications (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[25] A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman, and A. Alazab, "A Novel Ensemble of Hybrid Intrusion Detection System for Detecting Internet of Things Attacks," *Electronics*, vol. 8, no. 11, Oct 2019.

[26] A. A. Diro and N. Chilamkurti, "Distributed attack detection scheme using deep learning approach for Internet of Things," *Future Generation Computer Systems*, vol. 82, pp. 761–768, Aug 2018.

[27] C. Kolias, G. Kambourakis, A. Stavrou, and J. Voas, "DDoS in the IoT: Mirai and other botnets," *Computer*, vol. 50, no. 7, pp. 80–84, 2017.

[28] B. B. Rad, M. Masrom, and S. Ibrahim, "Evolution of computer virus concealment and anti-virus techniques: A short survey," *CoRR*, vol. abs/1104.1070, 2011. [Online]. Available: http://arxiv.org/abs/1104.1070

[29] T. L. Sperling, F. L. Caldas Filho, R. T. de Sousa Jr., L. M. C. Martins, and R. L. Rocha, "Tracking intruders in IoT networks by means of DNS traffic analysis," in *2017 Workshop on Communication Networks and Power Systems (WCNPS)*. Brasília-DF: IEEE, 2017, pp. 1–4.

[30] T. L. Sperling, B. A. França, F. L. Caldas Filho, L. M. C. Martins, R. O. Albuquerque, and R. T. de Sousa Jr., "Evaluation of an IoT device designed for transparent traffic analysis," in *2018 Workshop on Communication Networks and Power Systems (WCNPS)*. Brasília-DF: IEEE, 2018, pp. 1–5.

[31] K. Bhargavan, C. Fournet, M. Kohlweiss, A. Pironti, P.-Y. Strub, and S. Zanella-Béguelin, "Proving the TLS Handshake Secure (As It Is)," in *Advances in Cryptology – CRYPTO 2014*, J. A. Garay and R. Gennaro, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 235–255.

[32] Z. Allam, "On Smart Contracts and Organisational Performance: A Review of Smart Contracts through the Blockchain Technology," *Review of Economic and Business Studies*, vol. 11, no. 2, pp. 137–156, 2019.

Sesión de Investigación A1: Detección de intrusiones y gestión de anomalías I
http://doi.org/10.18239/jornadas_2021.34.02

A review of SSH botnet detection in initial stages of infection: a Machine Learning-based approach

José Tomás Martínez Garre, Manuel Gil Pérez, and Antonio Ruiz-Martínez
Faculty of Computer Science, University of Murcia, 30100 Murcia, Spain
Email: josetomas.martinez@um.es, mgilperez@um.es, arm@um.es

***Abstract*—Botnets are exponentially increasing because of new zero-day attacks, a variation of their behavior, and obfuscation techniques that are not detected by traditional defense systems. Botnet detection has been focused on intermediate phases of the botnet's life cycle during operation, underestimating the initial phase of infection. Using SSH-based High Interaction Honeypots, we have designed a Machine Learning-based system capable of detecting the botnet infection phase in near real time, which was trained with a real dataset of executed commands and the network data obtained during SSH sessions. This approach reached a very high level of prediction and zero false negatives, where all known and unknown SSH sessions aimed at infecting our honeypots were detected.**

***Index Terms*—Botnet, Machine Learning, Zero-day malware, Honeypot, High interaction**

Tipo de contribución: *Investigación ya publicada*

I. Introduction

A botnet is a network composed of hosts (*bots*) infected with malware, which are under human control through Command and Control (C&C) servers, whose life cycle is made up of the infection, communication, and attack phases. In the infection one, the host is infected with malware actively or passively (e.g., vulnerable SSH service or email phishing with a malicious document), through different attack vectors, which are used for downloading the malware binaries that will turn the victim into a potential bot. Secondly, communications take place between the infected computer and the C&C servers in the communication phase, to become a new member of the botnet and to update its behavior. Lastly, in the attack phase, the bots are the ones that carry out the malicious activity (e.g., sensible information theft or DDoS attacks).

The exponential rise in zero-day malware, or its variants that modify their behavior anytime, and the use of advanced obfuscation techniques help the malware to evade the traditional signature- and heuristics-based detection techniques. A practical method for detecting new malware is to use fake and monitored vulnerable systems (or honeypots), and analyze the generated log events to reveal intelligence information and malware behavior on current threats. Besides, the number of attacks is dramatically growing, and analyzing the massive number of log events generated by honeypots is a challenge for security analysts, despite the fact that bots exhibit certain behavioral patterns in their ways of proceeding.

Both facts lead to the current trend of developing bot detection methods based on Machine Learning (ML) techniques, which can recognize complex patterns and make qualified decisions based on experience to detect known and unknown malware with fewer false positives and false negatives, in comparison with other malware detection methods. Several ML solutions have been defined for botnet detection, but most of them are focused on identifying bots in the communication and attack phases, albeit bots may have caused certain damage. However, a detection during the infection phase would have the advantage that the bot can be identified and disabled before participating in any malicious activity.

To address this, we focused on the development of a novel approach that used ML to detect the infection phase through the SSH service performed by a known or unknown botnet. We focus on SSH because it is a very common botnet attack vector and, to the best of our knowledge, there is no solution focused on the detection of the infection process in SSH-based botnets. The present work was initially published in [1].

II. Proposal

We propose a ML-based solution using Supervised Learning algorithms to automate detection and prediction of incoming SSH security threats aiming to infect new hosts. Thus, our approach needs evaluated and tagged data to build the model and classify the SSH sessions' behavior, but to the best of our knowledge, there is no public dataset with executed commands and network traffic statistics generated during SSH sessions. We decided to use a high interaction SSH honeypot (HIH) to create a dataset for the effective and dynamic training of our ML model. Fig. 1 depicts the honeypot architecture used in our solution with a *Machine Learning-based Detection System*, in addition to entities such as the *Attacker* and the honeypots unfolding the *SSH Sensors*.

The workflow of our solution starts when an attacker seeks to inject the malware through (default SSH) port 22, breaking into a device through a brute-force or dictionary attack, among others. Malicious SSH behavior conducted by the attackers is captured by our SSH sensors. We focus on the executed commands and the network traffic statistics generated in each SSH session. After the attack, the SSH sensor forwards the captured intelligence information and its self-internal state to the ML-based detection system, where the records received from all SSH sensors are stored in a database. Next, our system generates a new ML-based infection detection model according to the records stored in our database. Finally, the new model is sent to the SSH sensors so that they can update and improve their previous detection model. This process is repeated every day automatically.

Correct labeling of the dataset is essential for an accurate evaluation of results, which is achieved thanks to our honeypots. An SSH session will be tagged as positive when:

- A change in the honeypot status or an outgoing network connection occurs when the attacker logs out.

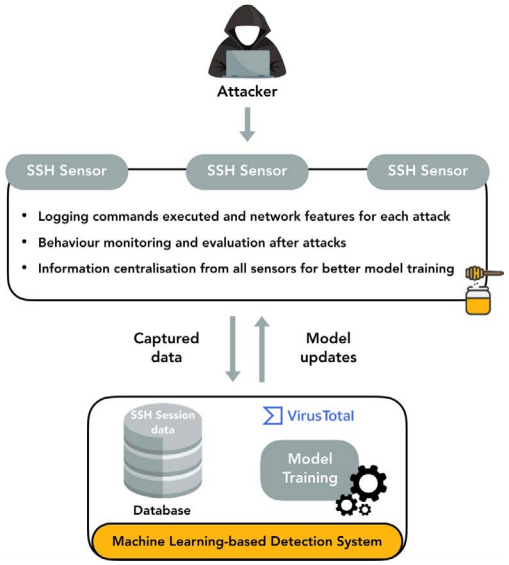


Fig. 1. The ML-based detection framework for the botnet's infection phase.

- The VirusTotal scanning indicates that the downloaded executable during the session corresponds to a malware.

After five months, we obtained a dataset with 2139 records: 337 positive (infected hosts) and 1802 negative SSH sessions (uninfected hosts). So, the dataset has a total of 93 features: 72 commands, 7 session states, and 14 network statistics. Where *commands* denote a list of different commands executed in all attacks received; *session states* refer to the number of executed commands (namely, Check software configuration, Download a file, Install a program, Run a program, Change the account password, Check the hardware configuration, and Change the system configuration); and *network statistics* are features like session duration, bytes sent, bytes received, etc.

III. Experimental Results

Table I depicts the results obtained by the most widely used ML algorithms in earlier research works. Although Random Forest (RF) and SVM achieve similar results, we chose the RF classifier due to a lower number of false negatives and, mainly, because RF is computationally less expensive.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Decision Tree	89.4	100	43.1	60.3
Random Forest	**98.1**	**95.7**	**93.9**	**94.8**
SVM	97.7	96.7	90.8	93.7
Naive Bayes	69.9	38.2	98.5	55.0

TABLE I
Performance of the classification algorithms proposed.

In the evaluation process, the classification accuracy and other error metrics were used to show the effectiveness of our model, tested with a testing dataset consisting of 427 examples (20% of the dataset): 54 positive SSH sessions (infected hosts) and 373 negative SSH sessions (uninfected hosts).

Our test results are presented in Table II. Not all uninfected SSH sessions were detected, but all malicious SSH sessions were successfully identified (false negatives = 0). Not having false negatives is very critical in an attack detection system.

A thorough analysis of the results shows that network statistics better profile the attacker's behavior than the commands executed by attackers, as shown in Figure 2. Thus, it confirms our hypothesis that commands executed by attackers can give a lot of information about the purpose of the SSH sessions. Still, since several commands are doing the same operation, further context information (e.g., network statistics) is needed to predict attacks accurately. Unfortunately, session states do not seem to represent the infection phase correctly via SSH service. In addition, the most relevant command is *chmod*, widely used to give execution permission to a file to execute it. *Scp* and *wget* allow the transfer and download a file into the victim. Other significant commands are *echo*, *rm*, and *./* (run an executable file). These commands correspond to Download, Installation, and Execution states, a sequence widely used in this attack type. These results do not mean that attackers cannot infect a host executing other commands, but it is the most frequently seen pattern in the received attacks.

Class	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Infected	99.59	96.87	100	98.41
Uninfected	99.59	100	99.53	99.76

TABLE II
Model performance for our testing dataset.

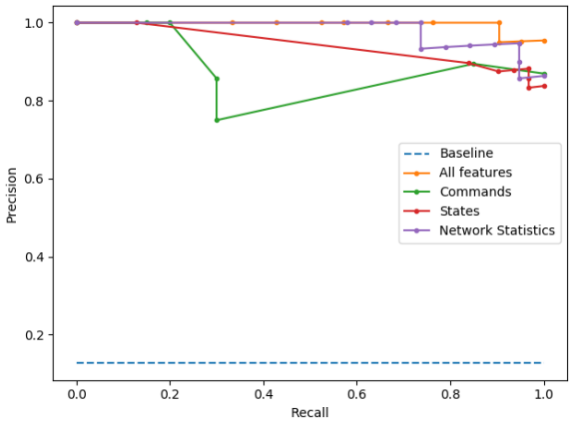


Fig. 2. Precision-Recall curve for classification.

As a summary, very positive results were achieved, but no real comparison can be made with other works since existing datasets with network traffic reflect phases other than that of infection and the commands executed by attackers only fit with our scenario. Instead, we have evaluated our approach with a dataset created by our honeypots for an SSH attack environment. Thus, we can confirm that ML techniques are feasible for our detection system.

Acknowledgements

The work is supported by the Spanish Ministry of Science, Innovation and Universities, FEDER funds, under grant no. RTI2018-095855-B-I00 and TIN2017-86885-R, the European Commission Horizon 2020 Programme under grant agreement no. H2020-SU-DS-2019/883335 – PALANTIR (*Practical Autonomous Cyberhealth for resilient SMEs & Microenterprises*), and the European Commission (FEDER/ERDF).

References

[1] J. T. Martínez Garre, M. Gil Pérez, A. Ruiz Martínez: "A novel Machine Learning-based approach for the detection of SSH botnet infection," *Future Generation Computer Systems*, vol. 115, pp. 387–396, 2021.

http://doi.org/10.18239/jornadas_2021.34.03

A Multi-agent Approach for Online Twitter Bot Detection

Jefferson Viana Fonseca Abreu
Computer Science Departament
University of Brasília
Brasília, DF, Brazil
Email: jeffvfa@hotmail.com
ORCID: 0000-0001-6609-5338

Célia Ghedini Ralha
Computer Science Departament
University of Brasília
Brasília, DF, Brazil
Email: ghedini@unb.br
ORCID: 0000-0002-2983-2180

João José Costa Gondim
Computer Science Departament
University of Brasília
Brasília, DF, Brazil
Email: gondim@unb.br
ORCID: 0000-0002-5873-7502

***Abstract*—Online social networks are tools that allow interaction between human beings with a large number of users. Platforms like Twitter present the problem of social bots which are controlled by automated agents potentially used for malicious activities. Thus, social bot detection is important to keep people safe from harmful effects. In this work, we approach the Twitter bot online detection problem with a multi-agent system (MAS). It is based on supervised classification with three machine learning algorithms and a reduced set of features. The MAS performance compared to the three algorithms applied separately - Random Forest, Support Vector Machine, and Naïve Bayes - presented similar results. Besides, interesting results for online bot detection with the MAS prototype suggested that 88.19% of bots detected were correctly labeled. The results indicate that the approach used is feasible and promising for the real-time bot detection problem.**

***Keywords*—Bot detection, Social bots, Twitter, Agents, Multi-Agent System, MAS.**

1. Introduction

Online social networks (OSN) are tools designed to facilitate interaction between human beings. Due to a large number of OSN users, it is natural that some individuals might be willing to carry out malicious actions seeking to gain undue advantage, the weakest link in information security [1], [2]. Social bots (hereinafter referred to only as bots), OSN profiles that impersonate human beings through automated interactions [3], are one of the methods used to carry out these abuses. These bots can be grouped into botnets acting in a coordinated manner, enabling more powerful and harmful attacks [4].

Bots can be used for malicious tasks such as: sharing spam [5], [6], [7], [8], vectors for phishing [3], [9] or spread of news [10]. This issue is very current in the Brazilian scenario after the 2018 elections, where there was a great prominence in electoral campaigns linked to the internet, where OSN provided a favorable scenario for the use of bots. The suspicion of the use of this device during the electoral period, and after it, culminated in the establishment of a Joint Parliamentary Commission of Inquiry (CPMI) at the Brazilian national congress to investigate, among other issues, the use of bots to influence the results of the 2018 elections [11].

Among the most well-known OSN, it is possible to highlight Twitter. The relevance of this social media is so great, that several notable people and heads of state adopt this platform as one of their main media. In the literature, several works have as one of their objectives the classification of bots on Twitter. In [12], bot classification approaches permeating three different classes are identified: using machine learning (ML) algorithms; based on graphs and emergent approaches; and one which encompasses techniques that do not fit into the other two categories. All the works listed in [12] were studied along with more recent publications. In [13] there is the conception of a reduced set of features that allows the detection of bots with quality comparable to one of the works that are state of the art Twitter bot detection [14].

Despite the thorough review of the literature, no work was identified involving the detection of bots using an approach based on intelligent agents (hereinafter treated only by agents). This fact represents an interesting research opportunity because the use of this approach can be very useful in this field of application. Agents are entities capable of performing activities in an automated, persistent manner, and having the ability to adapt to different contexts [15]. These are desirable attributes when it is envisaged to enhance the detection of bots for Twitter.

Thus, the main objective of this work is to develop a multi-agent system (MAS) capable of autonomously detect bots on Twitter. The MAS design phase uses Tropos methodology. The MAS applies a supervised classification approach with three ML algorithms using and a reduced set of features (presented previously in [13]). The contributions of this work includes:

- a bot classifier using a MAS approach that combines three ML algorithms presenting good performance with the average of AUC equal to 0.9856 and standard deviation of 0.0199; and
- proof of concept with the MAS capturing and clas-

sifying users provided by an online stream of tweets where 88.19% of bots detected were correctly labeled.

This paper is organized as follows: Section 2 presents the preliminary concepts that are needed to fully understand our work, Section 3 contains the works that were found in the literature and are most related to this research, Section 4 explains the methodology followed, Section 5 shows the results achieved and discusses them, Section 6 are presented the final considerations and the future work.

2. Preliminaries

In this section, some concepts applied in the work are presented, including intelligent agents and MAS design.

2.1. Intelligent Agents

According to [16] computational intelligence is the area of study that deals with the development of intelligent agents. [15] defines an agent as an entity that acts, but a computational agent is expected to do more like: operate under autonomous control, perceive its environment, persist for an extended period, adapt to changes, and be able to create and achieve goals. As all computer programs are developed to process something, we can differentiate an agent by the ability to perceive its environment with its sensors and act on it using its actuators.

According to [17], [18], a MAS is composed of a set of agents capable of interacting with the environment and with each other, cooperatively or competitively, in pursuit of achieving one or more individual or collective objectives. The agents of a MAS can play several roles to achieve their objectives, either through the independent execution of actions, adapting to changes in the environment, and interacting with other agents. A MAS project needs the modeling phase to define the behavior and reasoning of each agent, the definition of a communication and interaction protocol for the group of agents, as well as which tools will be used in the implementation of the system.

According to [15], a rational agent needs to perform actions seeking to achieve the best possible result using the set of available information. Acting rationally is an action that may involve the treatment of uncertainties, logical inferences, and reflexes. Agents can be classified according to their performance in the environment through their behavior at various levels of complexity, from the simplest to the most complex including simple reactive agent, model-based reactive agent, objective based agent, utility based agent, and learning agent. All types of agents can be extended using ML techniques.

2.2. MAS Design

An essential element in the design of MAS is the correct characterization of the environment in which agents are inserted. According to [15], [17], [18] the environment must be characterized according to six essential aspects: observable, deterministic, episodic, static, discrete, and multi-agent.

For an adequate definition of the task environment of the rational agent, [15] defined PEAS (Performance measure, Environment, Actuators, and Sensors). In the MAS design pre-project phase, the aspects of performance measurement, environment, perceptions, and actions for each agent should be specified as detailed as possible. The performance measure is not fixed for all agent tasks, it is defined by the designer and must be able to assess the agent's behavior in a specific environment. For each sequence of perceptions of the rational agent, the designer must select an action that will maximize the defined performance measure, given the evidence provided by the sequence of perceptions and any internal knowledge of the agent or received from other agents.

In this work, we applied the modeling methodology called Tropos to design agent oriented software systems. According to [19], Tropos is based on two key ideas. First, the notion of agent and all related mentalistic notions (for instance goals and plans) are used in all phases of software development, from early analysis down to the actual implementation. Second, Tropos covers also the very early phases of requirements analysis, thus allowing for a deeper understanding of the environment where the software must operate, and of the kind of interactions that should occur between software and human agents. The development of a project using Tropos is divided into five phases: initial requirements, late requirements, architectural design, detailed design, and implementation. For the initial modeling phases, the language i* (iStar) is adopted, which has a focus on the intentional (why?), social (who?) and strategic (how?) dimensions of the software [20].

3. Related Work

Considering the Twiter bot detection works available in the literature, we present some recent ones that achieved interesting results comparable to our proposal. In Table 1 the related work comparison is presented.

In [14] the idea of classifiers with only one class trained with data from legitimate users is explored obtaining a tool capable of detecting any type of bot. To demonstrate this thesis several ML algorithms were investigated: the Bayesian networks, J48, RF, Adaboost, bagging, k-nearest neighbors. The authors identified the most used multiclass algorithms in the literature review performed by them include the logistic regression, multilayer perceptron, Naïve Bayes (NB), and Support Vector Machine (SVM). The tested class algorithms were bagging-TPMiner, bagging-random miner, one-class k-means with randomly-projected features algorithm, one-class SVM, and NB. The classifiers were trained and tested offline with the public datasets of [23], [24]. As a result, there is a consistent classification of bots of different types without needing any prior information, with an AUC of 0.89.

In [4] a classifier is developed to detect retweeting bots. The method consisted of taking all the retweets during a time

TABLE 1. RELATED WORK COMPARISON

Reference	ML algorithm	Application Domain	Method
Rodríguez-Ruiz et al. (2020) [14]	BN, J48, RF, Adaboost, Bagging, KNN, LR, MLP, NB, SVM, BTPM, BRM, OCKRA, ocSVM	general	offline supervised classification
Mazza et al. (2019) [4]	LSTM	retweet botnets	online unsupervised classification
Begenilmis & Uskudarli (2018) [21]	RF, LR, SVM	election campaign	offline supervised classification
Varol et al. (2017) [22]	RF	general	offline supervised classification, with online data

window to analyze the retweets pattern for each account during the period. A data visualization technique, called ReTweet-Tweet, was developed, which found four patterns of retweets, one for human users and three for bots. When conceiving the classifier Retweet-Buster (RTBust) the unsupervised algorithm Long Short-Term Memory (LSTM) was applied. With 12 features, a F1-score = 0.87 was obtained better than those in [22], [25], [26].

In [21] an organized behavior classifier is built in Twitter. The period chosen here is the 2016 American elections, as it is believed that there was a high volume of propaganda spread and fake news using OSN. Three different classifiers were trained and tested offline: RF, LR and SVM. The algorithms classified the accounts as organized or organic, political or non-political, pro Trump or pro Hillary. The RF algorithm performed better than the others with an average accuracy and F1-score greater than 95% in each category. The source code and the datasets used are available in a public repository at GitHub. The main contribution of this work is the design of a basic model for the classification of organized behavior on Twitter.

In [22] a report is made about the classifier *bot or not* (called today as botometer). The bots classification is done using Random Forest (RF), where more than a thousand different Twiter account attributes are used to perform the classification. A dataset containing only bots obtained through a honeypot [27] and a dataset manually annotated by the researchers are used. Both datasets are then mixed to build the dataset used in the article that is available to the community. The authors also maintain a web tool that can classify any profile on Twitter in real-time. The measure of the accuracy of implementation is given by the Area under the ROC Curve (AUC) with a score of 0.95. Authors realized that user metadata and tweet content are the most important attributes to find out if the account is a bot or not. Also, it is shown that accounts controlled by software can be grouped according to their intentions or *modus operandi*.

In a previous work [13], a set of five features is defined and four classifiers - RF, SVM, NB, and one class SVM (ocSVM) - were implemented to use this reduced set of features. The performance of the algorithms is compared to the state-of-the-art bot detection work presented in [14]. The classifiers were trained and tested offline with the public datasets of [23], [24]. The accuracy of the classifier was considered homogeneous with an average of 0.8549 and 0.1889 of standard deviation. Also, all multiclass classifiers achieved AUC greater than 0.9, indicating a practical benefit for bot detection on Twitter.

4. Methodology

As presented in Section 1, the objective of this work is to develop a MAS that performs the detection of bots for Twitter automatically. In the MAS design, we used the Tropos methodology. The choice of this methodology is motivated mainly by the fact that it covers several different phases of the MAS project, which allows a greater understanding of the system and a more refined set of requirements. In this section, the development stages performed during the work will be reported.

4.1. Design Model

In the pre-project phase of the MAS design, we define the PEAS (see Section 2.2). In this work, the environment that the MAS will interact is Twitter. It consists of a microblog where users post short messages (hereinafter referred to as tweets) in their respective profiles [28]. According to [29], the tweets are up to 280 text characters long and may contain references to other profiles by typing @ and the username to be mentioned. Other multimedia elements are also possible, such as videos, images, surveys, geolocation, among others [30]. Thus, the environment shared by agents in the MAS can be defined as:

- partially observable: impossible to view the complete set of all tweets;
- non-deterministic: it is not possible to determine the contents of the tweet in advance;
- episodic: tweets can be treated loosely connected in episodes;
- dynamic: new tweets may appear at any time;
- continuous: tweets may appear in a continuous spectrum of values; and
- multi-agent: more than one agent can perceive and act in the environment.

Three types of agents were defined in the MAS. The first type is the collector which is responsible for identifying the profiles that are making tweets according to a pre-established domain, saving the respective data in a monitoring database. As soon as data is identified, the collector forwards it to the classifying agents. In turn, the classifiers perform the classification of the profiles and send the results to the referee. The training of the classifiers is carried

TABLE 2. PEAS FOR THE THREE AGENTS: COLLECTOR, CLASSIFIER AND REFEREE.

Agent Type	Performance Measure	Actuators	Sensors
Colletor	Number of collected tweets	Monitor tweets according to key-words, Collect information about tweets, Record information in database, Send data to classifiers.	Verifies the existence of new tweets in the tweet flow chosen
Classifier	Number of executed classifications	Receive tweet data from collectors, Extract the necessary features, Classify the user who posted the tweet.	Verifies the existence of new data from tweets collected by collectors
Referee	Number of accounts classified as bots	Count *votes* and decide about the classification.	Verifies the existence of new classification from classifier agents

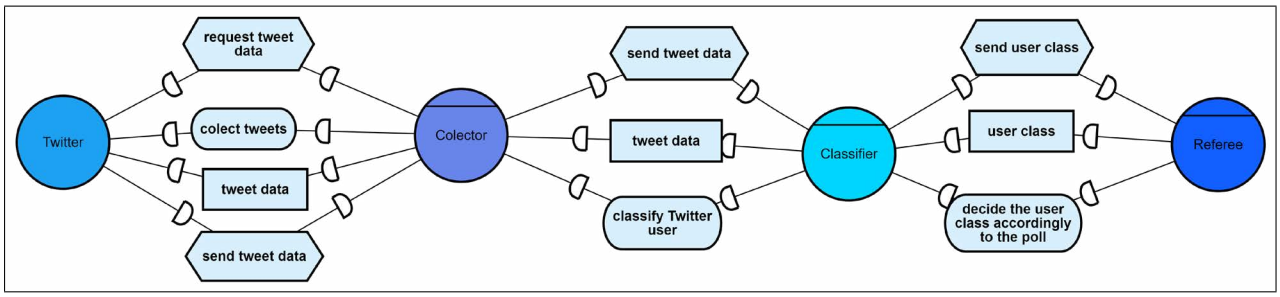


Figure 1. Late requirements diagram.

out offline, making its use faster during the execution of the system. Each classifying agent is implemented with a different ML algorithm, so they can classify the same profile into different classes. The choice of the approach using supervised ML algorithms was based on [12], where it was possible to glimpse the most widely used algorithms for the Twiter bot application domain.

The referee agent is responsible for making the final classification of the profile, according to the results of the classifying agents through a simple majority vote. In other words, the referee joins the decisions resulting in a group learning approach and records these decisions in a comma-separated values (csv) file. The experiment is only interrupted by the researchers' decision. The PEAS of the three agent types is available in Table 2. Since the environment is common to all agents, it is not included in the PEAS table.

During the MAS design project, four different models were built referring to development phases. The models designed include the initial requirements, late requirements, architectural design, and detailed design. Due to the adoption of the Tropos [19] methodology, the models were built using the language i* (iStar) [20]. To build the models, the online tool piStar implemented by [31] was used. Figure 1 illustrates the late requirements diagram for the MAS project, showing the relationships between actors, agents, tasks, resources, and objectives.

4.2. Agent Reasoning Model

Three types of intelligent agents were defined for the MAS. Each type has different properties and reasoning models. The reasoning model of each agent was chosen according to the actions that they would perform as presented in the PEAS definition of Table 2.

The first type of agent is the collector with the reactive reasoning model. This is the simplest type of reasoning without a sophisticated strategy but the choice of action that will be performed is based on the perception-action mapping. This reasoning model was chosen because the activity to be performed is trivial since the agent is responsible for accessing Twitter to collect tweets based on a set of pre-defined keywords and passing the tweet data to the classifying agents.

The classifier agent is defined as a learning agent. This agent is responsible for receiving data from the collector agent, applying a feature selection routine, and performing a data classification using a ML algorithm. This classification will be the agent's *vote* to assist the referee's decision. In the first MAS implementation, three classifying agents were built using one of the three most common algorithms in this application domain: RF, SVM, and NB. The same classifiers previously used in [13] intending to allow comparison with the MAS results.

Finally, the referee is an objective-oriented agent. Such agents consider, among the universe of actions to be taken, which is the best one to achieve their objectives. The referee is responsible for counting the classifier agents' *votes* referring to a Twitter profile. The referee's decision will allow reporting the profile to Twitter and monitor whether the profile has been banned.

4.3. Architecture

The MAS architecture is based on the reactive agent model that is horizontally layered as presented in [32]. The architecture layers connect directly the sensors (input) and the actuators (output). There are three layers and m possible actions suggested by each layer resulting in m^3 possible

interactions. The three layers are divided by well-defined activities and each has homogeneous agents. The activities include tweets capture, the tweets' authors classification, and arbitration. In other words, the architecture is horizontalized distributed into competency modules. Note that the architecture is naturally prepared to accommodate the online detection, since it establishes a direct pipeline for processing and classifying tweets as they are captured by the agents.

One of the facts that support the architecture choice is the high fault tolerance inherent in the model, failures that can occur at any of the levels. As the MAS aims to deal with a dynamic environment this property is interesting. Besides, the layered scheme is a good choice when we have several agents with different capabilities interacting in the same environment. Adversity faced in this architectural model is the bottleneck introduced by the communication between MAS and the environment since there are unique points of communication for both input and output. However, the positive points mentioned compensating for this negative characteristic for the time being. Also, the proposed architecture does not have many layers, which helps to soften the bottleneck. In Figure 2 we present a diagram of the proposed MAS architecture. It is possible to visualize the agent types and how they interact with each other.

4.4. Implementation

After the artifacts described, the implementation of the MAS was carried out. We used the language Python, version 3 and the framework for agent-oriented development called Python Agent DEvelopment framework (PADE) developed by [33]. The source code is open source and is available in a public repository of the research group at GitLab (https://gitlab.com/InfoKnow/SocialNetwork/jeffersonabreu-twitterbotdetection). The three proposed agents were developed and the interactions between them occur using communication protocols standardized by the Foundation for Intelligent Physical Agents (FIPA) [34].

The communication between the collector and the classifier agents occurs through the FIPA Agent Communication Language (ACL) with an interaction protocol called subscribe [35]. In this protocol, the agent's subscribers perform a subscription to an agent publisher, and after that, all subscribers receive notifications sent by the publisher. The role of subscribers while the collecting agent acts as a publisher. We believe this interaction protocol implements the desired behavior for communication between classifiers and the collector, through a simple message exchange using the FIPA inform performative.

For training and testing the MAS, the dataset used is provided by [23]. It is composed of several files in the csv format containing information related to user profiles and tweets. Each file is flagged with the type of profile produced by the present data. The types of profiles used in this work are the same as those of [13], including:

- Social spambots#1 (Social1) which includes accounts that retweeted the candidate for mayor of Rome in the 2014 Italian elections, where one of the candidates hired a marketing company that used 991 bots to run the campaign.
- Genuine accounts which the $3,474$ accounts were classified as legitimate using the following method: a simple question was asked in natural language for profiles chosen at random in the OSN, after this step the responses were analyzed by human beings to define whether the account is legitimate.
- Traditional spambots#1 (Traditional) is the same dataset provided by [24], which contains information on $1,000$ bots that have posted with malicious links.

The classification algorithms used are the same as those of [13], trained with 90% of the dataset Genuine

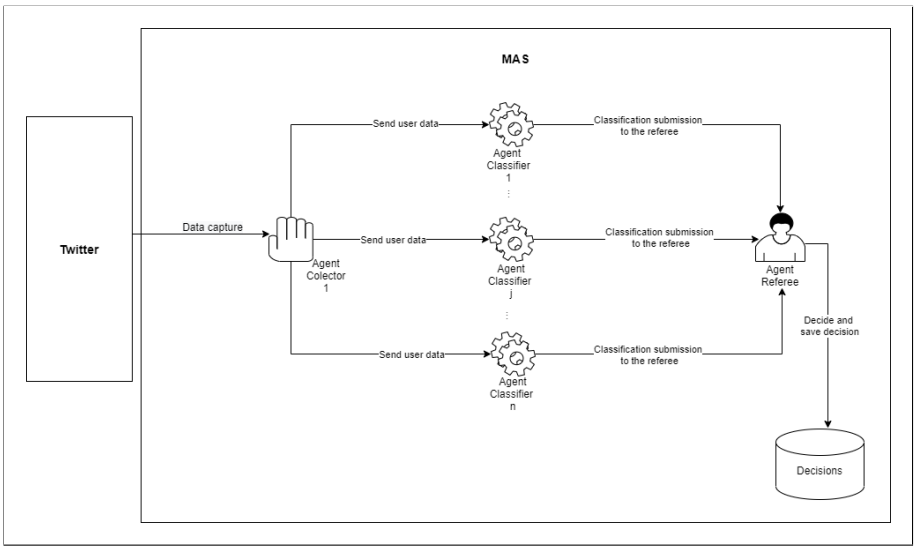


Figure 2. The MAS horizontally layered architecture.

accounts combined with 90% of social spambots#1 xor Traditional spambots #1, serialized with the help of the pickle [36], [37] library. The serialization is important to avoid training the three classifiers every time the MAS is executed. The features set adopted here is the same as [13], it includes the amount of tweets ($statuses\_count$), number of followers ($followers\_count$), number of friends ($friends\_count$), number of likes given by the account ($favorites\_count$), and number of lists that the account is included ($listed\_count$).

5. Experiments and Results

Two different tests were carried out during the development of this research. The first one intended to measure the performance of the MAS developed, using the same methodology as [13] and comparing the two works. The second experiment is a proof of concept involving the use of the MAS to perform the Twitter bot detection on an online stream of tweets. Both tests are better discussed in the following subsections.

5.1. MAS Performance

To validate the proposed solution, the MAS performance was compared with the three algorithms executed separately as previously presented in [13]. In Table 3 we present the comparison of the scores for the three algorithms - RF, SVM, and NB - using the Social1 and Traditional datasets. As in [13] the implementation of the ML algorithms are provided by [38]. Note that, the MAS presented the second-best performance among the other algorithms with an accuracy of 0.9795 in the dataset Social1 and 0.9840 in the Traditional dataset, only behind the RF results with 0.9821 and 0.9933 with Social1 and Traditional datasets, respectively.

According to [39], [40], the recall (also known as sensitivity) is the ratio between the correctly classified examples (true-positive - TP) and the incorrectly classified examples (false-negative - FN). In other words, is the proportion of true-positive cases that are correctly predicted. As we can see the recall score from the MAS is considerably high (it is only lower than the RF) in both datasets. This measure indicates that the MAS maintains the high sensitivity achieved by our best ML classifier.

According to [41], F1-score represents the harmonic average between precision and recall. The value of the F1-score is a number between zero and one, where values closer to one indicate high classification performance. Over again, our MAS keeps the good score achieved by our experiment performed at [13], which is a good indicator that the solution proposed applies to the task of classifying Twitter bots. The results related to the AUC score are better explained below.

In Table 4 there is a comparison between the AUC of the classifiers executed separately and combined in the MAS. For the calculus of the AUC, the probability of the MAS was measured considering the average of the probabilities of the algorithms. This procedure is the same used in [38] for the calculus of the probability for the Adaboost ensemble method. Since we have fewer examples for each type of bot than examples of legitimate accounts in the dataset, we faced a class imbalance problem. Therefore, choosing AUC as a performance measure is appropriate [14], [42]. The MAS achieved the second-best performance among the other algorithms with an average AUC of 0.9856 (standard deviation of 0.0199) only behind the RF results with 0.9878 (standard deviation of 0.0170). Figure 3 contains the results presented in Table 4, where for each classifier the first two bars represent the AUC values obtained using the two datasets (i.e., Social1 and Traditional) and the third bar contains the average value. The standard deviation is also shown in the first two bars.

TABLE 3. ACCURACY, AUC, RECALL AND E F1-SCORE FOR CLASSIFIERS (ALONE AND COMBINED IN MAS), ACCORDING TO TRAIN/TEST DATASET.

ML algorithm/ Dataset	Accuracy	AUC	Recall	F1-score
RF/Social1	**0,9821**	0,9758	0,9636	0,9737
RF/Traditional	**0,9933**	0,9999	0,9850	0,9902
SVM/Social1	0,8281	0,9382	0,6186	0,6421
SVM/Traditional	0,9821	0,9978	0,9778	0,9744
NB/Social1	0,5000	0,9011	0,6710	0,4980
NB/Traditional	0,8438	0,9937	0,8994	0,8145
MAS/Social1	**0,9795**	0,9716	0,9587	0,9682
MAS/Traditional	**0,9840**	0,9996	0,9776	0,9756

TABLE 4. COMPARISON BETWEEN AUC OBTAINED FOR MAS AND THE CLASSIFIERS ALONE AS IN [13].

Técnica de AM	Social1	Traditional	Average	Standard Deviation
RF	0,9758	0,9999	**0,9878**	**0,0170**
SVM	0,9382	0,9978	0,9680	0,0421
NB	0,9011	0,9937	0,9474	0,0654
MAS	0,9716	0,9996	**0,9856**	**0,0199**

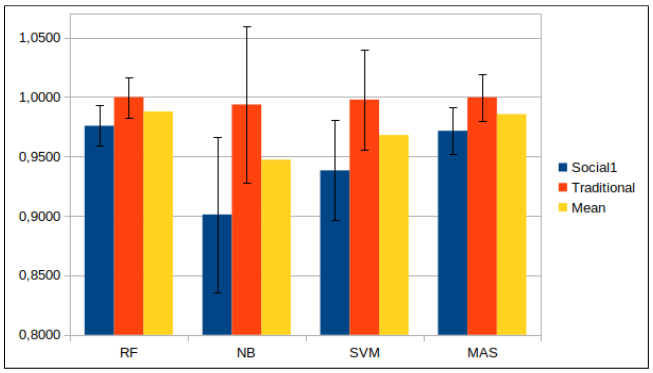


Figure 3. Parallel of the AUCs obteined for MAS and [13]

Considering the presented experiments, the MAS results are very close to the best algorithm that is RF meaning the approach can be used for classification with satisfactory performance. We might consider that the method on which the vote is taken is not complementary. Also, the RF algorithm presents better results considering the datasets

used, but we cannot guarantee it will perform better with any dataset. The ensemble methods in ML produce optimal predictive models, as pointed out in the literature and the MAS approach, can provide an ideal framework for the implementation of such methods. In this work, we provided the MAS proof of concept with three different ML algorithms and a simple decision process based on voting. However, other solutions can enhance the results of ensemble methods especially considering online tweet bot detection.

5.2. Online Twitter Bot Detection

An important attribute for a bot classifier is execution in near real time, thus becoming a detector. In [9] there is a demonstration that the detection of bots from Twitter allows very simple bots to continue working on the platform. Therefore, it is necessary to have tools that help the discovery of bots as close as possible to real time. A very slow tool can detect bots when they have already been excluded or have already caused damage. Thus, harmful bots can be discovered earlier by minimizing the damage caused by them. During the literature review it was noticeable that most articles do not perform the detection of bots in real time. In this way, a proof of concept (PoC) of the execution of the MAS was also carried out using real-time data collected from the Twitter API. The main objective of this PoC was to assess the MAS approach feaseability for real time detection.

The methodology was simple and used the MAS set up with the three ML classifiers (RF, SVM, and NB) trained with the dataset Social1 combined with the dataset genuine accounts (both provided by [23]), the same used in part of the first test. This dataset choice was done intending to discover the most sophisticated bots, as argued in [13], [23]. The stream of tweets was retrieved looking for the keyword "vacina" which is the Portuguese word that means vaccine. This keyword was chosen by the fact that the test was carried out during the COVID19 pandemic, and this theme was a hot topic in Brazil. It is notorious that it is more likely to capture bots in a larger sample of accounts, and the keyword is very helpful to allow for it. The test was executed using one notebook with the following specs: 7,8 GiB of RAM, CPU Intel® Core™ i7-3610QM, GPU NVIDIA® GeForce® GT 630M with 2GB DDR3 VRAM, Solid State Drive 250 GB, Operational System Xubuntu 20.04.2 LTS (64 bits).

The execution took place on February $12^{th}, 2021$, for approximately six hours (from 11:56 am to 6:00 pm). About 44,156 tweets were captured, corresponding to 31,271 different users, which were classified (representing a throughput of approximately 123 tweets per minute, roughly two per second). Among these, 183 tweets came from profiles classified as bots by MAS. These 183 suspicious tweets were published by 144 different Twitter accounts.

On March $25^{th}, 2021$, a request for data about these 144 profiles classified as bots was carried out. The Twitter API was able to return data about only 17 profiles, which means that 127 suspicious profiles were suspended by Twitter or excluded by their owners. It suggests that $88,19\%$ of the users classified as bots by our method were correctly labeled, using the same methodology as [27] to validate their results. This preliminary test, although quite simple, showed that the proposed methodology is on the right track and applies to real-time data.

6. Conclusion

The main objective of extending the work in [13] by developing a MAS capable of autonomously detecting Twitter bots was successfully achieved. The average AUC of 0.9856 and standard deviation equal to 0.0199 shows that MAS is useful for labeling bots on Twitter. The approach applyed in the MAS developement allows exploring ML classifiers by replacing, removing or adding algorithms.

The MAS architecture naturally accommodated online detection as it uses a direct pipeline for processing and classifying tweets captured by agents. When tested on a PoC with an online tweet stream it behaved well with an indication of good throughput and accuracy of detection. Nevertheless, more robust tests should be executed to evaluate the classification of data using online streams.

In future work, we intend to improve agents' reasoning capacity, especially the referee agent. This can be enhanced by making more rational decisions about the use of classification votes from the agents. Another idea would be to transform the referee into a coordinator agent that could perform a pre-classification by sending profiles with certain characteristics to specialized classifiers saving computational resources. Furthermore, we intend to carry out a more robust validation of online bot detection, since this is a very important feature for the adoption of the MAS as a real-time Twitter bot detector.

References

[1] B. Schneier, "Secrets & lies: Digital security in a networked world," *International Hydrographic Review*, vol. 2, no. 1, pp. 103–104, 2001.

[2] K. D. Mitnick and W. L. Simon, *The art of deception: Controlling the human element of security*. John Wiley & Sons, 2003.

[3] M. Shafahi, L. Kempers, and H. Afsarmanesh, "Phishing through social bots on twitter," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3703–3712.

[4] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi, "RTbust: Exploiting Temporal Patterns for Botnet Detection on Twitter," in *Proceedings of the 10th ACM Conference on Web Science*, ser. WebSci '19. Boston, Massachusetts, USA: Association for Computing Machinery, 2019, pp. 183–192. [Online]. Available: https://doi.org/10.1145/3292522.3326015

[5] A. H. Wang, "Detecting spam bots in online social networking sites: A machine learning approach," in *Data and Applications Security and Privacy XXIV*, S. Foresti and S. Jajodia, Eds. Berlin, Heidelberg: Springer, 2010, pp. 335–342.

[6] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proceedings of the 26th Annual Computer Security Applications Conference*, ser. ACSAC '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1–9. [Online]. Available: https://doi.org/10.1145/1920261.1920263

[7] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?" *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 6, pp. 811–824, 2012.

[8] G. Tavares and A. Faisal, "Scaling-laws of human broadcast communication enable distinction between human, corporate and robot twitter users," *PloS one*, vol. 8, no. 7, 2013.

[9] J. V. F. Abreu, J. H. C. Fernandes, J. J. C. Gondim, and C. G. Ralha, "Bot development for social engineering attacks on twitter," 2020.

[10] C. Freitas, F. Benevenuto, S. Ghosh, and A. Veloso, "Reverse engineering socialbot infiltration strategies in twitter," in *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2015, pp. 25–32.

[11] L. da Mata, "Plano de trabalho CPMI da Fake News," Congresso Nacional. Disponível em: http://legis.senado.leg.br/sdleg-getter/documento/download/d78bab7d-515f-45df-b785-03e364a7e138 Acessado em: 19/04/2020., 2019.

[12] M. Latah, "Detection of malicious social bots: A survey and a refined taxonomy," *Expert Systems with Applications*, vol. 151, p. 113383, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417420302074

[13] J. V. Fonseca Abreu, C. Ghedini Ralha, and J. J. Costa Gondim, "Twitter bot detection with reduced feature set," in *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2020, pp. 1–6.

[14] J. Rodríguez-Ruiz, J. I. Mata-Sánchez, R. Monroy, O. Loyola-González, and A. López-Cuevas, "A one-class classification approach for bot detection on twitter," *Computers & Security*, vol. 91, p. 101715, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167404820300031

[15] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. USA: Prentice Hall Press, 2009.

[16] D. Poole, A. Mackworth, and R. Goebel, *Computational Intelligence*. UK: Oxford university press, 1998.

[17] M. Wooldridge, *An introduction to multiagent systems*, 2nd ed. UK: John Wiley & Sons Ltd, 2009.

[18] G. Weiss, Ed., *Multiagent systems*, 2nd ed. MIT press, 2013.

[19] P. Bresciani, A. Perini, P. Giorgini, F. Giunchiglia, and J. Mylopoulos, "Tropos: An agent-oriented software development methodology," *Autonomous Agents and Multi-Agent Systems*, vol. 8, no. 3, pp. 203–236, 2004.

[20] F. Dalpiaz, X. Franch, and J. Horkoff, "istar 2.0 language guide," 2016.

[21] E. Beğenilmiş and S. Uskudarli, "Organized behavior classification of tweet sets using supervised learning methods," in *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, ser. WIMS '18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3227609.3227665

[22] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online Human-Bot Interactions: Detection, Estimation, and Characterization," in *11th International AAAI Conference on Web and Social Media (ICWSM)*, May 2017. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587

[23] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Proceedings of the 26th International Conference on World Wide Web Companion*, ser. WWW '17 Companion. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017, p. 963–972. [Online]. Available: https://doi.org/10.1145/3041021.3055135

[24] C. Yang, R. Harkreader, and G. Gu, "Empirical evaluation and new design for fighting evolving twitter spammers," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 8, pp. 1280–1293, 2013.

[25] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for sale: Efficient detection of fake twitter followers," *Decision Support Systems*, vol. 80, pp. 56–71, 2015.

[26] S. Liu, B. Hooi, and C. Faloutsos, "A contrast metric for fraud detection in rich graphs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2235–2248, 2019.

[27] K. Lee, B. Eoff, and J. Caverlee, "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter," in *International Conference on Weblogs and Social Media (ICWSM), AAAI Publications*, 2011.

[28] M. Fazil and M. Abulaish, "A hybrid approach for detecting automated spammers in twitter," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2707–2719, 2018.

[29] Twitter Inc., "Help center," https://help.twitter.com/en, 2020, accessed 09/17/2020.

[30] T. Gui, P. Liu, Q. Zhang, L. Zhu, M. Peng, Y. Zhou, and X. Huang, "Mention recommendation in twitter with cooperative multi-agent reinforcement learning," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 535–544.

[31] J. Pimentel and J. Castro, "pistar tool – a pluggable online tool for goal modeling," in *2018 IEEE 26th International Requirements Engineering Conference (RE)*, 2018, pp. 498–499.

[32] F. J. Mora Lizán and C. Rizo-Maestre, "Intelligent buildings: Foundation for intelligent physical agents," 2017-05.

[33] L. S. Melo, R. F. Sampaio, R. P. S. Leão, G. C. Barroso, and J. R. Bezerra, "Python-based multi-agent platform for application on power grids," *International Transactions on Electrical Energy Systems*, vol. 29, no. 6, p. e12012, 2019.

[34] P. D. O'Brien and R. C. Nicol, "Fipa—towards a standard for software agents," *BT Technology Journal*, vol. 16, no. 3, pp. 51–59, 1998.

[35] M. A. Karzan and N. Erdogan, "Topic based agent migration scheme via publish/subscribe paradigm," *International Journal of Information and Education Technology*, vol. 3, no. 3, p. 290, 2013.

[36] M. Lutz, *Learning python: Powerful object-oriented programming*. " O'Reilly Media, Inc.", 2013.

[37] P. S. Foundation, "pickle — python object serialization — python 3.9.2 documentation," https://docs.python.org/3/library/pickle.html, 03 2021, (Accessed on 03/18/2021).

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[39] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation," in *AI 2006: Advances in Artificial Intelligence*, A. Sattar and B.-h. Kang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1015–1021.

[40] D. M. Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[41] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2210832718301546

[42] Provost, F and Fawcett, T, "Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions," in *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, 1997, pp. 43–48.

http://doi.org/10.18239/jornadas_2021.34.04

Un resumen: Un enfoque para la aplicación de un clasificador dinámico de clases múltiples para sistemas de detección de intrusiones en la red

Xavier A. Larriva-Novo, Carmen Sánchez-Zas, Víctor A. Villagrá, Mario Vega-Barbas, Diego Rivera
Universidad Politécnica de Madrid (UPM). DIT, ETSI Telecomunicaciones. Avda. Complutense, 30. 28040 Madrid
{xavier.larriva.novo,carmen.ssaz,victor.villagra,mario.vega,diego.rivera }@upm.es

Resumen- **Actualmente, el uso de modelos de aprendizaje automático para el desarrollo de sistemas de detección de intrusos es una tendencia tecnológica que proporciona una gran mejora. Estos sistemas inteligentes se entrenan con conjuntos de datos etiquetados, que incluyen diferentes tipos de ataques y el comportamiento normal de la red. La mayoría de los estudios utilizan un modelo de aprendizaje automático único, identificando anomalías relacionadas con posibles ataques. En otros casos, los algoritmos de aprendizaje automático se utilizan para identificar cierto tipo de ataques. Sin embargo, estudios recientes muestran que ciertos modelos son más precisos para identificar ciertas clases de ataques que otros. Por lo tanto, este estudio propone un modelo adaptativo a cada tipo de ataque a partir de un conjunto de módulos razonadores. Además, este trabajo de investigación propone organizar estos módulos para alimentar un sistema de selección, un clasificador dinámico, permitiendo aumentar el rango de detección para cada modelo individual en términos de precisión.**

Index Terms- Clasificador Dinamico, Machine Learning, IDS
Tipo de contribución: *An Approach for the Application of a Dynamic Multi-Class Classifier for Network Intrusion Detection Systems, Investigación publicada en la revista ELECTRONICS*

I. Introducción

Recientemente, el uso de sistemas de detección de intrusiones (IDS) está integrado en la mayoría de las redes de las empresas. Este sistema monitoriza el tráfico para encontrar registros atípicos o patrones de ataque, y decide si representan un ataque o no, antes de que haya un daño real a los recursos.

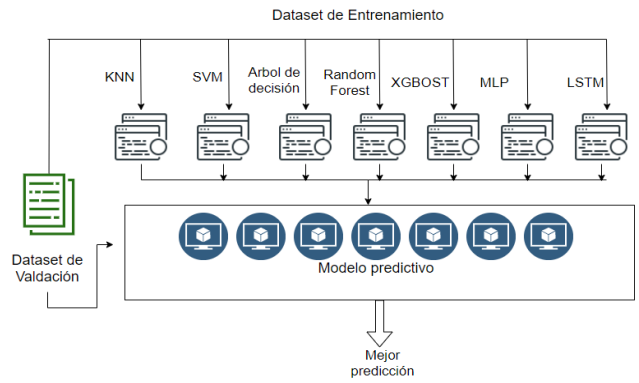


Fig. 1. Arquitectura modelo clasificador dinámico.

Un IDS analiza las anomalías detectadas en su ámbito de actuación, pero, tradicionalmente, este procedimiento es muy estático y es difícil de actualizar y adaptarse a nuevos tipos de ataques. Este proceso se está automatizando gracias a técnicas de Machine Learning, desplegando un clasificador binario o multiclase, entrenado con amplios conjuntos de datos que contienen anomalías y logs de tráfico normal, para ser lo más precisos posible.

Con todo esto, aún existen limitaciones para su detección, ya que los modelos pueden producir algunos resultados erróneos, debido a que no se adaptan por igual a todo tipo de datos. Esta contribución se presenta lo expuesto en [1] que pretende indicar la necesidad de mejorar estos sistemas y reducir el riesgo en los equipos monitorizados, mediante el uso de diferentes algoritmos y seleccionando el que dé mejor resultado a los datos de entrada recibidos en tiempo real.

II. Propuesta y Metodología

Este artículo propone un sistema dinámico en tiempo real para la autoselección de un algoritmo de Machine Learning. Para ello, se estudian diferentes algoritmos y configuraciones para realizar una selección de los candidatos y entrenar los modelos por separado, utilizando el conjunto de datos UNSW-NB15. Posteriormente, se procesarán datos en tiempo real con todos los candidatos, y se compararán las predicciones, para finalmente elegir el resultado más preciso, y así adaptar el sistema a los datos de entrada.

El clasificador dinámico propuesto en este artículo está diseñado para lograr el objetivo final: un sistema capaz de obtener los mejores resultados de predicción de varios algoritmos de aprendizaje automático basados en una clasificación multiclase. Para desarrollar el clasificador dinámico, se requieren modelos previamente optimizados [2]. El sistema propuesto se presenta en la Figura 1, en la que se propone una arquitectura compuesta por diferentes módulos: una serie de algoritmos de aprendizaje automático estáticos preconfigurados manualmente mediante un estudio de selección de hiperparametros y selección de características, y finalmente por un clasificador dinámico.

En primer lugar, se incluye el desarrollo de esa serie de modelos de aprendizaje automático estáticos, capaces de predecir posibles ataques. Cada modelo cuenta con su propia configuración de hiperparametros y una selección de características específicas. Estudios relacionados [3] indican que no todos los modelos pueden lograr una buena precisión de predicción para cada tipo de ataque. En cambio, cada modelo de aprendizaje automático parece ser más adecuado para determinadas categorías de ataques [3].

El clasificador dinámico funciona evaluando la información entrante por cada uno de estos modelos estáticos de aprendizaje automático. La salida de cada uno consiste en una predicción sobre el posible ataque basada en las categorías de ataque definidas por el conjunto de datos UNSW-NB15, el cual se utiliza luego como datos de entrada para el clasificador real. Posteriormente, el clasificador puede determinar la mejor predicción a partir de las generadas por los modelos de aprendizaje automático y seleccionarla independientemente de la categoría de ataque. Este método mejora la precisión de la predicción de ataques.

A. Preparación de datos

Para esta investigación, el conjunto de datos seleccionado fue el UNSW-NB15 [4]. Además, se establece la importancia de seleccionar las características más importantes entre todas las características disponibles teniendo un alto impacto en el algoritmo [4]. Las características se clasificaron según su tipo, es decir, los valores numéricos se transformaron en valores de z-score, y las características categóricas se transformaron en valores numéricos. Para evitar el desbalanceo, algunas investigaciones han utilizado SMOTE para equilibrar el conjunto de datos UNSW-NB15. Este algoritmo se aplicó en esta investigación con el objetivo de comparar los resultados entre un conjunto de datos equilibrado de SMOTE y un conjunto de datos no equilibrado como el conjunto de datos original.

Además, hemos realizado una selección de características basada en el coeficiente de correlación de Kendall mejorando los resultados de la tarea de selección. Para esta investigación, los mejores resultados se obtuvieron al eliminar las características con una correlación superior a 0.8 para conjuntos de datos balanceados y no balanceados.

B. Clasificador dinámico

El clasificador dinámico, como se presenta en la Figura 1, fue diseñado para agregar las predicciones de los modelos de ML individuales y hacer una selección automática de la predicción óptima obtenida de cada uno para una sola muestra, mientras que los modelos se ejecutan en paralelo para hacer predicciones. sobre el conjunto de datos de prueba.

Después de varias pruebas, el clasificador dinámico se diseñó con un modelo aprendizaje automático basado en XGBoost.

El objetivo del clasificador dinámico propuesto es combinar las opiniones de un conjunto de modelos. Cada experto asigna un coeficiente dinámico basado en su propia predicción, sobre la decisión tomada de los datos de entrada. Este coeficiente de peso (escala 0-1) depende de la precisión general de cada modelo individual. Para ello, se creó un conjunto de datos alternativo seleccionando como características de entrada, las predicciones de los modelos individuales mencionados a continuación y, como resultado, el ataque de clasificación deseado. El sistema propuesto clasifica las predicciones de los modelos individuales, y el resultado se basa en las relaciones que el modelo de conjunto ha encontrado exponiendo la más relevante de los algoritmos individuales.

III. DISCUSIÓN

La mayoría de los estudios que proponen IDS basados en algoritmos ML se basan en un modelo estático con un rendimiento mejorado en términos de precisión [5] con una variación sobre sus hiperparámetros.

Tabla I
COMPARACIÓN ENTRE MODELOS ESTÁTICOS Y CLASIFICADOR DINÁMICO EN TÉRMINOS DE PRECISIÓN PARA CONJUNTOS DE DATOS BALANCEADOS Y NO BALANCEADOS.

Modelo	Dataset Balanceado		Dataset Imbalanceado	
	Accuracy	F1-Score	Accuracy	F1-Score
KNN	0.739	0.7406	0.779	0.7673
SVM	-	-	0.739	0.685
DT	0.816	0.8066	0.801	0.8062
RF	0.823	0.824	0.828	0.820
XGBoost	0.795	0.796	0.824	0.803
MLPNN	0.709	0.715	0.811	0.784
LSTMNN	0.747	0.754	0.816	0.792
Clasificador Dinámico	0.876	0.879	0.851	0.829

Este artículo expone a través de varias pruebas que múltiples algoritmos pueden detectar significativamente mejor algún tipo de ataque sobre diferentes configuraciones de hiperparametros como se presenta en la Tabla I.

Siguiendo esta idea, esta investigación propone un clasificador dinámico de autoselección sobre diferentes modelos de ML con el objetivo de obtener las mejores capacidades de cada modelo individual para detectar ciberataques.

IV. CONCLUSIONES

El clasificador dinámico pudo mejorar los resultados en un 5,3% y un 2,3% en comparación con el mejor modelo estático para el conjunto de datos equilibrado y desequilibrado. La idea clave del modelo propuesto en esta investigación es seleccionar automáticamente la mejor tasa de detección, reuniendo las ventajas individuales de cada modelo. Usamos el método basado en un modelo de conjunto basado en XGBoost para mejorar la tasa de detección.

Aunque el modelo aumenta la tasa de detección, se necesita más tiempo en ejecución para detectar un ataque, ya que los datos deben ser procesados por cada algoritmo individual y el modelo de clasificador dinámico. En un escenario práctico, esto podría introducir un pequeño retraso en el tiempo de detección de un posible ataque.

REFERENCIAS

.[1] X. Larriva-Novo, C. Sánchez-Zas, V. A. Villagrá, M. Vega-Barbas, y D. Rivera, «An Approach for the Application of a Dynamic Multi-Class Classifier for Network Intrusion Detection Systems», *Electronics*, vol. 9, n.º 11, Art. n.º 11, nov. 2020, doi: 10.3390/electronics9111759.

[2] A. Aldweesh, A. Derhab, y A. Z. Emam, «Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues», *Knowledge-Based Systems*, vol. 189, p. 105124, feb. 2020, doi: 10.1016/j.knosys.2019.105124.

[3] W. Elmasry, A. Akbulut, y A. H. Zaim, «Empirical study on multiclass classification-based network intrusion detection», *Computational Intelligence*, vol. 35, n.º 4, pp. 919-954, 2019, doi: 10.1111/coin.12220.

[4] X. A. Larriva-Novo, M. Vega-Barbas, V. A. Villagra, y M. Sanz Rodrigo, «Evaluation of Cybersecurity Data Set Characteristics for Their Applicability to Neural Networks Algorithms Detecting Cybersecurity Anomalies», *IEEE Access*, vol. 8, pp. 9005-9014, 2020, doi: 10.1109/ACCESS.2019.2963407.

[5] P. Mishra, V. Varadharajan, U. Tupakula, y E. S. Pilli, «A Detailed Investigation and Analysis of Using Machine Learning Techniques for Intrusion Detection», *IEEE Communications Surveys Tutorials*, vol. 21, n.º 1, pp. 686-728, Firstquarter 2019, doi: 10.1109/COMST.2018.2847722.

http://doi.org/10.18239/jornadas_2021.34.05

Despliegue de técnicas SDNFV para la detección, gestión y mitigación de amenazas a la seguridad de centros de supercomputación (HPC)

Felipe Lemus Prieto
CénitS
Carretera Nacional 521 Km. 41,8
10.071 Cáceres
felipe.lemus@cenits.es

David Cortés-Polo
Universidad de Extremadura
Avda. Universidad S/N EPCC
10.071 Cáceres
dcorpol@unex.es

José-Luis González-Sánchez
CénitS
Carretera Nacional 521 Km. 41,8
10.071 Cáceres
joseluis.gonzalez@cenits.es

Jesús Calle-Cancho
CénitS
Carretera Nacional 521 Km. 41,8 10.071 Cáceres
jesus.calle@cenits.es

Luis Ignacio Jiménez Gil
CénitS
Carretera Nacional 521 Km. 41,8 10.071 Cáceres
luisignacio.jimenez@cenits.es

***Resumen*—Los incidentes de seguridad son cada vez más frecuentes en todo tipo de organizaciones que experimentan el impacto negativo provocado por este tipo de amenazas. Los centros de supercomputación que prestan servicios de computación de alto rendimiento no son ajenos a este tipo de incidentes que afectan a la alta disponibilidad que estos centros críticos deben garantizar. Por ello, la investigación y diseño de infraestructuras que permitan detectar, gestionar y mitigar las amenazas a la seguridad de este tipo de centros de datos es especialmente necesaria. Se presenta la investigación en la *softwarización* y virtualización de red para implementar una infraestructura de red basada en NFV y SDN. Se ha desplegado esta infraestructura de red donde se aplican tecnologías de nueva generación para mitigar los incidentes de seguridad mediante funciones de red virtualizadas y desplegadas en contenedores.**

***Index Terms*—Supercomputación, HPC, seguridad, SDN, NFV**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

Los centros de supercomputación ofrecen recursos de computación de altas prestaciones (High Performance Computing, HPC) escalables, asociados a una institución o grupo de usuarios y adaptados a las necesidades concretas de cada perfil de usuario. La partición de los recursos busca la reproducibilidad de los experimentos en múltiples infraestructuras de computación, para que cualquier investigador pueda, a partir de esos experimentos, continuar con el trabajo previamente realizado. Por ello, los entornos de software modulares [1], entornos virtuales [2] o software científico desplegado en contenedores [3] se han convertido en herramientas cotidianas en los centros de supercomputación, en los que los investigadores desarrollan sus aplicaciones que son ejecutadas sobre los recursos de cómputo. Estas herramientas garantizan, además, la reproducibilidad en los experimentos. Por contra, esta virtualización de recursos en un supercomputador hace mucho más compleja la gestión de la seguridad de la infraestructura. Mecanismos como los cortafuegos (software o hardware) o las listas de control de acceso (ACL) eran suficientes para gestionar de forma eficiente el acceso a la infraestructura y la utilización de los recursos. Sin embargo, el incremento de servicios virtualizados, la interconexión de máquinas en remoto o el uso de recursos en plataformas *cloud* para incorporarlos al cómputo local, han hecho que los mecanismos de seguridad tradicionales no sean suficientes para asegurar que los recursos de cómputo estén siendo usados de forma correcta y equitativa con respecto al resto de los usuarios.

Con estas necesidades, los protocolos de comunicaciones, así como el hardware de red que se está desplegando en los centros de supercomputación, no sólo atienden a métodos tradicionales de computación, sino que están migrando a tecnologías como las redes definidas por software (Software Defined Networks, SDN) o a la virtualización de funciones de red (Network Function Virtualization, NFV) como dos de las principales tecnologías.

Estas dos tendencias en las comunicaciones, basadas en software, proveen una mayor versatilidad a las redes de comunicaciones de forma que pueden ser fácilmente integrables en los centros de supercomputación, siendo, además, herramientas indispensables para gestionar la seguridad en estas infraestructuras [4].

Las arquitecturas SDN y NFV aportan numerosos beneficios a este paradigma de arquitecturas de computación y, por tanto, la integración de ambas en un único marco denominado red definida por software con funciones virtualizadas (Software Defined NFV, o SDNFV) [5] puede resultar especialmente útil. Es por esto por lo que, a través de una SDNFV, se pueden incluir mejoras importantes en los mecanismos de detección y mitigación de los problemas de seguridad en la red [6].

La sección II de este artículo introduce la arquitectura de red SDNFV usada en esta investigación, así como el diseño y el desarrollo de los protocolos que la componen. En la sección III, se presenta la detección y la mitigación de amenazas de seguridad en una red basada en SDNFV, describiendo el algoritmo para detectar un posible problema de seguridad así cómo los mecanismos para mitigarlo usando funciones de red virtualizadas en un escenario real. Finalmente, la sección IV resume las conclusiones del trabajo realizado.

II. DESPLIEGUE DE MECANISMOS DE SEGURIDAD SOBRE UNA RED SDNFV IMPLEMENTADA EN UN CENTRO HPC

En una arquitectura SDN pura el controlador tiene la función principal de gestionar el plano de control y la conmutación de los flujos en la red. Esta función se mantiene en una red basada en SDNFV que define servicios de red complejos que se ejecutarán en un hardware de propósito general, reemplazando el tradicional hardware específico creado para realizar las funciones de comunicación [7]. Dicha función debe complementarse con la orquestación de recursos virtualizados en la red, con el fin de implementar las funciones virtualizadas y administrar el ciclo de vida de la función.

Desde el punto de vista de la seguridad, la función del controlador/orquestador es muy importante porque gestiona servicios clave cuyo propósito es detectar, gestionar y mitigar un ataque, reservar recursos de red para crear *black holes* y/o analizar los flujos de datos de la comunicación.

La Fig. 1, muestra la arquitectura propuesta y cómo se interconectan SDN y NFV mediante el controlador/orquestador desplegado en la red. Como puede observarse, la arquitectura SDNFV se divide en diferentes tipos de nodos interconectados en la red, con dos tipos diferenciados de *switches* SDN.

A continuación, se detallan individualmente las tecnologías usadas en la red SDNFV, especificando la implementación de SDN y de NFV usadas para, a través de la conjunción de ambas, desarrollar un sistema de detección, gestión y mitigación de un ataque real.

II-A. Implementación de la red SDN

En la arquitectura propuesta, la red SDN está compuesta por dos tipos de *switches*. Los *switches OpenvSwitch* (OVS) [8] no son *switches* inteligentes y su función principal es conmutar los paquetes por la red siguiendo las reglas marcadas por el controlador. Este *switch* también se implementa en los diferentes clústeres de contenedores usados para desplegar los *dockers* que implementan la arquitectura NFV.

El *switch CPqD* [9], modificado por el grupo de trabajo del proyecto *BEhavioural BAsed forwarding* (BEBA) [10], implementa la propuesta *OpenState* [11], ampliando las funcionalidades principales de *OpenFlow* e incluyendo las capacidades para aplicar diferentes reglas basadas en la coincidencia con los estados descritos en las tablas de flujo SDN del *switch*.

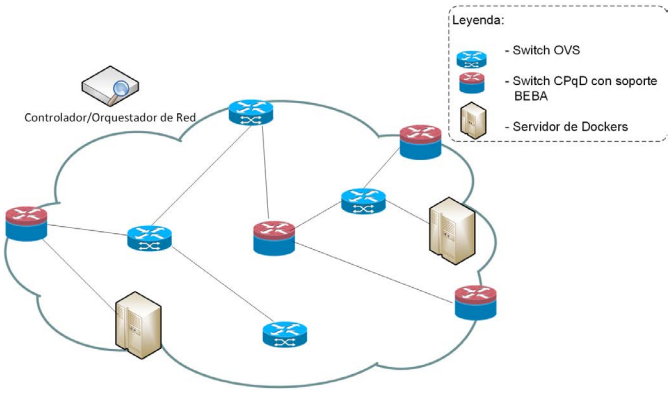


Fig. 1. Red SDNFV propuesta

Con esta funcionalidad, el *switch* adquiere la capacidad de reaccionar ante eventos a nivel de paquete, por ejemplo, analizando los flujos que se están conmutando. Si el resultado del análisis coincide con las reglas que el *switch* tiene en las tablas de flujo, entonces puede actuar de acuerdo con la regla.

La Fig. 2 muestra la arquitectura de *OpenState* con la que se ha conseguido que la implementación de las aplicaciones de red se simplifique y aumente el rendimiento de estas [11].

Se ha modificado el código del *switch CPqD/ofsoftswitch13* para implementar las capacidades *OpenState*, introduciéndolas como una extensión experimental de *OpenFlow*. Para poder realizarlo, la especificación *OpenFlow* define la estructura común de los campos experimentales, con un conjunto de mensajes y acciones que deben ser definidos.

II-B. Implementación de las funciones NFV

La arquitectura NFV se implementa mediante *dockers* en un clúster de contenedores. El controlador de red debe conocer la información del clúster la cual incluye información del hardware implementado, así como el rango IP usado por el clúster. Todas estas características definen el NFVI.

Con esta información, el controlador puede gestionar los recursos y desplegar las funciones virtualizadas. Como ya se ha indicado anteriormente, las funciones de la red están empaquetadas en contenedores *docker*, de modo que la implementación de la función es independiente de la plataforma utilizada para desplegarla.

El NFVI se interconecta con la red SDNFV a través de interfaces de red física (pNIC). Estas NIC intercambian los paquetes entre las funciones de red virtualizadas y la red SDN. Cada función virtualizada incluida en un contenedor tiene dos interfaces de red virtual, que están interconectadas a través de un *bridge*.

Con este enfoque, un pequeño clúster puede implementar un gran número de instancias, configurando la conexión de red y las interfaces que permiten la comunicación entre los contenedores.

Los *dockers* implementan su funcionalidad básica en un fichero conocido como *Dockerfile*. Este archivo incluye un conjunto de instrucciones para construir automáticamente el entorno en una imagen de *docker*. Esta imagen crea una instancia en el NFVI para ejecutar la función en la red SDNFV. La comunicación entre el controlador/orquestador y el NFVI se cifra mediante un canal seguro para enviar la configuración de la función de red virtualizada, crear el *Dockerfile* y los comandos necesarios para implementar la función virtualizada en la infraestructura.

II-C. Implementación del Controlador/Orquestador

El controlador desplegado en la arquitectura es Ryu [12] y está basado en componentes que son desarrollados para implementar las distintas funcionalidades.

Para incluir las funciones básicas de BEBA en Ryu, se deben implementar nuevos mensajes, acciones y campos de comprobación. De hecho, para que el controlador gestione las capacidades BEBA, su implementación básica debe ampliarse, incluyendo las capacidades de gestión de toda la información que proporcionan las diferentes tablas de gestión de flujo.

La implementación de BEBA utiliza la funcionalidad aportada por *OpenFlow* para incorporar mensajes experimentales,

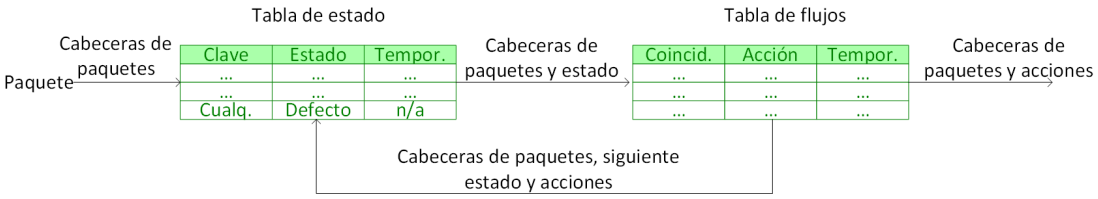


Fig. 2. Arquitectura *Openstate*

usados para implementar la lógica de control, así como definir el *payload* de los paquetes que deben transportar la información requerida para conmutar el tráfico, tomar decisiones sobre los flujos o ejecutar reglas en la red. Para esta investigación todas las modificaciones se han realizado sobre los *switches* CPqD, de forma que, a la instalación básica de Ryu, se le ha incorporado la implementación de BEBA.

Esta implementación ha sido extendida en su funcionalidad para incluir el orquestador de NFV, obtener la información del NFVI y desplegar las funciones virtuales sobre el clúster de *docker*, implementando la funcionalidad de cortafuegos y ser capaz de configurar la red de forma que redirija el tráfico a la función virtualizada de red.

La Fig. 3 describe la arquitectura completa del controlador/orquestador. Como se observa, la arquitectura de Ryu se amplía para integrar las primitivas de BEBA, de forma que la aplicación SDN puede hacer uso de esas primitivas para desarrollar nuevas funcionalidades como la propuesta.

La aplicación desarrollada para este despliegue se sustenta sobre BEBA para analizar los flujos y detectar los que son sospechosos. Con esta información el módulo de orquestación NFV elige el mejor servidor de *dockers* para desplegar el contenedor que implementa la función NFV. Para ello se crea un canal seguro y se ejecutan los comandos necesarios para su despliegue. Una vez se ha realizado la implementación, el controlador modifica las tablas de flujo para redirigir el tráfico seleccionado al contenedor NFV que mitigue el ataque.

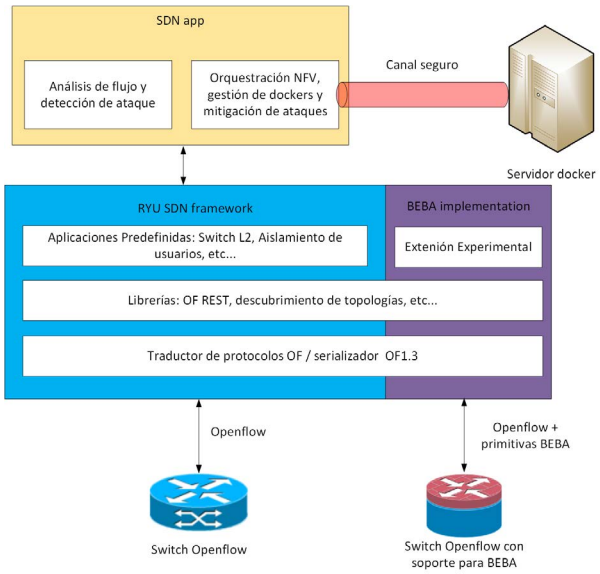


Fig. 3. Arquitectura controlador/orquestador

III. DETECCIÓN Y MITIGACIÓN DE ATAQUE EN RED SDNFV

Un centro de supercomputación es un escenario especialmente crítico por lo que ha sido necesario crear un entorno seguro para realizar las primeras pruebas sobre la red SDNFV propuesta en el presente artículo. Además, dicho entorno permite simular el despliegue de nuevas funciones NFV sin necesidad de exponer la infraestructura real del centro y comprobar de forma segura que su funcionamiento es el adecuado. Para ello se han desplegado dos servidores con el sistema operativo GNU/Linux.

En el primero de los servidores se encuentra instalado el controlador Ryu y los módulos encargados de la detección y mitigación de los ataques. En el segundo servidor se ha simulado la red mostrada en la Fig. 1, utilizando la herramienta Mininet [13] que permite crear redes virtuales muy realistas y soporta *Openflow*. En este caso en concreto, se ha configurado un *switch* OVS para que soporte BEBA y así poder reenviar al controlador estadísticas de tráfico. También se encuentra instalado *Docker* en este segundo servidor y es el que recibe las instrucciones del controlador mediante una conexión SSH para desplegar la función NFV que corresponda. El funcionamiento del escenario global se describe a continuación.

El controlador/orquestador implementa dos módulos para detectar y mitigar los ataques. La fase de mitigación requiere la información del flujo: @IP de origen, puerto de origen, @IP de destino, puerto de destino y el protocolo utilizado. Toda esta información se obtiene en la fase de detección. El módulo de detección es un subproceso que, periódicamente, envía una solicitud a los conmutadores *OpenFlow* para adquirir información estadística. El Algoritmo 1 describe el bucle principal del análisis de los flujos.

En este bucle, se solicitan las estadísticas de los flujos al conmutador, utilizando la primitiva implementada en BEBA: *OFPExpStateStatsMultipartRequestAndDelete*. Para recibir una respuesta del conmutador, se ha creado un controlador de eventos que recibe el mensaje *EventOFPExperimenterStatsReply* con la obtención de las estadísticas. Estas estadísti-

Algoritmo 1 Análisis de flujos

```
1: mientras verdad hacer
2:     Envia peticiones de estados a la tabla de estado.
3:     para datapath en datapaths hacer
4:         Obtener estadisticas del datapath
5:         Espera X segundos
6:     fin para
7: fin mientras
```

cas se analizan obteniendo la IP origen y destino del paquete, así como el puerto destino y el tipo de protocolo de nivel de transporte, TCP o UDP, para recontar los diferentes flujos que se están analizando.

Una vez analizados los flujos, se calcula la entropía. Esta se usa para detectar ataques DDoS midiendo las propiedades estadísticas del encabezado del paquete. En este caso, la muestra de datos reales analizados se basa en la comparación de los paquetes consecutivos de un flujo para identificar un ataque, de forma que la entropía puede ser tratada como se muestra en la Ec. (1):

$$E = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

Donde E es la entropía, n el número de elementos detectados en el análisis de los flujos y p_i la probabilidad de encontrar el elemento *iésimo* en la conjunción de elementos detectados en el análisis. En este trabajo, hemos utilizado la @IP de origen, de destino, el puerto de origen y el puerto de destino como elementos para detectar la entropía. Una vez calculada la entropía, se ejecuta el algoritmo de detección. El algoritmo se basa en las Ec. (2 a 4).

$$l_{inf} = \bar{x}_p - precision * \sigma_p \tag{2}$$

$$l_{sup} = \bar{x}_p + precision * \sigma_p \tag{3}$$

$$Ataque = \begin{cases} falso & Si\ l_{inf} < x < l_{sup} \\ verdadero & cualquier\ otro \end{cases} \tag{4}$$

Donde $\bar{x}_p$ es el valor medio de los elementos analizados y *precision* es el valor para definir la precisión en el algoritmo de detección. En este caso, los valores utilizados para la precisión son de un 68 % para precisión baja; de un 95 % para precisión media; y un 99.7 % para una precisión alta.

Si se detecta el ataque, el controlador/orquestador busca en una base de datos donde se encuentra la información sobre el clúster de *dockers* para elegir el servidor que puede implementar de forma eficiente la función NFV. Una vez que se elige el servidor, el controlador abre un canal seguro para desplegar el NFV en un contenedor *docker*. El despliegue se realiza mediante el archivo descriptor de *Dockerfile*. Este archivo describe la función NFV y el comportamiento del *docker*.

El *Dockerfile* define las reglas con las que se lanzará el *docker*. Una vez que se construye el contenedor *docker*, se crean las interfaces y se establece el *bridge* entre ellas. Todo ello es desplegado en el servidor elegido entre todos los del clúster de *dockers*. El controlador, una vez analizada la fase de despliegue, modifica las tablas de reenvío de los conmutadores involucrados en la comunicación del ataque DDoS para mitigarlo. Con este mecanismo, los conmutadores que reciben el ataque solo tienen que reenviar los paquetes al puerto de salida, y el ataque se mitiga utilizando un servidor diseñado para absorberlo sin que se produzca un DDoS en el nodo destino elegido por el atacante. En este caso, el destino es el *firewall* desplegado como una función NFV, el cual es una implementación software de *firewall* con las reglas pertinentes para filtrar el trafico malicioso.

IV. Conclusiones

Este trabajo presenta la investigación y el despliegue de una arquitectura SDNFV, aplicable a una red de computación, que admite la detección y mitigación de ataques DDoS utilizando funciones virtualizadas en un contenedor de aplicaciones.

La arquitectura presentada es una red simple y de bajo coste basada en SDN que analiza los flujos y detecta los ataques utilizando un mecanismo de entropía desarrollado en el controlador. Este mecanismo es más avanzado en términos de detección y mitigación, porque sólo se mitiga el flujo involucrado en el ataque. El mecanismo puede discriminar diferentes valores como IP, puertos o protocolos. Se pueden implementar diferentes funciones virtualizadas según la aplicación que se ejecute en la parte superior del controlador Ryu.

Se presenta la función virtualizada de un *firewall* y, además, las capacidades de la arquitectura desarrollada permiten implementar otras funciones como el conformado de tráfico, la inspección profunda de paquetes, analizadores de flujo de datos, etc. La simplicidad de la arquitectura desarrollada hace transparente el despliegue de las funcionalidades de la red para el usuario final. El controlador/orquestador propuesto gestiona el plano de control y administra los recursos de la red para aumentar el rendimiento.

V. AGRADECIMIENTOS

Esta investigación ha sido financiada en parte por el Ministerio de Ciencia, Innovación y Universidades bajo el proyecto Go2Edge (RED2018-102585-T).

Referencias

[1] J. L. Furlani and P. W. Osel, "Abstract yourself with modules," in {*USENIX*} *10th Systems Administration Conference ({LISA} 96)*, 1996.

[2] J. Smith and R. Nair, *Virtual machines: versatile platforms for systems and processes*. Elsevier, 2005.

[3] N. A. Snow, V. R. Dasari, and B. E. Geerhart, "Openflow experimenter labels for encoding adaptive network functions," in *2018 IEEE 39th Sarnoff Symposium*. IEEE, 2018, pp. 1–5.

[4] J. Higgins, V. Holmes, and C. Venters, "Securing user defined containers for scientific computing," in *2016 International Conference on High Performance Computing & Simulation (HPCS)*. IEEE, 2016, pp. 449–453.

[5] W. Wang, Y. Liu, Y. Li, H. Song, Y. Wang, and J. Yuan, "Consistent state updates for virtualized network function migration," *IEEE Transactions on Services Computing*, 2017.

[6] D. Cortés-Polo, F. Lemus-Prieto, J. Calle-Cancho, L. I. Jiménez Gil, and J. L. González-Sánchez, "Gestión de la seguridad de las comunicaciones para entornos HPC en centros de supercomputación," in *Jornadas SARTECO 2019*, 2019.

[7] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, 2015.

[8] B. Pfaff, J. Pettit, T. Koponen, E. Jackson, A. Zhou, J. Rajahalme, J. Gross, A. Wang, J. Stringer, P. Shelar *et al.*, "The design and implementation of open vswitch," in *12th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 15)*, 2015, pp. 117–130.

[9] R. Bifulco and A. Matsiuk, "Towards scalable SDN switches: Enabling faster flow table entries installation," *ACM SIGCOMM Computer Communication Review*, vol. 45, no. 4, pp. 343–344, 2015.

[10] BEBA, "BEBA SDN project home page," http://www.beba-project.eu/, Último acceso: 2021-02-16.

[11] G. Bianchi, M. Bonola, A. Capone, and C. Cascone, "Openstate: Programming platform-independent stateful openflow applications inside the switch," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 2, pp. 44–51, 2014.

[12] Ryu project team, "Ryu SDN framework," https://ryu-sdn.org/, Último acceso: 2021-03-11.

[13] Mininet, "Mininet: An Instant Virtual Network on your Laptop (or other PC)," http://mininet.org/.

Sesión de Investigación A1: Detección de intrusiones y gestión de anomalías I
http://doi.org/10.18239/jornadas_2021.34.06

Evaluation of state-of-art phishing detection strategies based on machine learning

F. Castaño*†[a], M. Sánchez-Paniagua*†[b], J. Delgado*[c], J. Velasco-Mata*†[d], A. Sepúlveda†[e], E. Fidalgo*†[f], E. Alegre*†[g]
*Department of Electrical, Systems and Automatics Engineering, University of León, Spain
†Researcher at INCIBE (Spanish National Cybersecurity Institute), León, Spain
Email:{**felipe.castano**, **manuel.sanchez**, javier.velasco, eduardo.fidalgo, enrique.alegre}@unileon.es,
jdelgs01@estudiantes.unileon.es, antonio.sepulveda@incibe.es
ORCID iDs: [a]0000-0001-9157-4111, [b]0000-0002-0788-916X, [c]0000-0002-3431-8565, [d]0000-0002-7923-1658,
[e]0000-0001-9488-8693, [f]0000-0003-1202-5232, [g]0000-0003-2081-774X

***Abstract*—phishing is one of the most common cyber-attacks. Machine Learning approaches can effectively deal with Phishing detection. However, models are trained on datasets with landing pages as legitimate samples without login forms, which is a situation closer to the real-world problem. In this work, we presented the Phishing Index Login URL (PILU-60K), a dataset with URLs of both index pages and login pages. Besides, five of the most used Machine Learning models were implemented and tested on PILU-60K and compared with well-known datasets. We used the models trained on index pages and tested on login pages to determine if the performance was affected when the models have to classify login URLs. Also, we reviewed the performance of the models over time, trained with datasets from 2016 and 2017, and tested them on recent ones. Results showed that models lose up to 14.5% of accuracy compared to the reported performance.**

***Index Terms*—Cybersecurity, Phishing Detection, URL, Artificial Intelligence, Machine Learning, NLP**

Contribution: *Extended summary of Impact of current phishing strategies in machine learning models for phishing detection* [1]

I. Introduction

One of the most frequent cyber-attacks on the internet is phishing [2]. According to the Anti-Phishing Working Group (APWP) report [3], there was an increasing number of phishing attacks since March 2020, reaching around 225.000 detected cases in the last quarter of 2020.

Phishing detection through blacklists can block phishing websites that have been previously registered [4]. However, these lists have to be constantly updated and they can not detect zero-day phishing attacks. Due to these flaws and the increasing number of phishing web pages, researchers rely in approaches based on Machine Learning [5], [6].

We consider that detecting if a login form of a website is phishing is closer to a real-world scenario than analysing landing pages. However, phishing datasets usually contains the landing page from known legitimate websites and, in many cases, the login form is not there. Besides, attackers constantly change their methods and URLs of the phishing web pages [7]. In the past, identifying a phishing website could be done just looking at its http protocol, but the APWG [8] reports that phishing websites under HTTPS have increased from a 10% in 2017 to 84% in 2020 [3]. For that reason, the performance of detecting methods trained with old data might be affected by the evolution of newer phishers strategies. In this work [1] we presented Phishing Index Login URL (PILU-60K), a dataset that contains URLs from legitimate login pages where we considered a more representative real-world scenario for Phishing Detection.

II. Methodology

A. Phishing Index Login URL

PILU-60K dataset [1] contains 60000 URLs classified into three balanced categories, with 20K samples each: legitimate index pages URLs, legitimate login pages URLs and phishing pages URLs. Legitimate samples were collected from Top Million Quantcast[2]. We used Selenium to reach the login page from the website. Phishing URLs were taken from Phishtank[3], from November 2019 to January 2020.

B. Models and Features

First, we extracted the 38 descriptors proposed by Sahingoz et al. [9] from each URL, including NLP features and URL rules. We used these descriptors to train the following five Machine Learning models, which are commonly used in the literature for Phishing detection [9], [10], [11]: k-Nearest Neighbours, Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machines. The parameters were set manually using the values that returned the best accuracy.

III. Experimental settings

A. Datasets

We tested the trained models using the following state-of-the-art datasets: 1 Million Phishing (1M-PD) [12], Ebbu2017 [9], Phishing Websites Dataset 2016(PWD2016) [13] and two subgroups of PILU-60K: PLU-40k, with legitimate login samples and phishing samples, and PIU-40k, with legitimate index samples and phishing samples.

We evaluated how current phishing strategies influenced the performance of machine learning models, training some models with outdated datasets. For this purpose, we divided the models into two groups, one was trained with PWE2016 and the second one was trained with Ebbu2017. Then we evaluated the models against more recent datasets, as showed

[1]Dataset available at: http://gvis.unileon.es/dataset/pilu-60k/
[2]https://www.quantcast.com/products/measure-audience-insights/
[3]https://www.phishtank.com/

in Table I. It is noticeable that URLs on PWD2016, PIU-40k, and 1M-PD do not contain paths since they use index page URLs, but in contrast, Ebbu2017 and PLU-40k have paths.

To check if current Machine Learning methods were useful in real-world scenarios, we took the models trained on a dataset with index URLs and phishing URLs and tested them on samples that included login URLs and phishing URLs. To this end, we created the initial models using the PIU-40k dataset and the widely used data split, 70% of its data for training and the remaining 30% for verifying the models, and then we evaluated them on PLU-40K.

We reported the results using accuracy as the main metric since it is a frequently used value on Phishing Detection methods [9], [7], [6], and besides we offered the F1-Score.

B. Results

Table I shows that the performance of all models, trained with datasets from 2016 and 2017, decreased when they were evaluated in more recent datasets. Among the models trained with PWD2016, the best result was obtained by kNN with 97.35% accuracy but its performance was the most affected over time, reduced to 82.85%. On the contrary, SVM was the model less affected with the reduction of performance, losing 6.89% of accuracy, from $93,46$.

TABLE I
PHISHING DETECTION ACCURACY OVER TIME (%)

Training set:	PWD2016			Ebbu2017	
Test set:	PWD2016	1M-PD	PIU-40K	Ebbu2017	PLU-40K
RF	97.31	90.64	86.04	**95.85**	67.94
kNN	**97.35**	87.22	82.85	93.77	64.77
SVM	95.62	**91.91**	**88.73**	93.17	**71.11**
NB	88.97	86.51	85.84	87.61	65.30
LR	93.46	88.73	85.96	79.51	64.42

Random Forest obtained the highest performance on PILU-60K dataset, as can be seen in Table II. It classified login and index URLs from Phishing with a 92.47% and 94.59% of accuracy, respectively. "PIU-40K throwback" is a test set for models trained with PIU-40k and verified on outdated samples of PWD2016, where Random Forest obtained the highest performance with a 91.22% of accuracy. Finally, another model was trained with PIU-40K and then tested on PLU-40K, showing that models trained with index URLs data lose accuracy when a login URL had to be classified as phishing or not, which was a more representative real-world scenario.

IV. CONCLUSIONS AND FUTURE WORK

In this work, we presented PILU-60K, a dataset that contains URLs from legitimate login pages, legitimate index pages and phishing pages, what we consider a more representative real-world scenario for Phishing Detection. We reviewed the performance of five machine learning models for phishing detection on PILU-60K. The results showed that models trained on a dataset with index URLs had a lower performance classifying login URL samples.

We reviewed the performance of five machine learning models for phishing detection over time. Systems based on these models lose accuracy due to the constant changes of both strategies and URLs used for phishing attacks. For this reason, we recommend that phishing detection models to be used in a real scenario should be trained using legitimate recent login websites instead of landing pages.

As future work, we will generate a larger dataset, including more information in the samples, such as HTML source code or screenshots, so researchers will have more possibilities to generate new methods using any feature extracted directly from the raw data. We will also work on finding new features for phishing detection due to the techniques of phishing web pages are constantly changing and this makes easy to bypass current methods.

TABLE II
PERFORMANCE USING THE PROPOSED DATASET (%)

	PIU-40K		PLU-40K		PIU-40K throwback		PIU-40K vs login URLs
	Acc.	F1Score	Acc.	F1Score	Acc.	F1Score	Acc.
RF	**94.59**	**94.60**	**92.47**	**92.49**	**91.22**	**92.03**	54.06
SVM	93.85	93.80	90.68	90.59	88.66	89.70	61.49
kNN	93.13	92.99	89.63	89.40	85.36	86.15	64.28
LR	92.55	92.45	85.40	85.11	87.02	89.02	**67.60**
NB	87.60	86.63	74.34	72.57	86.91	87.25	58.22

V. ACKNOWLEDGEMENT

This research was funded by the framework agreement between the University of León and INCIBE (Spanish National Cybersecurity Institute) under Addendum 01.

REFERENCES

[1] M. Sánchez-Paniagua, E. Fidalgo, V. González-Castro, and E. Alegre, "Impact of current phishing strategies in machine learning models for phishing detection," in *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020)*, Á. Herrero, C. Cambra, D. Urda, J. Sedano, H. Quintián, and E. Corchado, Eds. Cham: Springer International Publishing, 2021, pp. 87–96.

[2] A. Ferreira and S. Teles, "Persuasion: How phishing emails can influence users and bypass security measures," *International Journal of Human Computer Studies*, vol. 125, pp. 19–31, may 2019.

[3] Anti-Phishing Working Group, "Phishing Activity Trends Report 4 Quarter," 2020.

[4] S. Chanti and T. Chithralekha, "Classification of Anti-phishing Solutions," *SN Computer Science*, vol. 1, no. 1, pp. 1–18, jan 2020.

[5] L. Halgaš, I. Agrafiotis, and J. R. Nurse, "Catching the Phish: Detecting Phishing Attacks Using Recurrent Neural Networks (RNNs)," in *Lecture Notes in Computer Science*, vol. 11897 LNCS. Springer, aug 2020, pp. 219–233.

[6] R. S. Rao and A. R. Pais, "Jail-Phish: An improved search engine based phishing detection system," *Computers and Security*, vol. 83, pp. 246–267, jun 2019.

[7] M. A. Adebowale, K. T. Lwin, E. Sánchez, and M. A. Hossain, "Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text," pp. 300–313, jan 2019.

[8] Anti-Phishing Working Group, "Phishing Activity Trends Report 3 quarter," 2017.

[9] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Systems with Applications*, vol. 117, pp. 345–357, mar 2019.

[10] M. Moghimi and A. Y. Varjani, "New rule-based phishing detection method," *Expert Systems with Applications*, vol. 53, pp. 231–242, jul 2016.

[11] R. S. Rao and A. R. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," *Neural Computing and Applications*, vol. 31, no. 8, pp. 1–23, 2018. [Online]. Available: https://doi.org/10.1007/s00521-017-3305-0

[12] H. Yuan, Z. Yang, X. Chen, Y. Li, and W. Liu, "URL2Vec: URL modeling with character embeddings for fast and accurate phishing website detection," *Proceedings - 16th IEEE International Symposium*, pp. 265–272, 2019.

[13] K. L. Chiew, E. H. Chang, C. L. Tan, J. Abdullah, and K. S. Yong, "Building Standard Offline Anti-phishing Dataset for Benchmarking," *International Journal of Engineering & Technology*, vol. 7, no. 4.31, pp. 7–14, 2018.

Sesión de Investigación A1: Detección de intrusiones y gestión de anomalías I
http://doi.org/10.18239/jornadas_2021.34.07

Automating Intrusion Detection Systems in Smart Contracts

Xabier Echeberria-Barrio*‡, Francesco Zola*†§, Lander Segurola-Gil*, Raul Orduna-Urrutia*
* Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Paseo Mikeletegi 57, 20009 Donostia/San Sebastian, Spain
{xetxeberria, fzola, lsegurola, rorduna}@vicomtech.org
†Institute of Smart Cities, Public University of Navarre, 31006 Pamplona, Spain
ORCID: ‡0000-0001-6836-2890, §0000-0002-1733-5515

***Abstract*—Blockchain technology has gained a lot of relevance in recent years, specifically the smart connect functionality, due to its great potential to decentralize different day-to-day scenarios. This technology brings many advantages thanks to its immutability, specifically smart contracts give blockchain a very great versatility. However, these smart contracts can bring many threats to the blockchain, considering that vulnerabilities they may have, which are immutable once deployed, can be exploited to attack them. Several studies have focused on studying these vulnerabilities in order to identify and fix them, thus deploying smart contracts in a secure way.**

This study has focused on developing an intrusion detection system (IDS) based on the path study. This IDS is deployed with the smart contract to be defended, which will monitor the transactions received by the defended smart contract looking at where the transaction comes from. This generates a path in the blockchain network, which will have some characteristics that are extracted and analyzed. This will detect possible threats received. Our approach suggests the implementation of an automated IDS in the smart contracts to defend them from possible threats.

***Index Terms*—Cybersecurity analysis, Behavioural classification, Smart contract analysis, Intrusion Detection**

Tipo de contribución: *Investigación en desarollo*

I. Introduction

In 2009 blockchain technology was first introduced and has been gaining interest in recent years. This technology started as a distributed ledger for peer-to-peer payments, but has evolved by adding programmable transactions, thus introducing the concept of smart contracts.

By adding the functionality of smart contracts, Ethereum revolutionized blockchain technology. Smart contracts can be considered complete programs that run on the blockchain, where they are deployed by transactions and generate a new block for the blockchain. A smart contract can be developed for many purposes and can execute transactions with assets of considerable value. Each user can send a transaction to the contract address, executing the contract. Furthermore, all transactions and parameters are visible to everyone, so they are subject to intense hacking activity. These bugs are exploited as vulnerabilities to exploit for malicious purposes and may not be patched. In 2016, Ethereum suffered the DAO attack [1], where the adversary took advantage of a re-entry vulnerability, to manage to drain more than $150 million cryptocurrencies from some smart contract. Therefore, this kind of attacks increase the need to study smart contract vulnerabilities, in order to fix them before their deployment.

In this work, the goal has been to protect smart contracts after deployment by developing an IDS based on the ContractGuard idea of studying transactions as paths. However, the proposed IDS instead of relying on an administrator, relies on machine learning (automating the system) that will tell us if a path is anomalous or not through some features. In addition, to obtain important data for each route, several features have been calculated, from which the machine learning model learns. Thus, this paper presents an automated smart contract intrusion detection system.

The rest of the paper is divided as follows: Section II explains the presented intrusion detection system, detailing each component. Finally, the first results are presented in section III-B.

II. Methodology

A. Problem description

In recent years, many vulnerability detection studies have analyzed the security of smart contracts with the aim of increasing their resilience and their ability to prevent cyber-attacks. In particular, this need has increased since several attacks on different blockchains, such as Ethereum, and the millions of dollars in losses they have caused, were identified. These attacks were able to be carried out thanks to the existence of vulnerabilities in smart contracts. These attacks and vulnerabilities are compiled in [1], [2].

1) *Re-entrancy vulnerability*: This consists of the possibility to re-entering to smart contract by calling non-recursive functions before their completion.
2) *Gasless vulnerability*: This consists of the possibility to occur a out-of-gas, stopping the process.
3) *Overflow/underflow vulnerability*: This consists of the possibility to pass a combination of input values that generate a result larger/smaller than the maximum/minimum value that the data type can contain.
4) *Denial of service vulnerability*: This consists of the possibility to become the smart contract inoperable for a short period of time, or in some cases permanently
5) *Keep private vulnerability*: This allows information that should be private to be obtained from the smart contract.
6) *Superficial randomness vulnerability*: This appears when the smart contract makes decisions using state variables such as the timestamp of transactions, as these variables can be manipulated by miners.

7) *Dangerous delegatecall vulnerability*: appears when a contract uses *delegatecall* to carry out an external function, executing it in the context of the calling contract.

Many commonly used smart contracts can be vulnerable to serious malicious attacks, for this several methods are implemented to address the flaws and to update the software to improve its performance and functionality after or before delivery [3].

Most of them have focused on detecting vulnerabilities to avoid deploying smart contracts with them, since once deployed smart contracts are immutable. These developed tools are *offline checking methods* to prevent attacks that can exploit these vulnerabilities. However, the IDS developed in this work is an *online checking method* to avoid these attacks, since it defends the smart contract once deployed. This type of tool, has also been proposed [4]–[11], among which is the IDS on which this work is based [2].

B. Paths in Blockchain

A path in a graph is a finite or infinite sequence of edges that joins a sequence of vertices. When working on blockchain, concretely in Ethereum, it can be represented as a graph, where the edges are the transactions and the vertices or nodes are the used addresses. Thus, the term path makes sense in the context of the blockchain. The Ethereum nodes can be of two different types: user nodes or smart contract nodes. These nodes are the vertices of the paths that exist in the Ethereum. This work will focus on a type of path that can be found in the Ethereum. These paths are formed by a user node that will be the first one in the path and one or more smart contract nodes, which will be all but the first one. Mathematically, being $a_1, a_2, \ldots, a_n$, the nodes of the path A, for $n > 1$, then we will work with paths A, where a_1 is a user node and $a_2, \ldots, a_n$ are smart contract nodes.

The paths, from which the received transactions come, give extra information, which helps in detecting whether a received transaction is a threat or not. Thus, if the path of the received transaction is anomalous, it could mean that this call is malicious. When working with machine learning it is possible enhance the ability to describe the path adding new attributes of the paths to be studied. Thus, the machine learning model will be able to detect such attacks through the features of the paths. The Table I shows which of mentioned vulnerabilities can be detected through path study and which can not be detected.

TABLE I: Characterization of attacks on paths

Attacks	Perceptible by Paths
Re-entrancy	Yes. Generate a cyclic paths
Gasless	Yes. Gas costs can be represented
Overflow/Underflow	Yes. Suspicious values can be characterized
Denial of service	Yes. Loops are type of paths
Superficial randomness	None.
Keep private	None.
Dangerous delegatecall	Yes. The functions called can be characterized

C. Intrusion Detection System

The developed IDS defends smart contracts that contain a communication with it. This communication is executed each time the smart contract is called. In such cases, the smart contract sends the received transaction information to the developed IDS (Fig. 1). This information is received by the path monitoring which monitors where the call comes from, generating the path from the first user encountered. This generated path is passed to the feature extractor, computing the determined attributes, which are sent to the detector. Finally, the detector makes the decision whether the path is anomalous or not. If it is anomalous, the smart contract does not carry out the received call and if it is not anomalous, it executes the function normally.

1) Path monitoring: This component calculates the path from which the received transactions come from. It takes the address that sends the call and using the blockchain, adds the previous addresses if these addresses are related through transactions. In particular, it collects the historical information related to each address. This process is repeated until the related address is a user address, which represents the end of the process being the last address in the route. Once the path calculation is finished, it is reversed. Therefore, the generated paths start with the user address and continues with at least one smart contract address, where the last address is the one that sent the transaction. In addition, these generated paths contain more details about themselves. These details are the values of each transaction of the paths, the duration time of the paths and the function of the smart contract that has been called. Therefore, each time a transaction arrives at the smart contract, the path monitoring generates a path with these features and sends it to the feature extractor.

2) Feature extractor: This component receives the path from the path monitoring, and it extracts the features which the detector needs to compute the classification. The feature extractor uses every detail that the received path has in order to generate several features, representing that path in the most appropriate way. These attributes are used to detect generated threats by different vulnerabilities that smart contracts may have. Therefore, the appropriate attributes are those that characterize known threats. Thus, the feature extractor component generates a vector with these explained features of each path it received from path monitoring. Once the feature vector is computed, it is used as input for the next component, called detector.

3) Detector: The aim of this module is to detect path anomaly analyzing the input feature vector. The idea is to discover and classify the anomalous smart contract paths. Once this model predicts whether the input is anomalous or not, if the prediction is anomalous, the received call is not carried out. Otherwise, if the prediction is non-anomalous, the received call is executed, modifying the state of the defended smart contract. The life of this detector consists of two steps: training and prediction. In the first one, a part of the path dataset obtained in the previous phase is used to train the deep learning model. In the second one, the remained part of the dataset is used to test the classifier and evaluate the received path detecting anomalous activities.

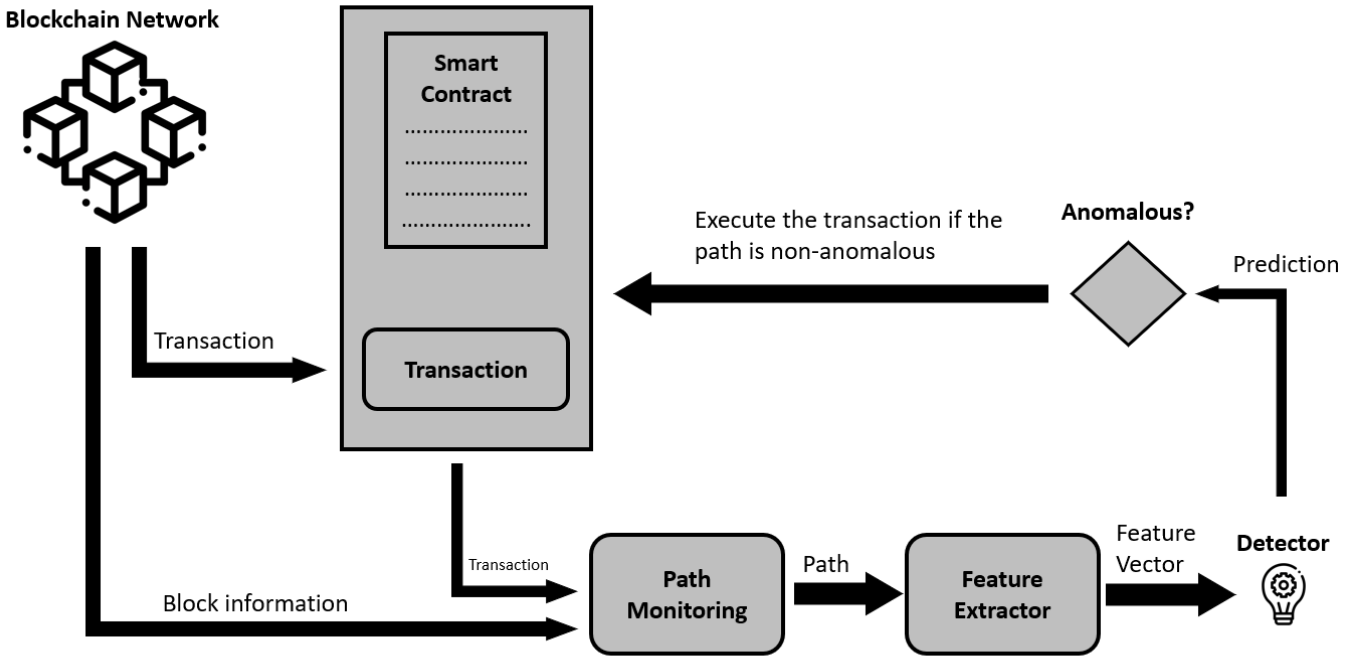


Fig. 1: Proposed IDS for smart contracts

III. Validation

A. Experiment

In this work, Contract Calls dataset from XBlock[1] platform is used for the train phase of the *Detector*. As described in [12], this dataset is an Ethereum dataset, which contains on-chain information about the smart contract activities. The Contract Calls is formed by historical activities of Ethereum smart contracts and each activity (row) is composed by 11 columns (Tab. II). The Contract Calls dataset contains attacks, which Ethereum suffered, such as mentioned DAO attack. Those attacks are represented in *isError* column. In particular, in this study, the generated paths have been labeled using this attribute. The paths are labelled as 0 if they contain at most one error or 1 if they contains two or more errors.

TABLE II: The dataset of the Contract Calls

Columns	First row
BlockNumber	6000000
Timestamp	1532118564
TransactionHash	0xea55aae509250286c601c22b5b6090f5c985b1d11d3cdc6ea00b919bc4641ad2
FromAddress	0xfbb1b73c4f0bda4f67dca266ce6ef42f520fbb98
ToAddress	0x88518ddba475934f1c35e4c03a126c851cd4b566
FromIsContract	0
ToIsContract	1
CallType	call
CallingFunction	0x6ea056a9
Value	0
IsError	None

On this experiment the first 10000000 lines of the dataset have been used. The experiment started generating the desired paths traversing the dataset lines. Those generated paths have different characteristics which were extracted. Once these characteristics were extracted for each path, the was divided in two groups, training data and test data. Those sets were generated randomly and both contained anomalous and no anomalous paths.

With this pre-processing of the data completed, work began on the detector. The model was trained with the training data and then evaluated with the test data. Therefore, the developed IDS was obtained, generating the first automated intrusion detector system for smart contracts, as far as we know.

B. Results

Taking the training data, the set used for the training phase, the model gave a result of 79.30% accuracy in the detection of anomalous and non-anomalous paths. However, when tested with the test data, the detector gave an accuracy of 70.20%. For more complete results, we will work with the entire dataset, taking into account the largest number of paths generated in the Ethereum history. Thus obtaining a more reliable model and a model that knows more types of paths, which could be unknown types of paths for the developed detector.

IV. Conclusions

In this work, an automatic intrusion detection system has been developed, which with historical data from the Ethereum blockchain obtains an accuracy of 0.70 in detecting anomalous paths. This demonstrates the possibility of implementing an intrusion detection system to defend smart contracts and thus deploy them defended. Future work can investigate more specialized detectors to improve detector accuracy. In addition, the selection of features made to characterize a path to be studied will be important to best represent a transaction path. It will also be possible to perform tests with the complete dataset to better generalize the learning of the intrusion

[1] http://xblock.pro/ethereum/

detection system detector in addition to including datasets from other blockchains that have smart contract functionality. This will enrich the results of this type of IDSs and improve the reliability of the results.

ACKNOWLEDGEMENT

This work has been partially supported by the Basque Country Government under the ELKARTEK program, project TRUSTIND (KK-2020/00054).

REFERENCES

[1] N. Atzei, M. Bartoletti, and T. Cimoli, "A survey of attacks on ethereum smart contracts (sok)," 03 2017, pp. 164–186.

[2] X. Wang, J. He, Z. Xie, G. Zhao, and S.-C. Cheung, "Contractguard: Defend ethereum smart contracts with embedded intrusion detection," *IEEE Transactions on Services Computing*, vol. 13, no. 2, pp. 314–328, 2019.

[3] J. Chen, X. Xia, D. Lo, J. Grundy, and X. Yang, "Maintaining smart contracts on ethereum: Issues, techniques, and future challenges," *arXiv preprint arXiv:2007.00286*, 2020.

[4] J. Lee, "Dappguard : Active monitoring and defense for solidity smart contracts," 2017.

[5] G. Ayoade, E. Bauman, L. Khan, and K. Hamlen, "Smart contract defense through bytecode rewriting," in *2019 IEEE International Conference on Blockchain (Blockchain)*. IEEE, 2019, pp. 384–389.

[6] S. Azzopardi, J. Ellul, and G. J. Pace, "Monitoring smart contracts: Contractlarva and open challenges beyond," in *International Conference on Runtime Verification*. Springer, 2018, pp. 113–137.

[7] F. Zhang, E. Cecchetti, K. Croman, A. Juels, and E. Shi, "Town crier: An authenticated data feed for smart contracts," in *Proceedings of the 2016 aCM sIGSAC conference on computer and communications security*, 2016, pp. 270–282.

[8] Z. Wang, W. Dai, K.-K. R. Choo, H. Jin, and D. Zou, "Fsfc: An input filter-based secure framework for smart contract," *Journal of Network and Computer Applications*, vol. 154, p. 102530, 2020.

[9] C. Ferreira Torres, M. Baden, R. Norvill, B. B. Fiz Pontiveros, H. Jonker, and S. Mauw, "Ægis: Shielding vulnerable smart contracts against attacks," in *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, 2020, pp. 584–597.

[10] H. Wang, Y. Li, S.-W. Lin, L. Ma, and Y. Liu, "Vultron: catching vulnerable smart contracts once and for all," in *2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*. IEEE, 2019, pp. 1–4.

[11] M. Rodler, W. Li, G. O. Karame, and L. Davi, "Sereum: Protecting existing smart contracts against re-entrancy attacks," *arXiv preprint arXiv:1812.05934*, 2018.

[12] P. Zheng, Z. Zheng, J. Wu, and H. N. Dai, "Xblock-eth: Extracting and exploring blockchain data from ethereum," *IEEE Open Journal of the Computer Society*, vol. 1, pp. 95–106, 2020.

http://doi.org/10.18239/jornadas_2021.34.08

Un resumen de: Un sistema de detección de intrusiones enfocado en el preprocesamiento de características de red para sistemas IoT

Xavier A. Larriva-Novo, Víctor A. Villagrá , Mario Vega-Barbas, Diego Rivera, Mario Sanz
Universidad Politécnica de Madrid (UPM). DIT, ETSI Telecomunicaciones. Avda. Complutense, 30. 28040 Madrid
{xavier.larriva.novo,victor.villagra,mario.vega,diego.rivera,mario.sanz}@upm.es

Resumen- **La seguridad en las redes de IoT es actualmente un aspecto crítico, dado la creciente adopción de estas tecnologías. Este artículo propone el estudio y evaluación de varias técnicas de preprocesamiento para un algoritmo de aprendizaje automático basadas en la categorización del tráfico. Esta investigación utiliza para su evaluación dos conjuntos de datos de referencia, UGR16 y UNSW-NB15, y uno de los conjuntos de datos más utilizados, NSL-KDD. Todos ellos se procesaron a través de diferentes conjuntos de características basadas en una categorización basada en cuatro grupos de características: características de conexión básica, características de contenido, características estadísticas y, finalmente, un grupo compuesto por características basadas en tráfico y características basadas en dirección. Esta propuesta demuestra que, aplicando la categorización del tráfico de red y varias técnicas de preprocesamiento, la precisión puede mejorar hasta en un 45%.**

Index Terms- Internet of Things, Detección de Intrusiones, Categorización de trafico

Tipo de contribución: *An IoT-Focused Intrusion Detection System Approach Based on Preprocessing Characterization for Cybersecurity Datasets, Investigación publicada en Sensors (2021)*

I. Introducción

Dentro del término Internet de las cosas (IoT) se incluyen diferentes dispositivos, aplicaciones y servicios vinculados al ciberespacio. Según el proceso de recopilación, transmisión y procesamiento de información de los sistemas de IoT, se estructuran en una arquitectura basada en entidades que se divide en tres capas principales: capa de percepción del terminal, capa de transporte de red y capa de servicio de aplicación [1].

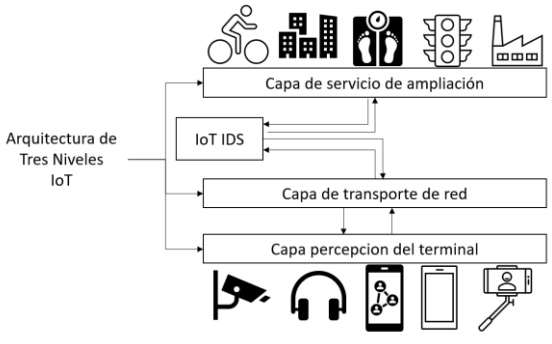


Fig. 1. Arquitectura de tres niveles IoT.

La capa de percepción del terminal está compuesta por la fuente de recopilación de datos de IoT. Las unidades involucradas en esta capa incluyen entidades físicas que representan unidades de dispositivos reales. La capa de transporte de red transmite la información recopilada por la capa de percepción a la capa de servicio de la aplicación. Finalmente, la capa de servicios de aplicaciones procesa los datos transmitidos desde la capa de transporte de red a diversas entidades, proporcionando servicios para diferentes usuarios en diferentes ámbitos, como redes inteligentes, hogares inteligentes y ciudades inteligentes [1].

Para detectar ataques de IoT en la capa de transporte de red, se han implementado sistemas de detección de intrusiones en redes (NIDS) como una segunda línea dec defensa después de los cortafuegos, antivirus y sistemas de control de acceso [2] para dispositivos inteligentes conectados.

Las técnicas de Machine Learning, específicamente los algoritmos de deep learning (DL), se posicionan como una solución eficaz para la detección de anomalías. Uno de los requisitos más importantes de los NIDS basados en técnicas de DL es la fase de preprocesamiento, que puede afectar la precisión de un algoritmo de manera significativa[3].

Este artículo de resumen hace referencia al presentado en [4] así como sus funciones de preprocesamiento. En el cual se toma en cuenta que los sistemas IoT intentan reducir al máximo el coste computacional, potenciando el modelo de aprendizaje y evitando el posible sobreajuste, se propone el uso de la categorización definida en [5] para incrementar la eficiencia de los modelos de aprendizaje. Esta categorización está compuesta por cuatro grupos de características que incluyen características básicas de conexión, de contenido, estadísticas y, finalmente, un grupo que está compuesto por características basadas en el tráfico y características del tráfico basadas en la dirección de la conexión. El objetivo de esta investigación es evaluar esta categorización mediante el uso de diversas técnicas de preprocesamiento de datos basadas en la transformación de valores categóricos en valores numéricos y mediante la aplicación de estandarización y normalización.

En el presente artículo se muestra la propuesta de la arquitectura propuesta para la aplicación del perceptrón multicapa (MLP) basado en [5] y los resultados obtenidos tras aplicar nuestra propuesta a los conjuntos de datos UNSW-NB15, UGR16 y NSL-KDD.

Tabla I
COMPARACIÓN DE LOS MEJORES MODELOS OBTENIDOS CON LAS TÉCNICAS DE PREPROCESAMIENTO PROPUESTAS

Conf	Dataset	Características Básicas	Características Contenido	Características Flujo	Características basadas en dirección	Accuracy
N01	NSL-KDD	z-score	min_max_0	z-score	-	0.997
N02	NSL-KDD	min_max_0	min_max_0	z-score	-	0.978
N03	NSL-KDD	z-score	z-score	z-score	-	0.978
N04	NSL-KDD	-	-	-	-	0.95
N05	UGR16	z-score	-	-	-	0.993
N06	UGR16					0.87
N07	UNSW-NB15	z-score	z-score	z-score	min_max_0	0.992
N08	UNSW-NB15	min_max_0	min_max_0	z-score	min_max_0	0.990
N09	UNSW-NB15	z-score	z-score	z-score	z-score	0.983
N10	UNSW-NB15	-	-	-	-	0.55

II. ARQUITECTURA Y DISEÑO

Esta propuesta se diseña para evaluar el conjunto de características propuestas en [5]. La arquitectura propuesta para el IDS consiste en la introducción de técnicas de preprocesamiento individuales basadas en una caracterización del contenido. El sistema se presenta en la Figura 2. Como se puede observar en la figura, está compuesto por algunas fases para la evaluación con el fin de obtener la mayor precisión.

A. Grupos de características bajo estudio

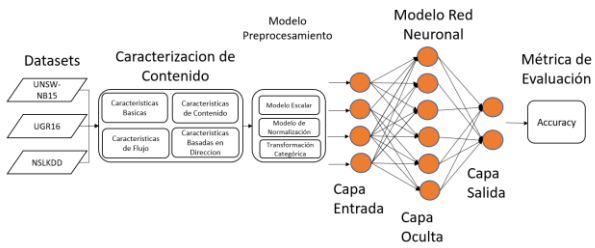


Fig. 2. Arquitectura del modelo propuesto

El algoritmo de red neuronal propuesto para este caso de estudio no permite el uso de variables de entrada de tipo texto, por lo que estas variables se transforman en vectores binarios mediante el método de codificación one-hot [36] y los ataques se transforman en vectores binarios.

Para el caso del dataset UGR16 las variables categóricas se componen de las siguientes:

- *protocol, flag*

Para el caso del dataset NSL-KDD las variables categóricas se componen de las siguientes:

- *protol_type, service, flag, land, num_failed_login, is_host_login, is_guest_login*

Para el caso del dataset UNSW-NB15 las variables categóricas se componen de las siguientes:

- *dur, proto, service*

Los resultados obtenidos del entrenamiento sin preprocesamiento muestran que el conjunto de datos NSL-KDD ofrece una mayor precisión, con un 95,5%, en comparación con el 87,68% del UGR16 y el 55,80% del UNSW-NB15. Cabe señalar que la arquitectura de la red neuronal es la misma para todos los conjuntos de datos.

B. Evaluación individual del conjunto de características

El objetivo de esta investigación propone el uso de la categorización definida en [5] definida por cuatro gripos de características. Además, se aplican tres técnicas de preprocesamiento individuales para cada conjunto individual de características, con el fin de comparar los resultados. Específicamente, se aplicó la técnica de estandarización (z-score), normalización (min-max) y sin preprocesamiento (-).

En el caso del dataset UGR16, dado que solo tiene 12 características básicas, se realizaron tres evaluaciones para cada una de las técnicas mencionadas anteriormente.

En el caso de los datasets NSL-KDD y UNSW-NB15, se tomó en consideración únicamente aquellas características disponibles. En la Tabla 1 se muestran las 10 principales variaciones de precisión con los resultados más favorables obtenidos, ordenados en orden descendente aplicando las variaciones de las técnicas de preprocesamiento de datos para cada uno de los grupos de características del conjunto de datos. En el presente .

III. CONCLUSIONES

Este trabajo de investigación presenta nuevas comparaciones en el uso de varios datasets como UNSW-NB15 y UGR16, permitiendo ampliar los diversos estudios en la aplicación de algoritmos ML para el caso de NIDS basados en anomalías. El estudio muestra la relevancia de la tarea de preprocesamiento de datos para que el algoritmo ML obtenga una mayor precisión

La conclusión más relevante que aporta este estudio es la importancia de las características de preprocesamiento, como las características básicas y las estadísticas de tráfico mediante técnicas de estandarización z-score, lo que permite incrementar la precisión ya que permite utilizar la desviación media de las variables. Se propone igualmente ampliar la investigación y aplicar el modelo propuesto a un entorno real, con datos reales recogidos de sistemas IoT, como la plataforma de servicios Smart City propuesta en [1], para demostrar la eficiencia de nuestra implementación.

REFERENCIAS

[1] H. Wu, H. Han, X. Wang, y S. Sun, «Research on Artificial Intelligence Enhancing Internet of Things Security: A Survey», *IEEE Access*, vol. 8, pp. 153826-153848, 2020, doi: 10.1109/ACCESS.2020.3018170.

[2] M. AL-Hawawreh, N. Moustafa, y E. Sitnikova, «Identification of malicious activities in industrial internet of things based on deep learning models», *Journal of Information Security and Applications*, vol. 41, pp. 1-11, ago. 2018, doi: 10.1016/j.jisa.2018.05.002.

[3] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, y P. Faruki, «Network Intrusion Detection for IoT Security Based on Learning Techniques», *IEEE Communications Surveys Tutorials*, vol. 21, n.º 3, pp. 2671-2701, thirdquarter 2019, doi: 10.1109/COMST.2019.2896380.

[4] X. Larriva-Novo, V. A. Villagrá, M. Vega-Barbas, D. Rivera, y M. Sanz Rodrigo, «An IoT-Focused Intrusion Detection System Approach Based on Preprocessing Characterization for Cybersecurity Datasets», *Sensors*, vol. 21, n.º 2, Art. n.º 2, ene. 2021, doi: 10.3390/s21020656.

[5] X. A. Larriva-Novo, M. Vega-Barbas, V. A. Villagra, y M. Sanz Rodrigo, «Evaluation of Cybersecurity Data Set Characteristics for Their Applicability to Neural Networks Algorithms Detecting Cybersecurity Anomalies», *IEEE Access*, vol. 8, pp. 9005-9014, 2020, doi: 10.1109/ACCESS.2019.2963407.

Sesión de Investigación A2:
Detección de intrusiones y gestión de anomalías II

http://doi.org/10.18239/jornadas_2021.34.09

Sobre las capacidades de detección de los IDS basados en firmas

J. Diaz-Verdejo*, F. J. Muñoz†, R. Estepa Alonso†, A. Estepa Alonso†, G. Madinabeitia†
* Dpt. Teoría de Señal, Telemática y Comunicaciones, CITIC, Univ. de Granada
C/ Periodista Daniel Saucedo Aranda, s/n, 18071 Granada (Spain)
ORCID: 0000-0002-8424-9932; E-mail: jedv@ugr.es
† Dpt. Ingeniería Telemática, Escuela Superior de Ingenieros, Univ. de Sevilla
C/ Camino de los Descubrimientos s/n, 41092 Sevilla (Spain)
ORCID: 0000-0001-8146-8438, 0000-0001-8505-1920, 0000-0003-1841-3973, 0000-0001-6376-4620
E-mail: {jvt,rafa,aestepa,german}@trajano.us.es

***Resumen*—Los sistemas de detección de intrusiones (IDS) pueden detectar actividades maliciosas y generar alertas a supervisar, por lo que constituyen el núcleo de los sistemas de monitorización de la seguridad de las redes. Tradicionalmente, se ha asumido que los IDS basados en firmas (SIDS) ofrecen una capacidad de detección y tasa de falsos positivos adecuadas, presentando limitaciones sólo en la detección de ataques 0-day. Sin embargo, estas capacidades están inequívocamente asociadas a la calidad de las firmas disponibles, que varían no sólo en el tiempo sino con la herramienta concreta utilizada. En este trabajo se exploran las capacidades de diversos sistemas SIDS ampliamente utilizados en un escenario real en el contexto de servicios web. Asimismo, se analiza la evolución de sus prestaciones a lo largo del tiempo considerando la actualización de las firmas. Los resultados de nuestras pruebas evidencian una gran variabilidad en las prestaciones en función de la herramienta seleccionada, así como una deficiente cobertura de ataques conocidos, incluso cuando se optimizan las reglas para ajustarse al sistema a proteger. Consecuentemente, es necesario revisar el papel de los SIDS como elementos de protección, ya que pueden proporcionar una falsa sensación de seguridad.**

***Index Terms*—Cybersecurity, Intrusion Detection, Signature-based IDS, Rules tunning**

Tipo de contribución: *Investigación original*

I. Introducción

Los sistemas de detección de intrusos (*Intrusion Detection Systems*, IDS, por sus siglas en inglés) [1] se consideran un elemento fundamental en la protección y monitorización de los sistemas y redes, constituyendo unos de los elementos clave del despliegue de la seguridad en capas. En particular, los sistemas de monitorización de la seguridad en redes (NSM, del inglés *Network Security Monitoring*) [2] incorporan uno o varios IDS como elementos generadores de alertas de incidentes. En ese sentido, la detección de ciberataques depende, de manera crítica, de las capacidades de los diferentes IDS desplegados, ya que el análisis de los incidentes se inicia a partir de las alertas existentes. Consecuentemente, resulta muy relevante que los IDS presenten una alta capacidad de detección (TP) y, a su vez, una muy baja tasa de falsos positivos (FP), esto es, que sean capaces de generar alertas para todos los eventos intrusivos, pero que no las generen para los eventos legítimos.

Los IDS se pueden agrupar en dos categorías principales en función de la forma en la que se realiza la detección [1]. Así, se distinguen los IDS basados en firmas (SIDS, *signature-based IDS*) y los basados en anomalías (AIDS, *anomaly-based IDS*). En los primeros se utilizan *firmas* o patrones extraídos de ataques conocidos mientras que en los segundos se establecen modelos de actividad que determinan la naturaleza de los eventos. Habitualmente se argumenta que los SIDS presentan altas tasas de detección para ataques conocidos con muy bajas tasas de FP, lo que los hace adecuados para su despliegue en sistemas en explotación. De hecho, la mayoría de los IDS en uso en escenarios reales son de este tipo. La principal desventaja de los SIDS es su teórica incapacidad de detectar ataques nuevos, esto es, ataques de día cero (*0-day*) que, a la postre, representan la mayor amenaza a la ciberseguridad precisamente por su carácter novedoso (se estima que en torno al 30 % de los ataques son de este tipo). Este escenario motiva la utilización de detectores basados en anomalías, teóricamente capaces de detectar este tipo de ataques, pero con unas prestaciones usualmente inadecuadas en relación a los FP. El resultado es un gran volumen de actividad investigadora en relación a los AIDS, mientras que apenas se desarrollan trabajos nuevos sobre SIDS[1].

Sin embargo, la supuesta incapacidad de detección de ataques de día cero es sólo parcial. Algunos estudios [3] apuntan a que el porcentaje de ataques de día cero detectados por los SIDS, aunque insuficiente, no es despreciable (en torno al 10 %). Por el contrario, este mismo estudio y resultados previos de los investigadores [4] evidencian una tasa de detección, incluso para ataques conocidos con suficiente antelación, claramente mejorable (54 % en [3] y 84 % en [4]) y que puede complementarse adecuadamente con técnicas basadas en anomalías. Por otra parte, el análisis de las alertas generadas por un SIDS usando reglas ajustadas al escenario de aplicación [5] revela la aparición de un número de FP similar al de TP, por lo que en torno al 50 % de las alertas generadas corresponden a actividad no intrusiva. Estos resultados nos llevan a plantearnos el estudio de la efectividad de los SIDS en escenarios reales, tanto en relación a sus capacidades de detección como a la generación de falsos positivos. Dada la experiencia previa del equipo y la disponibilidad de datos, elegimos un escenario de monitorización de ataques basados en URI de HTTP, idéntico al considerado en [5].

Debido a la presumible dependencia de los resultados con el conjunto de reglas y el detector seleccionado, se plantean

[1]En torno a 350.000 publicaciones sobre AIDS frente a 150.000 para SIDS en *google scholar*, de las cuales 150.000 y 2.000 corresponden, respectivamente, a los últimos 10 años.

varias dimensiones y cuestiones en este estudio. Así, el presente trabajo se articula en torno a tres ejes:

- *Capacidades de SIDS específicos*. Existen múltiples sistemas SIDS, cada uno con sus conjuntos de reglas específicos. ¿Tienen todos prestaciones similares? ¿Son complementarios? Si hay ataques que detecta una herramienta pero no otra, ¿tendría sentido combinar las capacidades de detección de varios de ellos?
- *Idoneidad/adaptación de las firmas*. Las reglas de detección son habitualmente desarrolladas para escenarios genéricos con el objetivo primario de posibilitar la detección del mayor número posible de ataques. Esto puede dar lugar a tasas de FP inaceptables, por lo que se suelen establecer diferentes *sensibilidades* en la detección (p.ej., niveles de paranoia, PL, en *modSecurity*) o reglas desactivadas por defecto (como es el caso, p.ej., en las reglas generadas por *Talos*[2] o *Emergency Threats*[3], que son ampliamente utilizadas). ¿Son suficientes las configuraciones por defecto? ¿Se modifican significativamente las prestaciones con el ajuste de las reglas?
- *Evolución temporal de las firmas*. Como hemos mencionado previamente, se asume que los ataques *0-day* no pueden ser detectados adecuadamente por los SIDS hasta que no están incluidos en las firmas. Sin embargo, en algunos escenarios (p.ej. APT) el tiempo medio de detección supera los 3 meses. ¿Cuánto tiempo tardan los ataques *0-day* en ser incorporados en firmas? ¿Son efectivamente incorporados o continúan sin ser detectados? En la misma línea, hay reglas que aparecen/desaparecen por diversos motivos, entre los que podríamos señalar la generación de tasas de FP inadecuadas a partir de algunas firmas. ¿Cómo evolucionan las capacidades de detección con el tiempo?

La enorme complejidad y multidimensionalidad del problema planteado nos lleva a una aproximación en la que exploraremos preliminarmente las cuestiones planteadas a partir de tres experimentos/contribuciones:

1. Estudio comparativo de las capacidades de detección de varios tres detectores basados en firmas ampliamente utilizados (*Snort/Suricata*, *modSecurity* y *Checkpoint*), evaluando las posibles combinaciones de las alertas generadas por los mismos mediante reglas simples.
2. Análisis del impacto del ajuste de las reglas en las capacidades de detección de Snort/Suricata, usando las reglas oficiales de *Talos*.
3. Estudio de la evolución temporal de las tasas de falsos positivos y verdaderos positivos para Snort/Suricata (usando también las reglas generadas por *Talos*) a medida que evolucionan las firmas.

Para la evaluación de las capacidades de detección se usará tráfico real y varios conjuntos de ataques con diferentes características, lo que permitirá estimar tanto la tasa de detección como la de FP, a partir de la inspección de las alertas generadas. Asimismo, esta aproximación posibilita el ajuste de las reglas de forma que se minimicen los FP sobre el escenario real. La elección de Snort/Suricata para las cuestiones 2 y 3 radica en el fácil manejo de las firmas asociadas y, especialmente, en la existencia de una herramienta propia que permite su aplicación a las trazas del servicio.

El resto de este artículo se estructura como sigue. En el Apartado II se presentarán los antecedentes y líneas de trabajo previamente exploradas en la bibliografía en relación a los 3 ejes planteados. En el Apartado III describiremos brevemente el escenario experimental considerado, describiendo los *datasets* a utilizar, así como las herramientas auxiliares relevantes y los resultados pre-existentes sobre este escenario y que han motivado el presente trabajo. En los Apartados IV, V y VI se procederá a desarrollar y analizar los experimentos previamente mencionados relativos a la comparación de las capacidades de detección de diversos SIDS, el ajuste de las reglas al escenario y la evolución temporal de las capacidades de detección con la evolución de las reglas. Finalmente, en el Apartado VII se presentarán las conclusiones y líneas de desarrollo futuras de la presente propuesta.

II. Antecedentes y trabajos previos

El análisis de las capacidades de detección de diferentes SIDS así como su combinación tiene un claro interés por sí mismo [6] y ha sido estudiado previamente en diversos estudios segmentados (abordan los problemas por separado). Así, en relación al primero de los aspectos, podemos encontrar diversos estudios comparativos con resultados dispares. A modo de ejemplo, [7] analiza *Snort* y *Suricata* en el contexto de redes de uso industrial, obteniendo diferencias significativas. Sin embargo, [8] compara ambos detectores en un escenario real, obteniendo resultados similares en cuanto a capacidad de detección. En cualquier caso, tanto [8] como la mayoría de los trabajos relacionados se centran principalmente en la comparación de los costes computacionales, asumiendo que las capacidades de detección son similares en todos los casos, es decir, ignoran o minimizan la influencia de los conjuntos de reglas usados en los resultados. Sin embargo, es evidente que la calidad de las firmas debe tener un fuerte impacto en las capacidades de detección y que no todos los detectores utilizan los mismos conjuntos. Como ya se ha mencionado, resultados previos de nuestro equipo [5] avalan esta afirmación, mostrando importantes diferencias para dos de las herramientas utilizadas.

Por otra parte, en cuanto a la combinación de varios sistemas IDS independientes, en la literatura se pueden encontrar numerosos trabajos que proponen sistemas tipo votación o multicapa, aunque están mayoritariamente enmarcados en el contexto de AIDS (p.ej. utilizando técnicas de clustering [9], SVM [10] o *ensemble learning* [11] [12] [13]). En el caso concreto de SIDS, podemos encontrar algunos trabajos en esta línea, p.ej. [14], aunque están fundamentalmente orientados a reducir la tasa de FP más que a mejorar las capacidades de detección y normalmente operan en base a la correlación de las alertas [15]. En particular, se asume que una alerta es correcta si un número suficiente de detectores coinciden en ella, lo que evidentemente puede resultar subóptimo, especialmente si los detectores no son genéricos o están ajustados para alguna tipología concreta de ataque. En esta línea, podemos encontrar varios trabajos que plantean sistemas colaborativos o distribuidos para determinar si una alerta es un FP e incluso para seleccionar las firmas a utilizar [16].

[2]https://www.snort.org/talos

[3]https://rules.emergingthreats.net/

Tabla I
CARACTERÍSTICAS DE LOS DATASETS UTILIZADOS.

Dataset	N. URIS	Tipo	Fechas	Fuente	Observaciones
Inves	14151496	Real	2018	Propia [17] [5]	Peticiones GET,POST,HEAD, PROPFIND con código respuesta correcto
ataques-800	832	Ataques	$<$ 2017	Propia [24]	Tipo 3 - CVE incluidos en firmas Snort
ataques-1100	1176	Ataques	$\leq$2017	Propia [24]	Tipos 1 (CVE 2016 y 2017), 3 y 4
Fwaf-bad	48126	Ataques	$<$ 2017	Pública [26]	Ataques WAF
Fwaf-2200	2200	Ataques	$<$ 2017	Pública [26]	Subconjunto con CVE identificados (reglas de Snort)
rdb	934	Ataques	$<$ 2009	Propia [4]	Ataques generados a partir de RDB
osvdb	6897	Ataques	$<$ 2009	Propia [25]	Ataques generados a partir de OSVDB
A-420	420	Ataques	2017-2018	Propia	Generados con CVE de 2017 y 2018 y CVSS$>$9

Esto, en cierta forma, correspondería a un sistema de tipo votación, aunque en algunos casos el detector es siempre el mismo y lo que cambia es su ubicación. Por el contrario, en [17] proponemos combinar *modsecurity* [18] y *Snort* [19] para limpiar el tráfico HTTP simplemente combinando las salidas (operación $\bigcup$), lo que está alineado con el objetivo de limpiar el tráfico (maximizar TP), aunque podría no resultar adecuado para minimizar las tasas de falsos positivos.

En relación a la idoneidad de las firmas para el escenario donde se aplica el SIDS, encontramos pocos trabajos orientados a analizar el impacto de una adecuada selección de las mismas en diferentes escenarios. Varios autores coinciden en la importancia de las reglas y de su calidad, pero lo hacen en contextos genéricos y mayoritariamente obviando los FP. Así, [20] propone una metodología y herramientas para optimizar las reglas orientada fundamentalmente a la evaluación de la calidad de las firmas, la detección de redundancias y la generación de ataques a partir de ellas para su evaluación en escenarios genéricos. Resulta interesante el análisis sobre el impacto que tiene usar reglas estrictas frente a reglas balanceadas, llegando a la conclusión de que es significativo incluso con ataques generados a partir de las propias reglas. Con un objetivo similar [16] propone una métrica para evaluar la calidad de las firmas. Sin embargo, no es un trabajo extensivo y sólo presenta buenas prácticas. Por otra parte, para ello incorpora información de otros sistemas, como analizadores de vulnerabilidades. Ninguno de estos trabajos considera el impacto de los FP, para lo que obviamente se requiere de tráfico real. En esta línea podemos mencionar [21], que plantea una selección colaborativa de las firmas a partir de la realimentación de los operadores. En este contexto, podemos encontrar múltiples trabajos que proponen mecanismos y herramientas para la administración de las reglas, pero que no analizan el impacto de las mismas.

Finalmente, en cuanto a la evolución de las capacidades de detección, el número de trabajos es muy reducido y con alcance realmente limitado. Así, [22] analiza la evolución de las reglas de Snort durante un corto periodo (6 meses), aunque no considera la evaluación de las capacidades asociadas. Por otra parte, [23] propone un AIDS que compara con Snort y Suricada usando tráfico y ataques reales, afirmando que su propuesta detecta ataques 4 semanas antes de su inclusión en las firmas. No hemos encontrado más trabajos en esta línea, lo que puede estar motivado por la dificultad de consecución de conjuntos de firmas datados por la continua actualización de los mismos y la inexistencia de históricos.

III. ESCENARIO EXPERIMENTAL Y HERRAMIENTAS

Las ideas a explorar en este trabajo tienen su antecedente en los resultados obtenidos en [5] y [17] durante el proceso de limpieza de dos *datasets* de tráfico real recolectados en la Universidad de Sevilla. Durante el mismo, se realizó una análisis exhaustivo y supervisado del tráfico utilizando diferentes herramientas y aproximaciones, determinando los falsos positivos generados por los SIDS utilizados. Este dato resulta relevante para la evaluación de las capacidades. Por tanto, el escenario base a considerar será el descrito en [17], cuyas características principales resumimos a continuación.

El tráfico real a considerar es el denominado *Inves*, recolectado de las trazas del servidor web del servicio de investigación de la Universidad de Sevilla (http://fama.us.es) durante el mes de mayo de 2018. Este servidor está principalmente orientado a la gestión de un repositorio de documentos y hace un uso extensivo de parámetros en las URI. Los datos disponibles corresponden a las trazas del servidor Apache completas, habiéndose seleccionado todas las líneas correspondientes a peticiones de interés, esto es, que contienen los métodos GET, POST, HEAD y PROPFIND y cuyo código de respuesta no se corresponde con un error. Las características más relevantes en cuanto a contenido se muestran en la Tabla I. Los resultados previos establecen que mediante las reglas *Talos* (*Snort/Suricata*) se pueden detectar al menos 1384 ataques reales en *Inves* sobre un total de 3093 alertas obtenidas eliminando las reglas más obviamente generadoras de FP.

En relación a los ataques a utilizar para evaluar las capacidades de detección, se considerarán varios datasets de ataques, algunos públicamente disponibles [26] y otros generados según diferentes procedimientos por los autores en trabajos previos [4], [24], [25]. Las características más relevantes se muestran en la Tabla I. Todos constan de un listado de URI que constituyen ataques en el momento de su generación/adquisición. En este sentido, resultan especialmente relevantes los conjuntos *A-420*, *ataques-800* y *Fwaf-2200*. El primero por contener ataques críticos con CVE de los años 2017 y 2018, lo que permitirá evaluar las capacidades de detección de ataques recientes (en la fecha de recolección de las trazas) y de día cero. Los otros dos, por contener ataques extraídos de otros conjuntos de ataques a partir del etiquetado del CVE asociado por las propias reglas de los SIDS. Sería de esperar, por tanto, un 100 % de detección para estos conjuntos.

IV. ANÁLISIS DE PRESTACIONES DE SISTEMAS SIDS

El primer estudio tiene como objetivo determinar las capacidades de detección de varios SIDS y, en su caso, determinar las posibles sinergias que pudieran mejorar dichas capacidades mediante la combinación de los mismos. Evidentemente, no

es factible comprobar todos los detectores, por lo que hemos seleccionado algunos de los más utilizados. Se considerarán 3 SIDS: *InspectorLog* [5], una herramienta desarrollada por los investigadores para aplicar las reglas de Snort/Suricata a las trazas de HTTP; *ModSecurity*, un WAF (*Web Application Filter*) ampliamente utilizado; y *CheckpointIPS*, un cortafuegos con IPS comercial de reconocido prestigio. En el caso de *InspectorLog* se utilizarán las reglas de *Talos* que afecten a los URI de HTTP a partir de su extracción de los *rulesets* oficiales mediante herramientas desarrolladas al efecto [5]. Para las otras herramientas, los conjuntos de reglas a utilizar son los vigentes a fecha de agosto de 2018, por ser ese el conjunto disponible para la herramienta comercial en el momento de recolección de las trazas.

Utilizaremos las trazas de tráfico real (*Inves*) para determinar tanto el número de ataques presentes en esta base de datos según cada SIDS, así como la tasa de falsos positivos. Para esto último, en el caso de *InspectorLog* utilizaremos como referencia [17], donde se determinan los 1384 ataques encontrados en dicha traza tras la inspección manual de las URI consideradas como ataques por *InspectorLog*. En el resto de SIDS se realizará por inspección manual de los ataques detectados. Las trazas que contienen sólo ataques serán utilizadas para evaluar la capacidad de detección (tasa de TP) a partir del análisis de las alertas obtenidas por los distintos SIDS.

Tanto *modSecurity* como *InspectorLog* permiten ajustar la sensibilidad de la detección, por lo que se evaluarán varias configuraciones. En el caso de *modSecurity* la sensibilidad se establece mediante el denominado nivel de paranoia (PL, *paranoia level*), que puede tomar valores entre 1 y 4, siendo 4 el más sensible. Los experimentos preliminares incluidos en [5] permitieron determinar que niveles de PL superiores a 2 eran inaceptables por el elevado número de FP, por lo que únicamente evaluaremos los niveles 1 y 2. En el caso de *InspectorLog* consideraremos las reglas activas por defecto (menor sensibilidad), In(D), la activación de todas las reglas (mayor sensibilidad posible), In(A), y las reglas utilizadas en el trabajo [17], In(DET), que permitieron determinar las URI consideradas ataques reales (1384) detectables y, consecuentemente, el de FP, en el caso de *Inves*.

Los resultados experimentales obtenidos se detallan en la Tabla II. En esta tabla se muestran las 6 configuraciones analizadas: In(D), In(A), In(DET) para *Inspectorlog*; M(PL1) y M(PL2), para *modSecurity* con los niveles de paranoia 1 y 2 respectivamente; y CP para *CheckpointIPS* en su configuración *Strict*.

Como era previsible, los resultados de detección dependen significativamente del nivel de sensibilidad, lo que hace imprescindible ajustar la misma al escenario analizado. En el caso de *Inves* y las reglas *Talos* podemos encontrar que, para un total de 1384 ataques efectivamente detectados [17], el número de FP varía entre 0 y 77769, aunque sólo se detectarían 57 de ellos si se selecciona la configuración por defecto (el punto de operación con 0 FP). Es evidente que es necesario encontrar un compromiso entre ambos parámetros a partir del ajuste de las reglas, como podría ser el caso de la configuración In(DET), que, detectando el mayor número posible de ataques tan sólo presenta 1709 FP. Cabe señalar que

Tabla II
RESULTADOS DE DETECCIÓN.

	Configuración					
Dataset	In(D)	In(A)	In(DET)	M(PL1)	M(PL2)	CP
	N. URI ataque					
Inves	57	79154	3093	2299	23820	96
	N. FP					
Inves	0	77769	1709	1431	22106	4
(FPR %)	0	0,55	0,51	0,007	0,0009	0,0002
	% TP					
ataques-800	0,48	92,07	86,06	8,40	41,66	1,92
ataques-1100	0,60	93,71	88,01	5,69	51,91	1,36
Fwaf-bad	0,57	53,77	43,14	1,44	20,43	0,10
Fwaf-2200	2,55	91,23	85,36	3,05	39,23	3,32
rdb	0,11	91,01	86,94	10,29	39,23	3,32
osvdb	5,87	66,99	53,75	2,60	67,42	0,61
A-420	47,14	81,67	76,90	0,95	69,05	7,14
Promedio (ataques)	8,19	81,49	74,31	4,63	46,99	2,54

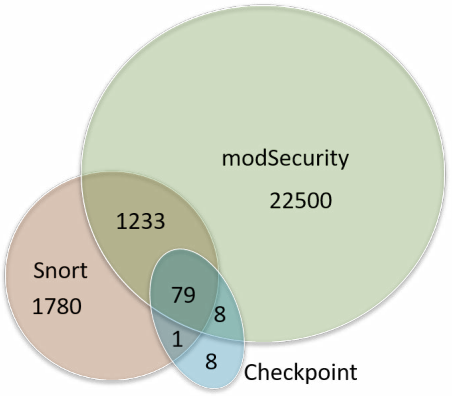


Figura 1. Relaciones entre alertas generadas (no a escala).

esta misma configuración obtiene capacidades de detección de los dataset de ataques muy similares a la configuración con todas las reglas activas –In(A)–.

Sin embargo, el resultado más relevante desde nuestro punto de vista es la diferente capacidad de detección de los tres detectores que, además, resulta insuficiente en muchos casos cuando se presentan las baterías de ataques. Como se puede comprobar en la Tabla II, las tasas de detección pueden llegar a ser realmente pobres en algunas configuraciones y en ningún caso superan el 94 %. Se observa también una gran variabilidad con el SIDS/reglas utilizados. Estos resultados dejan en evidencia a los SIDS y su supuesta capacidad para detectar todos los ataques conocidos. Además, esta capacidad máxima de detección correspondería a un ajuste que resultaría totalmente inadecuado por el elevado número de FP.

A continuación se procede a comparar las alertas generadas por los 3 SIDS en la mejor configuración disponible, lo que corresponde a la configuración PL2 para *modSecurity* y a In(DET) en el caso de las reglas *Talos*. La Figura 1 muestra las alertas detectadas de forma aislada por cada detector, así como las detectadas por varios a la vez. Como se observa en la figura, existen discrepancias significativas entre los URI clasificados como ataques por los diferentes sistemas. Sólo 79 ataques son detectados por los tres SIDS empleados. En el caso de *InspectorLog* se detectan 3093 ataques, de los que que 1312 son también detectados por *modSecurity* y 80 por CP. De los 23820 ataques detectados por *modSecurity* tan sólo

Tabla III
RESULTADOS DE COMBINAR LOS DETECTORES.

Operación	TP	FP	N alertas
In ∩ MS ∩ CP	79	0	79
In ∪ MS ∪ CP	1796	23813	25609
Votación	1315	6	1321
In ∩ MS	1307	5	1312
In ∩ CP	80	0	80
MS ∩ CP	86	1	87

comparte 87 con CP. En el caso de CP, de los 96 ataques detectados 8 de ellos no han sido detectados por ningún otro SIDS.

Análogamente al proceso realizado en [5], se clasifican las alertas generadas en TP y FP mediante inspección manual. Así, se ha establecido que existen al menos 1796 ataques (TP), cantidad significativamente superior a la que proporciona cualquiera de los detectores por separado. Por otra parte, el análisis de los falsos positivos revela que ninguno ha generado alertas en los tres sistemas a la vez. De los 1709 falsos positivos de *InspectorLog*, tan sólo 5 han generado alertas en *modSecurity*, mientras que de los 22106 falsos positivos detectados por *modSecurity* sólo 1 ha sido detectado por CP.

Estos resultados sugieren que los falsos positivos tienden a ser específicos del IDS, mientras que las detecciones correctas presentan más coincidencias, lo que implicaría que una combinación de detectores podría mejorar los resultados, no sólo en cuanto a la mejora en la detección de ataques sino también en relación a la generación de FP.

En consecuencia, consideramos operaciones simples para combinar las alertas generadas por los tres detectores, obteniéndose los resultados mostrados en la Tabla III. Como se puede observar, la intersección de los tres detectores merma mucho la capacidad de detección debido a la baja capacidad que presenta CP, aunque proporciona el menor número de FP. Por el contrario, la unión de los mismos dispara el número de falsos positivos en una proporción muy superior a la mejora obtenida en capacidad de detección, que sería la máxima en este caso. En este escenario, los resultados más equilibrados se obtienen para la intersección de In y MS y para la votación, ya que permiten reducir la tasa de FP aunque con una reducción en los ataques detectados. Como se evidencia en la Tabla III y se ha indicado previamente, la combinación de los detectores permitiría detectar hasta 1796 ataques (TP), lo que supone una mejora relevante. Sin embargo, ninguna de las operaciones simples analizadas parece proporcionar un punto de operación adecuado. En este sentido, consideramos que el resultado está fuertemente influenciado por el claramente diferente comportamiento de los 3 detectores, siendo *Snort/Suricata* el más equilibrado, *CheckpointIPS* el más estricto para generar alertas y *modSecurity* el que genera mayor número de FP. Consideramos, por tanto, que es necesario avanzar en el estudio de mecanismos que permitan combinar las capacidades de detección teniendo en cuenta las características de cada uno de ellos.

V. IMPACTO DEL AJUSTE DE LAS REGLAS

A continuación, procedemos a estudiar con mayor detalle cómo la optimización de las reglas para un escenario considerado influye sobre las capacidades de detección y la tasa de generación de FP. Los resultados del apartado anterior apuntan a que este ajuste desempeña un papel relevante. Para esta experimentación únicamente consideraremos las reglas *Talos*, que aplicaremos con *InspectorLog*, en diferentes configuraciones que denominaremos *modos* (la Tabla IV detalla el número de reglas incluidas en cada uno de los 5 modos definidos).

Tabla IV
MODOS UTILIZADOS PARA OPTIMIZAR LAS REGLAS

Modo	M0	M1	M2	M3	M4	M5
N. reglas	133	2343	2205	2201	2194	2189

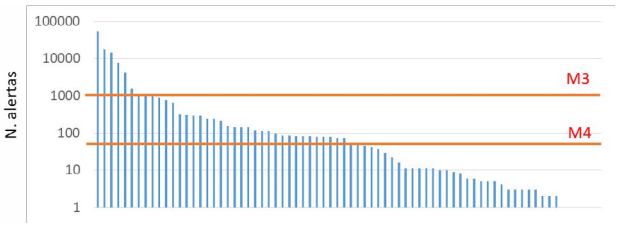


Figura 2. Histograma de alertas por SID.

En primer lugar, tras utilizar únicamente las reglas activas por defecto (*modo 0*), que genera unos resultados claramente insuficientes, procedemos a activar todas las reglas (*modo 1*). Estas dos configuraciones corresponden con las utilizadas en el apartado anterior en los experimentos *In(D)* y *In(A)*. Evidentemente, el modo *M1* será el que proporcionará un mayor número de alertas y una mayor capacidad de detección de ataques, pero también el mayor número de FP. En la Tabla V se muestran las 10 alertas (SID) más frecuentes en este modo. Como se puede observar, algunas de ellas presentan un número muy elevado de apariciones, por lo que podrían corresponder a FP o a la activación de reglas inadecuadas para el escenario (p.ej., la detección de actividad de robots). De esta forma, parece aconsejable una selección de las reglas activas. Consideraremos *Inves* para el ajuste de las reglas, por disponer del análisis de FP.

Para ir ajustando y reduciendo el número de reglas procedemos a eliminar las que se consideran inadecuadas a partir de, primero, la eliminación de las marcadas como *DELETED*[4] por el propio equipo *Talos*, lo que da lugar al conjunto de reglas que llamaremos *M2*. A continuación, analizamos las alertas obtenidas y su distribución por código SID (*Signature Identifier*) (Figura 2) e iremos eliminando reglas progresivamente dando lugar a las configuraciones *M3* a *M5*, siendo el modo *M5* el que menos reglas contiene (modo más optimizado). Para el modo *M3* eliminamos todas las reglas que generan más de 1000 alertas, mientras que para el modo *M4* eliminamos todas las que generan más de 50 alertas. Finalmente, en el modo *M5* eliminamos todas las reglas que únicamente generan FP a partir del análisis manual de las alertas generadas. Hemos de aclarar, en relación a la Figura 2, que la reducción de M3 a M4 (líneas rojas en la figura) parte de la eliminación previa de las reglas DELETED, que se encuentran incluidas en el histograma, pero que, obviamente, no deben ser contabilizadas.

[4]Reglas incorporadas en algún momento en la distribución oficial pero que a posteriori se consideran inadecuadas por generar un número excesivo de FP.

Tabla V
REGLAS QUE GENERAN MAYOR NÚMERO DE ALERTAS (*M1*).

SID	Frec.	Desc
1852	53772	SERVER-WEBAPP robots.txt access
41742	17352	POLICY-OTHER external admin access attempt
17410	14521	OS-WINDOWS Generic HyperLink buffer overflow attempt
873	7621	DELETED WEB-CGI scriptalias access
21515	4270	SERVER-APACHE Apache Tomcat Web Application Manager access
23362	1546	[SERVER-IIS tilde character file name discovery attempt
1113	1002	DELETED WEB-MISC http directory traversal
38336	1002	DELETED SERVER-WEBAPP possible directory traversal attempt
1122	960	SERVER-WEBAPP /etc/passwd file access attempt
13990	885	SQL union select - possible sql injection attempt - GET parameter

Tabla VI
URI CLASIFICADOS COMO ATAQUES EN LOS DIFERENTES MODOS.

Dataset	M0	M1	M2	M3	M4	M5	$\Delta(\%)$
Inves	57	79154	73002	2386	1510	1413	1,78
ataques-800	3	765	723	721	713	711	6,61
ataques-1100	5	1102	1051	1045	1035	1033	5,87
Fwaf-bad	275	25879	23148	22668	22064	22040	7,98
Fwaf-2200	56	2007	1908	1886	1883	1883	5,64
rdb	1	850	824	818	807	804	4,93
osvdb	405	4620	4027	3787	3719	3704	13,28
A-420	198	343	341	335	324	322	5,00
Total ataques	924	35567	32024	31260	30545	30497	8,37

Los resultados de detección en función del modo se muestran en la Tabla VI. En dicha tabla, en la columna $\Delta(\%)$ se muestra el porcentaje de URI etiquetados como ataque en el modo más optimizado (*M5*) respecto del menos (*M1*) para el caso de los dataset de tráfico real, y la variación en la tasa de detección para los datasets de ataque (i.e., la diferencia entre el porcentaje de ataques detectados con *M5* y *M1* respecto a los indicados en la Tabla I). Se observa que, incluso en el modo optimizado (*M5*) se generan un reducido número de FP (1413 URI detectados frente a 1384 TP). También resulta relevante que la mayor parte de los FP que han sido eliminados durante la optimización, una vez eliminadas las reglas *DELETED*, proceden de un reducido número de reglas (16, véase la Tabla VII), de las cuales 11 pueden ser directamente identificadas por el volumen de alertas que generan. Asimismo, a partir de *M2*, hemos comprobado que se siguen detectando los mismos ataques reales (TP), en algunos casos por activarse otras reglas para los mismos. Es decir, no se ha reducido la capacidad de detección para el tráfico real. Como resultado, se observa una reducción muy significativa en el número de alertas generadas tras la optimización (sólo el 1,78 % de las originales para *Inves*) sin pérdida de capacidad de detección.

Adicionalmente, se constata que el impacto de la optimización final a partir del análisis detallado de todas las alertas (de *M4* a *M5*) es mínimo y podría obviarse sin afectar significativamente a las tasas de FP. Esto supone una simplificación del procedimiento a aplicar, ya que únicamente habría que eliminar las SID que generen un cierto volumen de alertas, para lo que bastaría la utilización de histogramas como los mostrados en la Figura 2. Esto resultaría consistente con la aproximación utilizada en algunos sistemas no supervisados que asumen que el número de ataques es porcentualmente insignificante en el tráfico real y, consecuentemente, no debería generar un número elevado de alertas.

Tabla VII
CONJUNTOS DE REGLAS OFICIALES *Talos* EVALUADAS.

Nombre	Fecha	N. Reglas	Reglas URI	DEL	Por defecto	N. CVE
2017	18/05/2017	43336	2343	138	133	1191
2018	03/01/0219	48921	2951	148	221	1503
2019	17/12/2019	52208	3219	150	247	1614
2020	23/04/2020	53233	3262	151	260	1637
2021	28/01/2021	56004	3351	153	257	1700

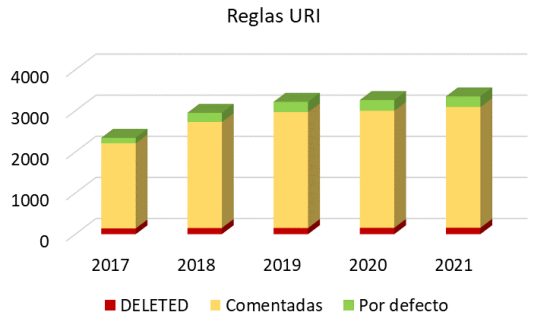
Por otra parte, se observa un impacto en algunos casos significativo del ajuste de reglas sobre las baterías de ataque (Tabla VI), tomando como referencia la máxima capacidad de detección posible (*M1*). Este se debe fundamentalmente a las reglas *DELETED*, es decir, al paso al modo *M2*. Sin embargo, se observa que la eliminación de reglas adicionales afecta a todas las baterías de ataque. El resultado es que en todos los casos se produce una reducción significativa (Δ) en el número de ataques detectados, oscilando entre el 4,93 % y el 13,28 % de diferencia absoluta. Sobre el total de ataques, el ajuste de las reglas origina una reducción de 8,37 puntos porcentuales sobre la máxima capacidad de detección posible. En cualquier caso, el ajuste de las reglas es necesario para evitar un número inmanejable de FP y produce un mayor impacto sobre estos que sobre la tasa de detección.

VI. EVOLUCIÓN TEMPORAL DE LAS CAPACIDADES DEL SIDS

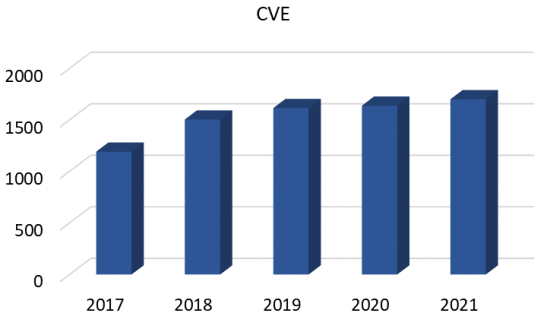
Finalmente, procedemos a evaluar el impacto de las actualizaciones de los conjuntos de reglas sobre los resultados de detección. Al igual que en el apartado anterior, únicamente utilizaremos *Inspectorlog* (Snort) para este análisis, dado que las actualizaciones son más frecuentes y están disponibles públicamente.

En la Tabla VII se detallan las características más relevantes de los *rulesets* oficiales de *Talos* obtenidos en diferentes fechas que se utilizarán en esta experimentación. En particular, se indican el número de reglas que afectan a los URI, el número de reglas marcadas como borradas (*DEL*), el número de reglas activas por defecto y el número de CVE directamente asociados a dichas reglas. Como puede observase, consideramos un conjunto de reglas previo a la adquisición del tráfico real, con más de un año de diferencia, y varios conjuntos posteriores hasta uno reciente. De esta forma, se analiza la evolución de las reglas entre mayo de 2017 y enero de 2021.

En la Figura 3 se muestra gráficamente la composición de los conjuntos de reglas considerados así como la evolución



a) Número y distribución de reglas en cada conjunto



b) Número de CVE asociados a al menos una regla

Figura 3. Evolución de las reglas *Talos* (sólo reglas URI).

del número de códigos CVE incorporados en las firmas. Esta información ha sido extraída de las propias firmas a partir del campo *reference*, por lo que es posible que exista una cobertura de CVE mayor que la indicada. Hemos de señalar que en el periodo considerado se ha producido un incremento del 30 % en el número de CVE para los que se incorporan reglas.

El primer análisis se realiza sobre el tráfico real en *Inves*, constatándose la inexistencia de variaciones en el número de URI de ataque detectados, a excepción de los modos *M1* y *M2*, donde llegan a detectarse 91791 y 85641 URI de ataque con las reglas más recientes. Estas variaciones son debidas a dos reglas nuevas que, de acuerdo a los criterios de ajuste, deben ser eliminadas por generar un alto número de alertas. El análisis de las mismas muestra que son realmente FP. Sí se observan pequeñas variaciones, porcentualmente insignificantes, en el número de alertas generadas, si bien, como se ha mencionado, no afectan al número de URI etiquetados. Por tanto, la evolución de las reglas no afecta a la clasificación obtenida con las reglas vigentes en el momento de la captura e, incluso, con reglas de un año antes. Este resultado sólo se puede justificar por la inexistencia de ataques desconocidos a la fecha de adquisición o la no inclusión de las firmas correspondientes a los posibles ataques de día cero presentes. En este sentido, resulta relevante indicar que el análisis realizado con *modSecurity* y *CheckpointIPS* ha identificado 412 ataques adicionales y que durante el proceso de limpieza aplicado en [5] se identificaron algunos ataques más que habían pasado inadvertidos por los SIDS, así como numerosas anomalías presumiblemente asociadas a ataques, por lo que cabría esperar que una actualización de las reglas hubiese incrementado el número de detecciones.

A pesar del incremento significativo en el número de

Tabla VIII
EVOLUCIÓN TEMPORAL DEL PORCENTAJE DE URI CLASIFICADOS COMO ATAQUES (CD)

Dataset	Modo	2017	2018	2019	2020	2021
ataques-800	M1	91,95	92,07	92,07	92,07	92,07
	M5	85,10	85,46	85,46	85,46	85,46
ataques-1100	M1	93,54	93,71	93,71	93,71	93,71
	M5	87,50	87,84	87,84	87,84	87,84
Fwaf-bad	M1	50,56	53,77	53,77	53,77	53,81
	M5	42,52	45,80	45,80	45,80	45,85
Fwaf-2200	M1	91,23	91,23	91,23	91,23	91,23
	M5	85,59	85,59	85,59	85,59	85,59
rdb	M1	91,01	91,01	91,01	91,01	91,01
	M5	86,08	86,08	86,08	86,08	86,08
osvdb	M1	66,86	66,99	66,99	66,99	66,96
	M5	53,39	53,70	53,73	53,76	53,75
A-420	M1	80,71	81,67	81,67	81,67	81,67
	M5	75,24	76,67	76,67	76,67	76,67
Promedio	M1	56,12	58,71	58,71	58,71	58,74
	M5	47,68	50,34	50,34	50,35	50,38

reglas y de CVE cubiertos por las mismas, los resultados experimentales obtenidos para las baterías de ataque (Tabla VIII) muestran un comportamiento que no corresponde, en la mayoría de los casos, con una mejora significativa de los resultados. A este respecto, mostramos los resultados para los modos *M1* y *M5*, por corresponder a la máxima tasa de detección posible y a la óptima. Como se observa en la tabla, en las baterías de ataques utilizadas se mantiene constante la tasa de detección (casos de *Fwaf-2200*, y *rdb*) o los cambios son insignificantes. En todos los casos en los que hay cambios, estos tienen lugar para el paso de las reglas *2017* a *2018* y únicamente se produce un ligero incremento para *Fwaf-bad* en el paso de las reglas *2020* a *2021* y en *osvdb* para las reglas optimizadas. Resulta significativo que tanto *ataques-800* como *ataques-1100* incorporan ataques correspondientes a vulnerabilidades descubiertas en 2016 y 2017, por lo que sería de esperar un incremento significativo de ataques detectados al pasar de *2017* a *2018*. El caso de *A-420* (Figura 4) es aun más relevante, puesto que todos los ataques incorporados corresponden a vulnerabilidades de 2017 y 2018, además con un valor de criticidad alto (CVSS>9). Sin embargo, las reglas de 2017 ya detectan un porcentaje relevante de los ataques (75 % con reglas optimizadas), que apenas varía en 2018 y que luego se mantiene constante. Por tanto, podemos concluir que la respuesta a los ataques *0-day* resulta muy limitada y que, en todo caso, se hace en un breve plazo después de la publicación de los mismos. En este sentido, *rdb* y *osvdb* contienen ataques relativamente antiguos (de antes de 2009) que no son suficientemente cubiertos por las reglas y para los que no ha habido evolución significativa en la capacidad de detección. Como se puede observar en la Figura 4, la evolución en la tasa de detección sobre el total de ataques es reducida.

VII. CONCLUSIONES

Los resultados obtenidos con los SIDS tienen una clara dependencia con el sistema concreto y, obviamente, con las reglas utilizadas. En nuestro estudio exploratorio, obtenemos que la combinación de diferentes SIDS puede ofrecer una mejor capacidad de detección (ya que ofrecen resultados

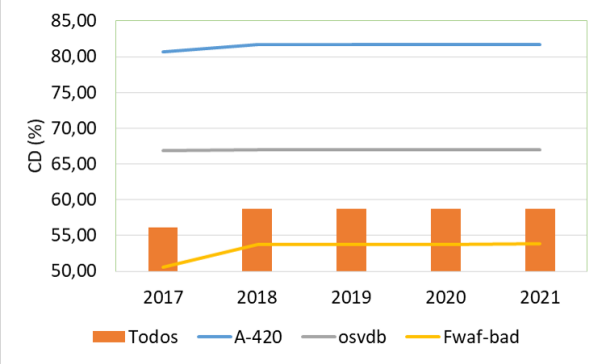


Figura 4. Capacidad de detección en el mejor caso

complementarios) a costa de generar un número excesivo de FP. También hemos demostrado que es posible optimizar las reglas de manera sencilla de forma que se alcance un equilibrio entre el número de FP reducido y la capacidad de detección perdida. Por ello, al igual que ocurre con los AIDS, es necesario ajustar y adaptar los SIDS al escenario de aplicación para evitar que el número de FP mine la utilidad del detector. Aunque esta idea no es nueva, en la práctica los usuarios suelen desplegar los sistemas tal como se distribuyen, por lo que pensamos que es necesario insistir en la difusión de esta idea. Es más, en su configuración por defecto, los resultados obtenidos son claramente insatisfactorios.

Finalmente, la respuesta ante los ataques de día cero es insuficiente y muy limitada en el tiempo. La evolución temporal de las reglas respecto a los resultados obtenidos muestra que los nuevos ataques sólo se consideran durante un breve periodo durante su aparición y son generalmente insuficientemente cubiertos.

Por todo lo anterior, concluimos que los SIDS distan de ser una protección suficiente en la actualidad debido a su limitada cobertura y a que, en la práctica, no suele realizarse una fase adecuada de ajuste de reglas por parte de los usuarios.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto 2020/00000172 dentro del programa de *Proyectos singulares de actuaciones singulades de transferencia en los CEI en las áreas RIS3* de la Junta de Andalucía y los proyectos PI-1921/22/2019 y PI-1786/22/2018 de la Universidad de Sevilla.

Referencias

[1] Moustafa, N., Hu, J., Slay, J. (2019). A holistic review of Network Anomaly Detection Systems: A comprehensive survey. Journal of Network and Computer Applications, 128, 33–55. https://doi.org/10.1016/j.jnca.2018.12.006

[2] I. Ghafir, V. Prenosil, J. Svoboda, M. Hammoudeh, (2016). A Survey on Network Security Monitoring Systems, 2016 IEEE 4th Int. Conf. on Future Internet of Things and Cloud Workshops (FiCloudW), 77-82. doi: 10.1109/W-FiCloud.2016.30.

[3] Holm, Hannes, (2014). Signature based intrusion detection for zero-day attacks: (Not) A closed chapter?, Proc. of the Annual Hawaii Int. Conf. on System Sciences, 4895–4904.

[4] Garcia-Teodoro, P., Diaz-Verdejo, J. E., Tapiador, J. E., Salazar-Hernandez, R. (2015). Automatic generation of HTTP intrusion signatures by selective identification of anomalies. Computers and Security, 55, 159–174. https://doi.org/10.1016/j.cose.2015.09.007

[5] Díaz-Verdejo, J.E., Estepa, A., Estepa, R., Madinabeitia, G,. Muñoz-Calle, F.J. ()2020). A methodology for conducting efficient sanitization of HTTP training datasets", Future Generation Computer Systems, 109, 67-82.

[6] Agrawal, S., Agrawal, J. (2015). Survey on anomaly detection using data mining techniques. In Procedia Computer Science 60, 708–713). https://doi.org/10.1016/j.procs.2015.08.220

[7] Waagsnes, H., Ulltveit-Moe, N. (2018). Intrusion detection system test framework for SCADA systems. In ICISSP 2018 - Proc. of the 4th Int. Conf. on Information Systems Security and Privacy 2018-January, 275–285). SciTePress. https://doi.org/10.5220/0006588202750285

[8] E. Albin and N. C. Rowe, (2012). A Realistic Experimental Comparison of the Suricata and Snort Intrusion-Detection Systems,"2012 26th Int. Conf. on Advanced Information Networking and Applications Workshops, 122-127, doi: 10.1109/WAINA.2012.29.

[9] Chitrakar R., Chuanhe H., (2012). Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k-Medoids Clustering and Naïve Bayes Classification, In Proc. of 8th IEEE Int. Conf. on Wireless Communications, Networking and Mobile Computing (WiCOM), 1-5.

[10] Singh, S., Sharma, P. K., Moon, S. Y., Park, J. H. (2017). A hybrid layered architecture for detection and analysis of network based Zero-day attack. Computer Communications, 106, 100–106. https://doi.org/10.1016/j.comcom.2017.01.019

[11] Gulshan Kumar, Kutub Thakur, Maruthi Rohit Ayyagari (2020). MLE-sIDSs: machine learning-based ensembles for intrusion detection systems—a review. The Journal of Supercomputing (2020) 76:8938–8971. https://doi.org/10.1007/s11227-020-03196-z1

[12] Peddabachigari, S., Abraham, A., Grosan, C., Thomas, J. (2007). Modeling intrusion detection system using hybrid intelligent systems. Journal of Network and Computer Applications, 30(1), 114–132. https://doi.org/10.1016/j.jnca.2005.06.003

[13] Ying Zhong, Wenqi Chen, Zhiliang Wang, Yifan Chen, Kai Wang, Yahui Li, Xia Yin, Xingang Shi, Jiahai Yang, Keqin Li (2020). HELAD: A novel network anomaly detection model based on heterogeneous ensemble learning, Computer Networks, 169, 107049, https://doi.org/10.1016/j.comnet.2019.107049.

[14] Spathoulas, G. P., Katsikas, S. K. (2013). Enhancing IDS performance through comprehensive alert post-processing. Computers and Security, 37, 176–196. https://doi.org/10.1016/j.cose.2013.03.005

[15] Neminath Hubballi, Vinoth Suryanarayanan, (2014). False alarm minimization techniques in signature-based intrusion detection systems: A survey, Computer Communications, 49, 1-17. https://doi.org/10.1016/j.comcom.2014.04.012.

[16] Raftopoulos, E., Dimitropoulos, X (2013). A quality metric for IDS signatures: in the wild the size matters. EURASIP J. on Info. Security, 7. https://doi.org/10.1186/1687-417X-2013-7

[17] Estepa, R., Diaz-Verdejo, J.E., Estepa, A., Madinabeitia, G. (2020). How Much Training Data is Enough? A Case Study for HTTP Anomaly-Based Intrusion Detection. IEEE Access, 8, 44410-44425.

[18] Modsecurity: Open source web application firewall, 2021, https://modsecurity.org/. (Último acceso 7 marzo 2021).

[19] Snort, 2021, https://www.snort.org. (Último acceso 7 marzo 2021).

[20] Z. Afzal and S. Lindskog, (2016). IDS rule management made easy. 2016 8th Int. Conf. on Electronics, Computers and Artificial Intelligence (ECAI), 1-8. doi: 10.1109/ECAI.2016.7861119.

[21] John Sonchack, Adam J. Aviv, Jonathan M. Smith, (2015). Cross-domain collaboration for improved IDS rule set selection. Journal of Information Security and Applications, 24–25, 25-40. https://doi.org/10.1016/j.jisa.2015.10.001.

[22] Asad H., Gashi I. (2018). Diversity in Open Source Intrusion Detection Systems. In: Computer Safety, Reliability, and Security. SAFECOMP 2018. LNCS, 11093, 267-281. https://doi.org/10.1007/978-3-319-99130-6_18

[23] D. Bekerman, B. Shapira, L. Rokach, A. Bar (2015). Unknown malware detection using network traffic classification. 2015 IEEE Conference on Communications and Network Security (CNS), 134-142, doi: 10.1109/CNS.2015.7346821.

[24] J. Díaz Verdejo, R. Estepa, A. Estepa, G. Madinabeitia, D. Rodríguez (2018). Metodología para la generacion de conjntos de datos de ataques basados en URI de HTTP. Actas de las IV Jornadas Nacionales de Investigación en Ciberseguridad, 119-126. ISBN: 978-84-09-02697-5.

[25] Salazar-Hernández, R., Díaz-Verdejo, J. E. (2010). Hybrid detection of application layer attacks using Markov models for normality and attacks. In LNCS, 6476, 416–429. https://doi.org/10.1007/978-3-642-17650-0_29

[26] Fwaf-Machine-Learning-driven-Web-Application-Firewall, https://github.com/faizann24/Fwaf-Machine-Learning-driven-Web-Application-Firewall Último acceso 7 marzo 2021)

http://doi.org/10.18239/jornadas_2021.34.10

Diseño y evaluación de modelos de aprendizaje automático no supervisado para la detección de anomalías en un sistema Spark.

Farid Bagheri-Gisour Marandyn, Xavier. Larriva-Novo, Víctor A. Villagrá
Universidad Politécnica de Madrid (UPM). DIT, ETSI Telecomunicaciones. Avda. Complutense, 30. 28040 Madrid
farid.academ@gmail.com, {xavier.larriva.novo, victor.villagra }@upm.es

Resumen- En el presente artículo se muestran los resultados obtenidos al diseñar y evaluar un modelo de aprendizaje automático no supervisado como es el K-Means para la detección de anomalías en tiempo real sobre múltiples sensores dentro de un sistema Spark utilizando un threshold para delimitar esas posibles anomalías. Los resultados obtenidos del modelo (aún en fase de desarrollo y mejora) demuestran la capacidad de poder detectar tres tipos de eventos: eventos no anómalos, eventos anómalos por características y eventos anómalos por aspectos temporales. Este comportamiento presenta características estimulantes para poder aplicar este tipo de algoritmos en un entorno real donde los datos no tienen ningún tipo de etiquetado. Todo ello, sumado a la capacidad que ofrece Spark para realizar el procesado de grandes volúmenes de datos en tiempo real, da como resultado un sistema prometedor capaz de clasificar eventos procedentes de diversos sensores de manera inmediata.

Index Terms- machine learning, K-Means, threshold, Spark, detección anomalías, ciberseguridad

Tipo de contribución: Investigación en desarrollo

I. Introducción

Hasta hace no mucho tiempo, la tarea de realizar un análisis y posterior extracción de información de la vasta amalgama de datos que una entidad empresarial posee era una tarea faraónica e inabarcable para los equipos, metodologías y mecanismos que clásicamente se venían usando en el área de ciberseguridad, más aún si las necesidades requerían que el trabajo de detección de anomalías se realizase de manera casi instantánea, a tiempo real.

Por ello, para poder solventar esos requisitos de seguridad, desde el mundo de la ciberseguridad se empezó a plantear utilizar métodos y mecanismos de una de las ramas de la computación que durante los últimos años ha experimentado un crecimiento exponencial, el campo del machine learning y el big data.

En el campo del machine learning, los modelos de aprendizaje no supervisado son los que más interés suscitan, ya que los datos que se les suministra no poseen ningún etiquetado, por lo que el conocimiento tiene que ser adquirido de la propia metaestructura de los datos entregados. Esta característica es fundamental para el análisis de la gran mayoría de los datos que se generan por los equipos, ya que dichos datos no poseen etiquetado alguno.

Dentro del big data se desarrollaron diversos mecanismos y modelos que permiten realizar el manejo y procesado de grandes volúmenes de datos a tiempo real. Dentro de las diversas propuestas se destaca Apache Spark, sistema bien conocido open-source, escalable, que permite procesar grandes volúmenes de datos mediante la paralelización del trabajo y en tiempo real gracias a Spark Streaming.

Por tanto, el objetivo del proyecto se centrará en utilizar las bondades que ofrece la herramienta Apache Spark para poder realizar un sistema de autoaprendizaje no supervisado que permita detectar anomalías de un conjunto de datos provenientes de varios sensores a tiempo real. A través de una de las API de python que ofrece Spark, denominada PySpark y que ofrece a su vez una librería de modelos de machine learning, se puede ejecutar e integrar con todo el ecosistema de Spark el algoritmo de aprendizaje no supervisado K-Means, para realizar esa tarea de detección de anomalías en tiempo real. Así mismo, el entorno provee todas las herramientas necesarias para el preprocesamiento de los datos y la posterior conexión a un sistema ELK (ElasticSearch, Logstash y Kibana) donde se podrá visualizar y analizar los resultados. Se propone ese modelo, ya que es uno de los modelos de aprendizaje no supervisado mejor conocidos que permita agrupar los datos en clústeres (conjunto de datos que se agrupan como similares).

El entrenamiento del modelo se realizará con datasets sin anomalías, que serán generados por cada fuente de datos. El correcto modelado del sistema requerirá del preprocesamiento de los datos a través de las herramientas de Pyspark mencionadas, para que el algoritmo K-Means pueda ser correctamente entrenado.

La arquitectura de detección de anomalías se basa en el diseño de un sistema de threshold junto al algoritmo K-Means que permita clasificar, en base a los clústeres formados por el propio algoritmo previamente, qué datos provenientes de las distintas fuentes de datos son anómalos. Se harán uso de técnicas que permitan optimizar el modelo a través de sus hiperparámetros y tras el diseño del sistema, se pasará a evaluar su comportamiento.

Las pruebas que se realizan determinan si el modelo efectúa una correcta clasificación de los datos en anomalía/no anomalía en función de las características de los datos, así como de la marca temporal del mismo.

Por último, se integra al sistema, sensores que proporcionan las fuentes de datos, así como un sistema ELK en donde pueden ser almacenados, gestionados y analizados los resultados. Se comprobará que el sistema cumpla los requisitos planteados en el escenario real, exponiendo aquellas ventajas y desventajas ofrecidas por la arquitectura, así como posibles mejoras y líneas futuras que puedan surgir del proyecto.

II. Metodología

El sistema de aprendizaje automático propuesto se enmarca en una arquitectura de adquisición, procesamiento y tratamiento de datos de un sistema de conciencia cibersituacional y es el

encargado de la detección de anomalías a partir de los datos de entrada del sistema. Los datos de entrada se componen de diversos conjuntos de valores numéricos en diversos formatos provenientes de distintos sensores de actividad. Estos datos no pueden ser inyectados directamente al algoritmo K-Means, sino que deben ser tratados con anterioridad para que puedan ser procesados. A su vez, el algoritmo requiere de una fase previa de entrenamiento y validación para que el sistema no muestre resultados aleatorios. Por último, los datos de salida del sistema están formadas por el conjunto de características respectivas al evento de entrada más el etiquetado por un valor 'False' o 'True' indicando si se ha detectado como anomalía o no

Por todo ello, se realizó una arquitectura compuesta por diferentes fases y subsistemas que se listan a continuación:

- Recepción de datos de los sensores.
- Subsistema de generación de datos sintéticos.
- Subsistema de preprocesamiento.
- Subsistema de entrenamiento y validación.
- Subsistema de procesamiento a tiempo real.
- Envío/almacenado de los resultados obtenidos.

Cada uno de los subsistemas está compuesto a su vez de diferentes módulos para poder realizar su tarea asignada.

La Figura 1 muestra la arquitectura del sistema realizado, con las conexiones entre los diferentes subsistemas que conforman el sistema completo.

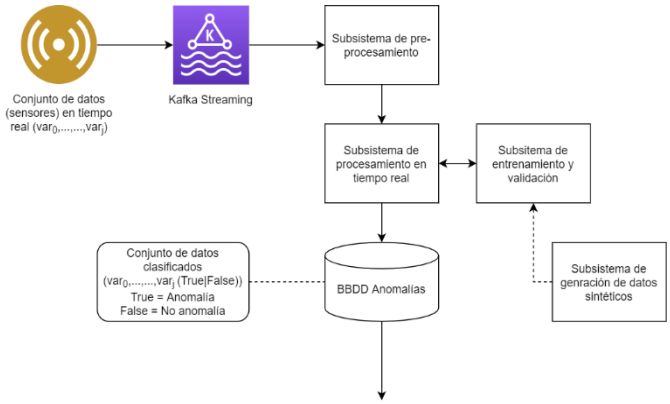


Figura 1: Arquitectura modular del sistema de aprendizaje automático

A. Subsistema de generación de datos sintéticos

Los datos que recibe la arquitectura son datos con formato json donde, dependiendo de su procedencia, tendrá campos diferentes propios de cada sensor. Estos datos serán consumidos por la arquitectura mediante los Topic de Kafka.

B. Subsistema de generación de datos sintéticos

El subsistema de generación de datos sintéticos está destinado a la creación de datasets sintéticos en aquellos modelos en los que no se posea un dataset acorde a las posteriores necesidades del modelo, ya que hay ciertos sensores que no pueden recolectar las suficientes entradas de datos distintas por estar desplegadas en un entorno no real.

La arquitectura del subsistema se muestra en la Figura 2, donde se puede observar los diferentes módulos que la componen.

El subsistema utiliza los datos que se quiera proporcionar a su entrada y comienza a establecer las relaciones entre los atributos necesarias para que no se produzcan entradas no válidas. A su vez, se definen los perfiles de generación de eventos, estableciéndose el número de eventos que se producen por paso de reloj y seguidamente, se añade la probabilidad de que cada perfil se ejecute por paso de reloj.

La configuración del reloj se establece con anterioridad, indicando el tiempo de simulación, así como el tiempo entre pasos. Esta configuración estará condicionada por un perfil temporal, en donde se indica las características de generación de eventos según la hora (p. ej. mayor cantidad de pasos en horas laborables).

Por último, se definen los atributos a generar ya con todas las configuraciones, relaciones, perfiles y se pasa a su generación. Cabe indicar que se establecen dos configuraciones distintas relativas a la generación de un dataset de tipo normal y otro de tipo anómalo.

Todas estas configuraciones son relativas a la librería para la generación de datasets realistas Trumania [1], librería que ofrece las herramientas necesarias para poder generar sintéticamente datos, definir estructuras internas del dataset e impedir o permitir la generación de eventos que se ajusten a las siguientes necesidades:

- Tipos adecuados para todos los valores del dataset, facilitando su posterior modificación.
- Estructura de creación de eventos no uniforme
- Estructura temporal verosímil respecto a escenarios reales normales y anómalos.
- La no posibilitación de creación de eventos que no puedan darse en entornos reales.

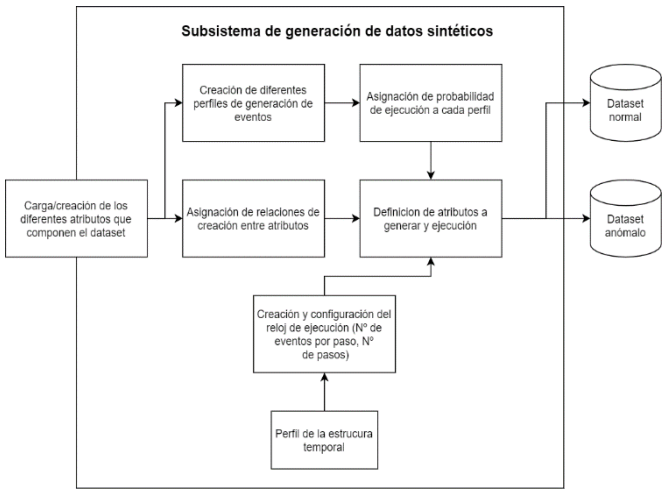


Figura 2: Arquitectura modular del subsistema de generación de datos sintéticos

Los datos de salida del subsistema corresponden a la pareja de dataset generados para un modelo específico, un dataset con características normales y un dataset con características anómalas. El tamaño de estos datasets creados son del orden de 100.000 entradas.

C. Subsistema de preprocesamiento de datos

El subsistema de preprocesamiento de datos es el encargo de aplicar modelos algorítmicos y funciones matemáticas de acuerdo con los datos con los que trabajen los diferentes modelos de machine learning de cada sensor.

El subsistema de preprocesamiento de datos queda definido en la Figura 3, donde se puede observar los diferentes subpartes que la componen.

Este preprocesamiento es previo al entrenamiento de los algoritmos de machine learning y al procesamiento en tiempo real. El subsistema se compone de:

- Estructurado de los datos.

- Aplicación de los módulos y funciones de preprocesamiento.
- Confección de vector de características final.

Los datos de entrada del subsistema de preprocesamiento se componen del conjunto de valores de datos respectivos a cada uno de los sensores, que deben ser procesados por los algoritmos de machine learning.

Esos datos son primeramente estructurados según el tipo de valor que tengan, teniendo así una definición del tipo de datos que aparecerá en los datos de entrada.

Seguidamente entran en juego los módulos de preprocesamiento ofrecidas por Pyspark y las funciones matemáticas definidas. Entre los diferentes módulos usados (dependiendo del dato con el que se lidie) se encuentran:

String Indexer. Módulo de preprocesamiento que codifica strings a índices numéricos.

Min Max Scaler. Módulo de preprocesamiento que normaliza por defecto valores numéricos al rango [0, 1].

One Hot Encoder (OHE). Módulo de preprocesamiento que asigna a valores categóricos un vector binario. La longitud del vector binario dependerá de la cantidad de valores categóricos distintos que posean los datos a procesar.

Regex Tokenizer. Módulo de preprocesamiento que realiza la separación del texto en múltiples subgrupos según la expresión regex definida.

Count Vectorizer. Módulo de preprocesamiento que transforma un conjunto de strings a vectores de token.

TF-IDF. Módulo de preprocesamiento que refleja la importancia de los términos de un texto dentro de un corpus.

Módulos de preprocesamiento específicos de cada modelo. Este conjunto hace referencia a aquellas transformaciones concretas necesarias para adecuar los datos de entrada de cada modelo.

Vector Assembler. Módulo de preprocesamiento que combina una lista de columnas dadas en un único vector columna. Esto es el paso final del preprocesamiento.

Los datos de salida del subsistema de preprocesamiento corresponden a vectores unidimensionales que recogen las características preprocesadas de los datos de entrada ofrecida por los sensores y/o datasets.

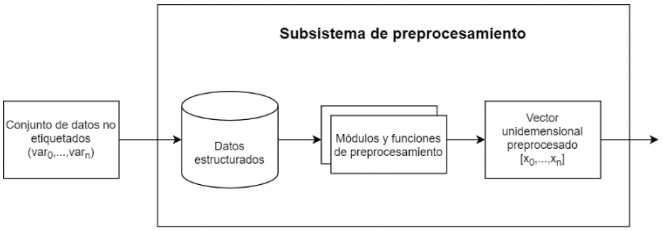


Figura 3: Arquitectura modular del subsistema de preprocesamiento

D. Subsistema de entrenamiento y validación

El subsistema de entramiento y validación es el encargado de la preparación y la supervisión del correcto funcionamiento del modelo de machine learning para la detección de anomalías por cada tipo de sensor.

Para ello, se utilizan los módulos de selección de hiperparámetros, métricas y datos de validación que se pueden observar en la Figura 4, en el que se muestra la arquitectura completa de este subsistema.

Los datos de entrada serán esos vectores ofrecidos por el subsistema de preprocesamiento, que pasarán a ser usados para entrenar y validar las diferentes propuestas de configuración del modelo de machine learning.

Seguidamente, el subsistema de selección de hiperparámetros y métricas permite optimizar y analizar el rendimiento que el algoritmo de machine learning muestra al variar los hiperparámetros, en un proceso iterativo, que lo componen. Son dos las métricas que se usan para medir el grado de mejora:

WSSSE. Mide por cada punto del dataset cuan lejano está del centroide de los clústeres. Se busca reducir esta medida lo máximo posible sin llegar al overfitting.

$$\sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 \qquad (1)$$

donde,
k = Clúster
S_k = Datos del cluster k,
x_{kj} = valor j del vector centroid del clúster k

Silhouette. Métrica que permite conocer la cohesión dentro de un clúster a la vez que la separación con otros clústeres, midiendo así cuan bien un dato está clasificado en un clúster. El rango de valores esta entre [-1,1] donde altos valores indican que los datos están bien asignados (alta cohesión a su clúster, alta separación con otros clústeres).

$$\mathbf{s(i)} = \frac{\mathbf{b(i)} - \mathbf{a(i)}}{\mathbf{max(b(i), a(i))}} \qquad (2)$$

donde,

a(i): Distancia media con todos los puntos de su clúster
b(i): Distancia media con todos los puntos del clúster más cercano
s(i): Coeficiente Silhouette para el iésimo punto

La selección del hiperparámetro optimo será, por tanto, el resultado del compendio del valor la métrica WSSSE y el valor que proporcione la métrica Silhouette.

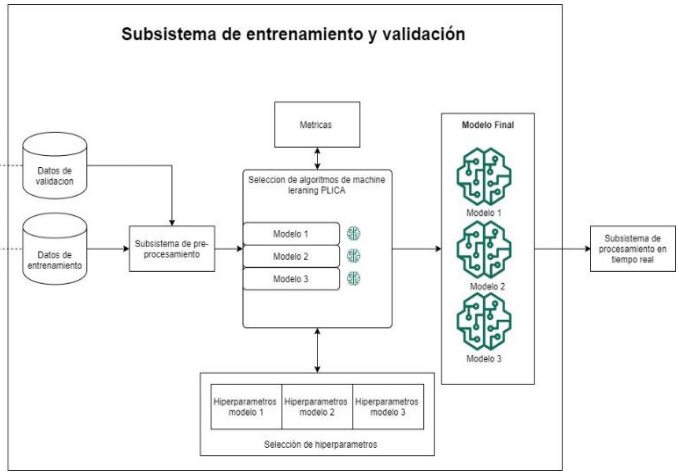


Figura 4: Arquitectura modular del subsistema de entrenamiento y validación

C.1. Modelo de Machine Learning.

El algoritmo que se usará para la realización del modelo de identificación de anomalías será el K-Means. Este algoritmo permite realizar agrupar los diferentes datos en clústeres de características similares de manera no supervisada.

La elección de este modelo radica principalmente a su amplio uso en el estado del arte del clustering y la posibilidad de su uso en la librería de Mlib de Pyspark.

Para la detección de anomalías en este modelo, se pasará a usar un threshold por cada clúster formado que permita delimitar que datos son lo suficientemente diferentes al resto de datos del clúster para que pueda ser considerado como una anomalía.

El threshold se considera como el límite de cada uno de los clústeres formados. El threshold es calculado como el punto más lejano al centroide de cada uno de los clústeres (aunque puede ser ajustado como el i-ésimo punto más lejano). Todos los puntos más lejanos a esos thresholds serán clasificados como anomalías.

Por tanto, el modelo es entrenado con un conjunto de datos que se consideran no anómalos para que en la detección de nuevos eventos se pueda comparar y observar si esos nuevos datos son lo aptos para detectarse como anomalías o no. Esta implementación permite entre otras características, tener en cuenta desplazamientos temporales de los datos, ya que la marca temporal es una característica propia de los datos.

E. Subsistema de procesamiento a tiempo real

El subsistema de procesamiento en tiempo real es el sistema que realiza finalmente la funcionalidad de la identificación de anomalías a tiempo real, mediante la integración de todas las funciones y modelos de preprocesamiento descritos, y el cargado del modelo de machine learning ya entrenado y optimizado en el paso anterior.

El sistema obtendrá los eventos de los sensores mediante un Topic de Kafka (un Topic por sensor), los preprocesará y los enviará el modelo de machine learning donde este lo clasificará según sea una posible anomalía o no.

Por último, los resultados se enviarán a elasticsearch, donde se almacenarán y podrán ser visualizados mediante kibana.

III. Resultados

Para observar el funcionamiento del sistema, y por tanto, que bien se comporta a las competencias que se le quiere asignar, se realizaron tres tipos de pruebas por cada modelo asociado a cada sensor. Se consideraron 6 sensores distintos, proporcionando datos de actividad referentes a señales Wifi, Bluetooth, redes de telefonía móvil, señales de radiofrecuencia, así como logs de firewalls y SIEMs. Las pruebas son:

- Identificación como datos normales aquellos datos que se ha visto durante el entrenamiento.
- Identificación como posibles datos anómalos aquellos datos nunca vistos y que se diferencien lo suficiente de los datos de entrenamiento.
- Identificación como anomalías aquellos datos vistos por el modelo durante el entrenamiento, pero desplazados temporalmente a una hora no común (por ejemplo, madrugada).

Los resultados arrojados por las pruebas realizadas para la confección del modelo y la medición de su desempeño son las siguientes:

De la Tabla 1 se puede aprecia que lo hiperparámetros de la tolerancia como de las máximas iteraciones son iguales. Esto se debe a que se comprobó que tales hiperparámetros no marcaban ninguna diferencia al variarse, por lo que se dejo por defecto. Por tanto, el hiperparámetro que influye considerablemente es el número de clústeres que tiene.

Tabla I
HIPERPARÁMETROS ÓPTIMOS POR CADA MODELO

Modelo del sensor	N.º de clústeres	Tolerancia	Máximas iteraciones
RM	25	1e-4	100
RF	17	1e-4	100
Bluetooth	15	1e-4	100
Wifi	17	1e-4	100
Firewall	18	1e-4	100
Siem	18	1e-4	100

Tabla II
PRECISIÓN (%) EN LA DETECCIÓN DE LOS DIFERENTES DATOS

Modelo del sensor	Normales	Anómalos por características	Anómalos temporales
RM	100.00 %	96.00 %	56.65 %
RF	94.56 %	100.00 %	71.84 %
Bluetooth	100.00 %	100.00 %	94.44 %
Wifi	100.00 %	100.00 %	72.89 %
Firewall	100.00 %	100.00 %	76.97 %
Siem	99.0 %	100.00 %	46.51 %

La Tabla 2 arroja el resultado de medir la precisión del sistema en un entorno en el que se definió que datos son normales y que datos son posibles anomalías. El resultado muestra que casi todos los datos considerados normales y posibles anomalías por características son clasificados como tal. En contrapartida, se puede ver que el rendimiento del modelo a la hora de definir anomalías temporales es mucho menor. No obstante, estos valores se deben en parte a que la distribución de detección de anomalías no es uniforme. Casi todas las anomalías detectadas son en un rango de horas de 1 a 5 AM, siendo el número de eventos aquí menor. Por tanto, no se podría hablar de un mal funcionamiento del modelo, sino un incompleto reflejo del desempeño del modelo.

IV. Conclusiones

A la vista de los datos observados se puede ver que los modelos realizados por cada sensor no tienen más complicación en realizar una detección de lo que se considero para las pruebas datos normales y posibles datos anómalos por características.

Por contraparte, el desempeño en la detección de posibles datos anómalos temporales, aún teniendo en cuenta esa distribución no uniforme, cae bastante con respecto a los otros dos y es punto de mejora de siguientes versiones.

Cabe recalcar que hay que coger estos datos con cuidado, ya que se está en fase de pruebas y se necesitan más tests para corroborar estos resultados.

V. Referencias

[1] Sv3ndk, Milanvdm, FHachez, Thomas-jakemeyn, Petervandenabeele: "Trumania", https://github.com/RealImpactAnalytics/trumania, 2020

[2] Erdem, Yadigar & Ozcan, Caner: "Fast Data Clustering and Outlier Detection using K-Means Clustering on Apache Spark", 2017

[3] Berthold, Michael & Höppner, Frank: "On Clustering Time Series Using Euclidean Distance and Pearson Correlation", 2016

[4] Peng, Kai & Huang, Qingjia: Clustering Approach Based on Mini Batch Kmeans for Intrusion Detection System over Big Data", 2018

[5] Zhang, Tao & Li, Haibin & Xu, Lexi & Gao, Jie & Guan, Jian & Cheng, Xinzhou: "Comprehensive IoT SIM Card Anomaly Detection Algorithm Based on Big Data", 2019

[6] Kumari, R. & Sheetanshu, & Singh, M. & Jha, R. & Singh, N.K.: "Anomaly detection in network traffic using K-mean clustering", 2016

[7] Han, Li.: "Using a Dynamic K-means Algorithm to Detect Anomaly Activities", 2011

[8] Lima, M.F., Zarpelão, B., Sampaio, L.D., Rodrigues, J., Abrão, T., & Proença, M.L: "Anomaly detection using baseline and K-means clustering", 2010

[9] Wazid, M., Das, A.K.:" An Efficient Hybrid Anomaly Detection Scheme Using K-Means Clustering for Wireless Sensor Networks", 2016

VI. Agradecimientos

La investigación presentada en este artículo ha sido parcialmente financiada por el proyecto PLICA dentro del programa coincidente del Ministerio de Defensa del Gobierno de España.

http://doi.org/10.18239/jornadas_2021.34.11

Aplicación de transfer learning a la detección de malware

David Escudero García
Research Institute on Applied Sciences in Cybersecurity
Campus de Vegazana s/n, 24071 León, España
descg@unileon.es
ORCID:0000-0002-3776-3920

Ángel Luis Muñoz Castañeda, Noemí DeCastro-García
Universidad de León
Campus de Vegazana s/n, 24071 León, Spain
{amunc, ncasg}@unileon.es
ORCID:{0000-0001-6993-9110, 0000-0002-5610-0153}

***Resumen*—La detección de malware es un problema que ha adquirido una mayor importancia en los últimos años con el incremento en la distribución de archivos maliciosos. El uso de machine learning es una de las soluciones propuestas en la literatura por su capacidad de detectar nuevas muestras maliciosas respecto a soluciones tradicionales como los antivirus. No obstante, el uso de machine learning en este ámbito se encuentra con problemas como el desbalanceo entre ficheros maliciosos y benignos y la necesidad de actualizar los modelos para hacer frente a nuevo malware, que dificulta la construcción de una solución eficaz. En este trabajo planteamos el uso de técnicas de transfer learning para intentar paliar estos problemas y llevamos a cabo una evaluación de diferentes algoritmos de transfer learning para determinar si permiten construir modelos que sean capaces de detectar malware en diferentes horizontes temporales aprovechando datos antiguos.**

***Index Terms*—Ciberseguridad, Transfer learning, Machine learning, Detección de malware**

Tipo de contribución: *Investigación original*

I. Introducción

La detección de malware es un problema al que se le está prestando más atención recientemente [1] a causa del incremento año a año de las amenazas [2]. En la literatura, una de las soluciones más comunes pasa por aplicar técnicas de machine learning, que no dependen de firmas ya conocidas como en el caso de soluciones antivirus tradicionales.

No obstante, la aplicación de machine learning a la detección de malware se encuentra con varios obstáculos. Por un lado, el etiquetado de archivos como maliciosos o benignos es un proceso laborioso y que requiere de conocimiento experto, por lo que la obtención de datasets a gran escala es complicada. Esto implica que deben construirse modelos usando muestras ya etiquetadas, que pueden no corresponderse con el comportamiento de malware más reciente [1]. Esta situación se conoce como deriva conceptual [3]. En la práctica, esto suele significar que las características de las muestras maliciosas cambian en el tiempo, por lo que los modelos construidos sobre muestras antiguas no podrán detectar eficazmente las nuevas. Este obstáculo es producto de la formulación general del machine learning. Una de las principales suposiciones realizadas en el campo es que los datos usados están extraídos de distribuciones probabilísticas independientes e idénticas [4]. Esto quiere decir que aunque se combinen muestras de malware más antiguas con otras modernas, es probable que sus distribuciones sean distintas, así que no se podrán usar conjuntamente sin arriesgar una degradación del modelo.

Por otro lado, mucha de la literatura está centrada en la detección de malware sobre ficheros ejecutables bien de Windows ([5], [6], [7]), bien de Android ([8], [9]). Sin embargo, un 45 % de la distribución de malware se realiza a través de ficheros de Microsoft Office [10]. En principio, el problema de detección de malware en ficheros no ejecutables tendrá una naturaleza distinta del de detección de ejecutables, por lo que no se podrá simplemente aplicar los mismos modelos sin que se degrade su capacidad predictiva.

Finalmente, en situaciones realistas es común encontrarse con que los datos usados para construir el modelo están desbalanceados [1], en general con una proporción de malware muy baja respecto al total de muestras benignas.

Un posible método para paliar estos problemas es el uso de técnicas de transfer learning. El transfer learning es una rama del machine learning que tiene como objetivo aprovechar datos de un conjunto de datos, denominado fuente, para mejorar el aprendizaje y la capacidad predictiva de un modelo sobre otro conjunto de datos distinto pero relacionado, denominado objetivo. Las técnicas de transfer learning asumen que los dominios fuente y objetivo pueden tener distribuciones distintas.

Esto supone varias ventajas. En primer lugar, permite construir modelos para predecir sobre una cantidad de datos limitada. Por ejemplo, en un sistema de detección de malware, es posible que al inicio de un trimestre no se dispongan de suficientes muestras para entrenar un modelo eficaz, pero si que se pueden aprovechar las muestras ya acumuladas para realizar predicciones. Esto además tiene cierta utilidad en el caso de que los datos estén desbalanceados ya que se puede incrementar el número de instancias de la clase minoritaria, lo que mejorará el modelo.

Por otro lado, como no se asume que las distribuciones de fuente y objetivo sean las mismas, se podrá detectar malware sobre diferentes tipos de ficheros y sortear el problema de la deriva conceptual pudiendo aprovechar los datos antiguos. El aprovechamiento de los datos es de particular interés en el caso de la deriva conceptual, ya que bastantes trabajos se basan en la detección pasiva de su ocurrencia [11] y el reentramiento del modelo con muestras más recientes, descartando muestras más antiguas que sí podrían contener información de interés para la construcción del modelo.

En este trabajo realizaremos una evaluación de diferentes técnicas de transfer learning para determinar su eficacia a la hora de mejorar la predicción de los modelos sobre muestras de malware en diferentes horizontes temporales, en los que

cabe esperar que exista deriva conceptual por la evolución del malware en el tiempo y con clases desbalanceadas con una proporción de malware baja.

En la sección II se describe formalmente el concepto de transfer learning y la deriva conceptual; en la sección III se describe el trabajo relacionado. A continuación, en la sección IV se explica todo el protocolo experimental. En la sección V se exponen los resultados y finalmente se enuncian las conclusiones en la sección VI.

II. Formalización de transfer learning y la deriva conceptual

La principal caracterización formal de transfer learning viene dada en [4]. Un dominio $\mathcal{D} = \{\mathcal{X}, P(X)\}$ se define por dos componentes. Un espacio de features $\mathcal{X}$ que es básicamente el conjunto de features de los datos y una distribución de probabilidad marginal $P(X)$ sobre $\mathcal{X}$. Las features son los campos con los que se caracteriza cada dato. Por ejemplo, se puede caracterizar un ejecutable con el número de llamadas que hace a cada función de la API de Windows durante su ejecución. Formalmente la distribución de los datos es una función que indica la probabilidad de que las features del conjunto de datos tengan un valor concreto. Por ejemplo, en un conjunto de muestras de ransomware cabrá encontrase con que el número de llamadas a APIs relacionadas con el sistema de ficheros será mayor que en muestras que infecten un equipo para formar parte de una botnet. Hay que notar que según esta definición, dos conjuntos de datos referidos al mismo problema se consideran pertenecientes a distintos dominios si difieren en su espacio de features $\mathcal{X}$ o en su distribución $P(X)$.

Por otro lado, una tarea $\mathcal{T} = \{\mathcal{Y}, P(y \mid X)\}$ se compone de un espacio de etiquetas $\mathcal{Y}$ que son las diferentes clases a las que pueden pertenecer las instancias (malware y benigna por ejemplo) y una función predictiva $f(.)$ que asigna a un evento $X = x_i$ una etiqueta y de $\mathcal{Y}$. Esta función se interpreta de forma probabilística como $P(y \mid X = x_i)$. Es decir, una función que indica la probabilidad, dados los valores de sus features, de que una instancia pertenezca a una clase. El motivo de esta notación es que hace más explícita la relación entre los datos y su etiqueta según el problema que se quiera resolver. Por ejemplo, una muestra que durante su ejecución sobrescribe un ejecutable en la ruta del sistema C:\Windows\System32 probablemente tenga características maliciosas.

El objetivo de las técnicas de transfer learning es mejorar la función predictiva del dominio objetivo, $f(.)_T$, usando la información obtenida del dominio y la tarea fuente D_S y $\mathcal{T}_S$. El principal obstáculo es que los dominios y/o las tareas de ambos dominios pueden ser distintas ($D_S \neq D_T$ o $\mathcal{T}_S \neq \mathcal{T}_T$). Varios de los algoritmos propuestos en la literatura se basan en aplicar algún tipo de transformación sobre los datos que disminuya las diferencias entre las distribuciones de dominios fuente y objetivo.

Por su parte, la deriva conceptual es el cambio en la distribución de los datos a lo largo del tiempo. Consideramos que en el instante t los datos tienen una distribución marginal $P_t(X)$ y la distribución de las etiquetas condicionada a los datos $P_t(y \mid X)$. La deriva conceptual se produce cuando en un instante de tiempo posterior $t + \Delta t$ alguna de las dos distribuciones anteriores cambia. Es decir, $P_t(y \mid X) \neq P_{t+\Delta t}(y \mid X)$ o $P_t(X) \neq P_{t+\Delta t}(X)$ [3]. Dada esta definición, es fácil considerar que los datos procedentes del instante t conforman un dominio fuente del que se puede extraer conocimiento y los datos presentes en el instante $t+\Delta t$ constituye un dominio objetivo con una distribución diferente en el que realizar predicciones. Creemos, por lo tanto, que la aplicación de transfer learning a este problema puede dar buenos resultados.

El transfer learning resulta particularmente útil cuando el número de instancias en el dominio objetivo es mucho menor que las que se tienen en un dominio fuente. Para construir un buen predictor se necesita una cantidad considerable de datos etiquetados de la que no siempre se puede disponer y con muy pocos datos en el dominio objetivo se corre el riesgo de sobreajuste, que implica una menor capacidad de predicción del modelo entrenado.

III. Trabajo relacionado

En la literatura, existen numerosos trabajos dedicados a la detección de malware desde diferentes enfoques, usando features procedentes de las llamadas a API durante la ejecución de la muestra [12], su tráfico de red [13] o combinaciones de features de diferentes fuentes [14]. Buena parte de los trabajos se centran en detección de malware en archivos ejecutables, ya sea de Windows o Android, dejando a un lado otros tipos de archivos que añadirían complejidad al análisis.

Dentro de la adaptación a la deriva conceptual en el campo de la detección de malware, hay dos enfoques principales. Por un lado la detección y reentranamiento del modelo con muestras más recientes cuando se produce la deriva [11] y por otro el uso de modelos que soportan entrenamiento incremental, de modo que el modelo se actualiza con cada nueva instancia clasificada [15].

El método propuesto en [11] detecta la deriva usando las propias predicciones del modelo en lugar de monitorizar directamente los cambios observados en las features. El algoritmo detecta la deriva usando un análisis estadístico del ajuste de las muestras al modelo calculada a partir de la salida del clasificador. Cuando el ajuste de nuevas muestras con otras clasificadas en la misma clase se reduce, el algoritmo lanza una alarma, con la idea de que el modelo usado se reentrene con nuevas muestras para ajustarse a la nueva distribución. En [16] se propone un ensemble de clasificadores, entrenados sobre diferentes subconjuntos de features. Cuando se detecta que un clasificador del ensemble produce predicciones significativamente peores que las del ensemble completo, se reentrena usando datos recientes obtenidos sobre una ventana deslizante de tamaño configurable. Este enfoque tiene el problema de que se pueden descartar datos útiles en el reentrenamiento; en este trabajo buscamos evaluar si es beneficioso aprovechar los datos más antiguos para la construcción del modelo. Por otro lado, en [15], se presenta un esquema de detección online, que hace frente a la deriva mediante la actualización constante del modelo con nuevas muestras y el ajuste de la importancia de cada feature a lo largo del tiempo. De este modo, no es necesario reentrenar el modelo constantemente para ajustarse a la deriva. Los resultados son positivos, pero no todos los modelos pueden entrenarse de forma incremental. Además, en el caso de que los conjuntos de datos estén desbalanceados,

cabe la posibilidad de que este modo de operación degrade el rendimiento del modelo al condicionar el clasificador a darle menos importancia a la clase minoritaria [17].

Por otro lado, la aplicación de técnicas de transfer learning a la detección de malware no ha recibido tanta atención como su aplicación a otros problemas; generalmente relacionados con procesado de imágenes [18], [19].

Uno de los enfoques más usados consiste en la transformación del problema de detección de malware a la clasificación de imágenes que se resuelve usando redes de convolución preentrenadas. Este es el caso de [20], [21], [22]. La idea se basa en usar ciertas características de la muestra para construir una imagen, ya sean los bytes en crudo [20] o un volcado de memoria durante la ejecución [23]. La idea fundamental es que estas imágenes presentan un perfil de la muestra que será similar entre malware de la misma familia y diferente del de las benignas. El transfer learning en esta situación se refiere al uso de ciertas arquitecturas de redes neuronales preentrenadas sobre grandes conjuntos de datos de problemas diferentes como VGG-16 [24]. Este proceso, conocido como *fine-tuning* se basa en conservar las capas externas de la red preentrenada y sustituir las capas finales. En esta clase de arquitecturas las capas externas de la red realizan implícitamente un proceso de extracción de features; las capas más externas producen features más generales y las internas más específicas al problema. Como la red está preentrenada sobre millones de ejemplos, las features extraídas serán eficaces de forma general y se reducen los requerimientos de datos, ya que únicamente será necesario ajustar en proundidad las capas finales. Estos trabajos se han centrado principalmente en la clasificación de archivos ejecutables por la prevalencia del uso de bytes en crudo o código como fuente para la generación de imágenes. Además, para poder aplicar estas técnicas es necesario que la fuente de datos usada para la creación de imágenes sea lo suficientemente uniforme y discriminante, ya que solo se podrá aplicar *fine-tuning* en estas arquitecturas si el problema se transforma de forma eficaz a la clasificación de imágenes. Esto limita en principio el uso de diferentes tipos de features.

Por otro lado, el trabajo en [25], sí que propone un algoritmo de transfer learning más tradicional, basado en la transformación de los datos para minimizar las diferencias en la distribución de dominios fuente y objetivo. No obstante, el algoritmo está específicamente diseñado para la detección de tráfico malicioso usando la representación del tráfico en forma de bolsa de flujos, lo que limita su aplicabilidad a otros problemas y tipos de features.

En este trabajo evaluamos el uso de algoritmos de transfer learning más generales, que son aplicables a cualquier tipo de features, al problema de la detección malware en presencia de deriva conceptual y datos desbalanceados. La hipótesis que planteamos es que las técnicas de transfer learning pueden mejorar la predicción en estos casos porque sirven para sortear las diferencias en las distribuciones y permiten construir un modelo eficaz incluso en situaciones en las que se dispone de pocos datos en el dominio objetivo.

IV. Materiales y métodos

El objetivo de esta investigación es realizar una evaluación del rendimiento de técnicas de transfer learning sobre el problema de clasificación binaria de distinguir entre muestras benignas y maliciosas de diferentes horizontes temporales. Se busca determinar si las técnicas de transfer learning pueden servir para usar el conocimiento extraído de muestras antiguas para mejorar la capacidad de detección de malware de los modelos sobre muestras más recientes.

En primer lugar se describe el conjunto de datos usado; a continuación los algoritmos de transfer learning seleccionados; despues el procedimiento de los experimentos y las métricas que se van a considerar. Finalmente de describe el conjunto de features que se usará para caracterizar las muestras.

IV-A. Conjunto de datos

Se han recolectado de VirusShare [26] muestras de malware subidas en los años 2015, 2017, 2019 y 2020. En total se obtienen 196608 muestras de cada año para obtener un conjunto más o menos representativo de las muestra de cada intervalo temporal. Como el estudio está enfocado a la detección de malware en situaciones con datos limitados y desbalanceo de clases, se selecionan aleatoriamente un número limitado de muestras maliciosas de cada año para los experimentos.

Además, se obtienen 9996 muestras benignas obtenidas de Digital Corpora [27], [28] que se combinarán con las muestras maliciosas como se explicará más adelante. Estas muestras no están catalogadas por año, pero consideramos que esto no debería ser un gran problema ya que nuestro objetivo es la detección de malware y cabe esperar que la variabilidad en el tiempo del comportamiento de muestras benignas sea menor.

En los conjuntos de ficheros incluimos diferentes tipos de archivo: documentos de Word, archivos PDF, páginas HTML y ejecutables de Windows. El número de muestras de cada tipo se presenta en la *Tabla I*.

Tabla I: Distribución de tipos de ficheros

	Año	Tipo de archivo	Cantidad
Benignas		Word	2999
		HTML	2000
		PDF	2499
		Ejecutable	498
Maliciosas	2015	HTML	225
		Ejecutable	73
		PDF	2
	2017	HTML	225
		Ejecutable	74
		Word	1
	2019	HTML	190
		Ejecutable	62
		Word	32
		PDF	13
		Excel	3
	2020	HTML	132
		Ejecutable	62
		Word	47
		PDF	44
		Excel	15

IV-B. Modelos y algoritmos utilizados

En los experimentos se utilizan diferentes modelos y algoritmos de transfer learning. La optimización de los hiperparámetros de los modelos y los algoritmos de transfer learning se lleva a cabo de forma conjunta, usando el algoritmo de optimización SMAC [29].

Se usan 3 modelos de machine learning diferentes obtenidos de la librería scikit-learn [30]:

- Random Forest (RF). Se dejan los valores por defecto salvo para los hiperparámetros que se optimizan.
- K-nearest neighbors(KNN). Se dejan los valores por defecto salvo para los hiperparámetros que se optimizan.
- Perceptrón multicapa (MLP). Se fija *learning_rate* a *adaptive* y el número de neuronas en las capas ocultas a $[100, 100]$, es decir, dos capas ocultas con 100 neuronas cada una.

Se decide usar estos tres modelos porque cada uno de ellos sigue paradigmas diferentes de aprendizaje: Random Forest es un ensemble de modelos de tipo árbol de decisión; el perceptrón multicapa es un tipo de red neuronal y K-nearest neighbors clasifica mediante cercanía al resto de instancias. Por lo tanto se podrá obtener información sobre si algún tipo de modelo en particular se beneficia más o menos de la aplicación de transfer learning. En los experimentos se optimizan los hiperparámetros de cada uno de los tres modelos. En la *Tabla II* se muestran los hiperparámetros que se optimizan así como el intervalo de valores aceptado.

Tabla II: Valores posibles de los hiperparámetros de los modelos

Modelo	Hiperparámetros	Valores
KNN	número de vecinos	$[1, 100]$
	pesos	[uniformes, distancia]
	distancia	[euclidiana, chebyshev, manhattan]
RF	número de estimadores	$[1, 250]$
	criterio de división	[ganancia de información, gini]
MLP	número de iteraciones	$[1, 10]$
	alpha	$[0, 1]$
	función de activación	[sigmoide, tangente, ReLU]

Por otro lado, se aplican 5 algoritmos diferentes de transfer learning que han sido implementados en Python. El código está disponible en https://github.com/Descug01/Algoritmos.

- **TrAda** [31]. Se trata de una generalización de AdaBoost para aplicarse en problemas de transfer learning. Simplemente se ajusta el esquema de asignación de pesos de las instancias para disminuir la importancia del dominio fuente cuando se cometen demasiados errores. Al contrario que el resto de algoritmos usados, TrAda no transforma los datos, sino que genera un modelo de tipo ensemble.
- **CORAL** [32]. Este método se basa en igualar la covarianza de las features del dominio fuente a las del objetivo. Esto se logra blanqueando los datos (reduciendo la correlación entre los datos a 0) y luego induciendo la covarianza del dominio objetivo. No se modifica el espacio de features original.
- **DTS** [33]. Este método está basado en ideas del *manifold learning* [34], que se basa en proyectar los datos originales sobre un nuevo espacio de features conservando ciertas propiedades del espacio de features original como la cercanía entre instancias. En este algoritmo se busca una proyección tal que cada instancia puede reconstruirse como una combinación lineal de las vecinas.
- **DAE** [35]. Este método se basa en el uso de autoencoders. Los autoencoders [36] se suelen utilizar para obtener un nuevo espacio de features de dimensión reducida que conserva la máxima información de los datos originales. En este método simplemente se modifica la función objetivo del autoencoder para que la información conservada en el nuevo espacio de features sea común a ambos dominios.
- **TIT** [37]. Este método está basado en la formulación de análisis de componentes principales (PCA) [38]. La idea es proyectar los datos a un nuevo espacio de features, maximizando la varianza de los datos (como en PCA) pero añadiendo restricciones al problema de optimización basadas principalmente en que las instancias cercanas en el espacio de features original lo sigan siendo en la proyección.

Todos estos algoritmos tienen sus propios hiperparámetros que también se optimizan. Estos se muestran en la *Tabla III*. Los intervalos de los hiperparámetros se han fijado a partir de los resultados de unos experimentos preliminares. El significado de estos hiperparámetros está descrito en las referencias correspondientes a cada algoritmo.

Tabla III: Valores posibles de los hiperparámetros de los algoritmos de transfer learning

Algoritmo	Hiperparámetros	Intervalo de valores
DTS	número de iteraciones	$[5, 15]$
	α	$[0, 1]$
	β	$[0, 1]$
	λ	$[0, 1]$
CORAL	λ	$[0, 5]$
DAE	número de iteraciones	$[150, 800]$
	α	$[0, 1]$
	β	$[0, 1]$
	γ	$[0, 1]$
	función de activación	[sigmoide, tangente, ReLU]
	número de features	$[5, 280]$
TrAda	número de iteraciones	$[10, 150]$
TIT	número de iteraciones	$[1, 10]$
	γ	$[0, 1]$
	β	$[0, 1]$
	λ	$[0, 1]$
	número de vecinos	$[3, 50]$
	número de features	$[5, 280]$

IV-C. Experimentos

A continuación se describe el desarrollo de los experimentos. Como caso de referencia, el proceso descrito también se llevará a cabo sin usar ningún algoritmo de transfer learning para poder evaluar la posible mejora que pueda darse en los modelos.

En estos experimentos se busca simular una situación realista en la que el volumen de datos en el dominio objetivo (las muestras más recientes) es mucho menor y, por lo tanto, resultaría difícil construir un modelo eficaz usando únicamente datos nuevos. El volumen de datos del dominio objetivo es una décima parte del dominio fuente.

A continuación se describe el procedimiento de los experimentos.

- Se seleccionan las muestras de un año como dominio fuente y las muestras de año posterior como dominio objetivo. En total habrá 6 pares (fuente, objetivo) $\{(2015, 2017), (2015, 2019), (2015, 2020), (2017, 2019), (2017, 2020)$ y $(2019, 2020)\}$. A cada

uno de los dominios se le añaden muestras obtenidas del conjunto de muestras benignas, extraídas de forma aleatoria y sin remplazo para evitar que la misma muestra aparezca en ambos dominios. Al dominio fuente se le añaden 3700 muestras benignas. Para que el dominio objetivo sea de menor tamaño se añaden solo 370 muestras benignas y se seleccionan aleatoriamente 30 de las maliciosas. En total habrá 4000 muestras en el dominio fuente y 400 en el objetivo, ambas con una proporción de malware del 7.5 %.

- Se aplica el algoritmo de transfer learning concreto a los dominios fuente y objetivo y se obtienen los dominios transformados. La excepción es TrAda, que no realiza ningún tipo de transformación de los datos. Los hiperparámetros de los modelos y algoritmos de transfer learning se optimizan usando el algoritmo de optimización de hiperparámetros SMAC [29], con 50 iteraciones.
- Finalmente se usa el modelo para predecir las etiquetas de los datos transformados usando validación cruzada de 10 iteraciones y el coeficiente de correlación de Matthews (MCC) [39] como métrica. Se elige porque es una métrica más informativa que la tasa de acierto o F1 score en conjuntos de datos desbalanceados [40]. Se define en la *Ec. 1*

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (1)$$

donde TP, TN, FP y FN denotan verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Se aplica la validación cruzada únicamente al dominio objetivo, pero en el entrenamiento del modelo se añaden todos los datos del dominio fuente. Esto significa que las predicciones se realizarán únicamente sobre los datos del dominio objetivo con modelos entrenados usando los datos del dominio fuente como apoyo.

El proceso se repite para cada uno de los tres modelos de machine learning seleccionados (Sección IV-B).

IV-D. Descripción de los conjuntos de features utilizados

Para caracterizar las muestras usamos una variedad de conjuntos de features basados en la literatura. Principalmente buscamos que las features sean lo bastante descriptivas pero de baja complejidad, de forma que el proceso de extracción sea relativamente rápido. Usamos el mismo conjunto de features que en [41], ya que provee una caracterización del comportamiento de la muestra independiente del tipo de archivo.

Al igual que los propios modelos de machine learning, bastantes algoritmos de transfer learning requieren que las diferentes muestras tengan el mismo espacio de features. Por este motivo, optamos por usar un conjunto de features basado en el análisis dinámico, ya que provee una caracterización del comportamiento de la muestra independiente del tipo de archivo. La extracción de features procedentes de análisis estático requeriría de un conjunto de features diferente para cada tipo de archivo debido a las diferencias estructurales que existen ellos.

Para el análisis dinámico usamos la sandbox Cuckoo [42] v2.0.7; las features se extraen del reporte en formato JSON y la captura de tráfico de red en formato pcap que resultan del análisis. En total hay 1135 features. A continuación se describen cada uno de los conjuntos de forma individual.

IV-D1. API: El reporte JSON de Cuckoo contiene todas las llamadas realizadas a la API de Windows durante la ejecución de la muestra. En primer lugar, todas las llamadas a APIs se asignan a la categoría que les asigna la clasificación de Cuckoo (https://github.com/cuckoosandbox/cuckoo/wiki/Hooked-APIs-and-Categories), que agrupa la funciones según su funcionalidad: interacción con el sistema de ficheros, procesos, etc. Por lo tanto, la secuencia de llamadas de la muestra pasa a ser una secuencia de categorías de llamadas.

Como features se extraen la frecuencia de cada categoría en la secuencia de llamadas y la frecuencia de los 2-grams de categorías presentes en la secuencia. Un n-gram es una secuencia contigua de elementos de longitud n. Por ejemplo, los 2-grams de la secuencia $A_1A_2A_1A_3$ son A_1A_2, A_2A_1, A_1A_3. Hay 18 categorías y 18^2 posibles 2-grams, así que el total es de 342 features.

IV-D2. Red: Este conjunto de features se extrae de la captura de red en formato pcap proporcionada por Cuckoo. Las features extraídas están basadas en las propuestas en [43]. Estas features reflejan propiedades estadísticas del tráfico como el tamaño medio de paquete, el número de bytes enviados por segundo, etc. Se incluyen features tanto a nivel de paquete como de flujo de red, que es una secuencia de paquetes bidireccional, caracterizada por una tupla (IP origen, puerto origen, IP destino, puerto destino). La idea fundamental es que esta clase de features son aplicables a diferentes muestras ya que no están enfocadas en ningún protocolo concreto. Antes de extraer las features filtramos el tráfico de fondo propio de Windows que aparece en todas las muestras: protocolos NETBIOS, IGMP, LLMNR y paquetes sin capa IP.

En la *Tabla IV* se resumen las features extraídas. En total son 199 features.

IV-D3. Firmas: Ya que se usa Cuckoo para analizar las muestras, se decide usar las propias firmas proporcionadas por Cuckoo como features binarias (valor 1 si la firma se activa y 0 si no). Creemos que pueden ser útiles porque proporcionan una visión a alto nivel del comportamiento de la muestra. En total hay 563 firmas.

Por otro lado, y basándonos en [6], [44], también añadimos como features conteos simples que reflejan la interacción con el sistema de ficheros y el registro de Windows. Hay trabajos que se centran más en profundidad en este ámbito usando técnicas de procesado de texto [7], pero imponen demasiada complejidad adicional.

Las features extraídas de interacción con el sistema de ficheros son:

1. Frecuencias de ficheros creados, leídos, escritos, borrados y accedidos.
2. Número de extensiones diferentes de ficheros creados.
3. Número de ficheros creados en rutas predefinidas del sistema. Se consideran %APPDATA %, %TEMP % , and %PROGRAMFILES %.
4. Número de procesos creados al ejecutar un fichero creado.

Del registro se obtienen:

1. Frecuencias de claves creadas, leídas, escritas, borradas

Tabla IV: Descripción de las features de red extraídas

Nivel	Métrica	Media	Máximo	Mínimo	Desviación	Varianza	Mediana	Tercer cuartil	Suma	Conteo	Entropía
Paquete	Bytes por segundo										
	Paquetes por segundo										
	Tamaño de paquete	X	X	X	X	X	X	X	X		X
	Packets entrada/salida									X	
	Out/In ratio de bytes										
	Out/In ratio of paquetes										
	Tamaño de cabecera	X	X	X	X	X	X	X	X		X
	Tamaño de payload	X	X	X	X	X	X	X	X		X
	TTL	X	X	X	X	X	X	X	X		X
	IPs únicas									X	
	IPs únicas en HTTP									X	
	Tiempo entre paquetes	X	X	X	X	X	X	X	X		X
	% de HTTP, TCP, UDP y encriptado										
	flags TCP									X	
	Volumen tráfico encriptado	X	X	X	X	X	X	X	X		X
Flujo	Duración	X	X	X	X	X	X	X			X
	Bytes intercambiados en HTTP	X	X	X	X	X	X	X	X	X	X
	Out/In ratio de bytes HTTP										
	Out/In ratio de bytes en HTTP										
	Bytes intercambiados	X	X	X	X	X	X	X			X
	Paquetes intercambiados	X	X	X	X	X	X	X			X

y accedidas.

2. Número de claves de diferente tipo creadas y escritas. Se definen 7 tipos en https://docs.microsoft.com/en-us/windows/win32/sysinfo/registry-value-types.

En total hay 12 features obtenidas de interacción con el sistema de ficheros y 19 del registro. Unido a las features de las firmas de Cuckoo hay un total de 594 en este grupo.

V. Resultados

En primer lugar, hay que destacar que aunque es un algoritmo de transfer learning, TrAda es una modificación de un modelo de machine learning, por lo que el propio algoritmo realiza las predicciones. Por este motivo los resultados de TrAda serán los mismos para los diferentes modelos de machine learning.

En la *Tabla V* se muestran los coeficientes de correlación de Matthews obtenidos con los diferentes modelos y algoritmos de transfer learning al predecir sobre el dominio objetivo. Este es el principal caso de uso del transfer learning, así que los resultados en este caso determinarán su aplicabilidad al problema. Al referirse a los dominios, 2015-2017 quiere decir que se usa como dominio fuente las muestras de 2015 y como dominio objetivo las muestras de 2017.

Para los algoritmos de transfer learning TrAda, CORAL y DAE hay una mejora más o menos general en todos los dominios fuente y objetivo y modelos. DTS y TIT son la excepción. Aunque en algunos casos alcanzan resultados bastante competitivos, suelen estar por detrás de los otros tres algoritmos, e incluso a veces empeoran respecto al caso en el que no se aplica transfer learning. Una posibilidad es que los resultados de TIT y DTS sean un producto del fenómeno de la transferencia negativa [45], en el que la aplicación de un algoritmo de transfer learning empeora las predicciones. No obstante, el resto de algoritmos tienen buen rendimiento. Nuestra hipótesis es que TIT y DTS son más sensibles a la configuración de hiperparámetros aplicada, por lo que sería necesario invertir más tiempo en la optimización de hiperparámetros o extender el intervalo de valores aceptado de estos para obtener mejores resultados. Para TIT, el MCC

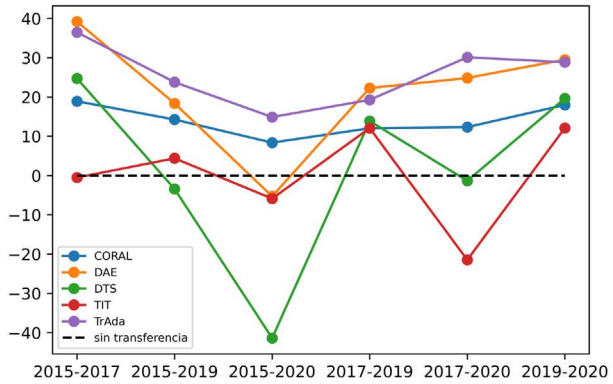
Tabla V: MCC obtenido en la validación cruzada

Dominios	Modelo	No transfer	TrAda	CORAL	DAE	DTS	TIT
2015-2017	KNN	0.704	0.961	0.837	0.980	0.878	0.700
	MLP	0.557	0.961	0.922	0.980	0.472	0.483
	RF	0.897	0.961	0.936	0.965	0.632	0.815
2015-2019	KNN	0.764	0.946	0.873	0.905	0.738	0.798
	MLP	0.607	0.946	0.912	0.876	0.632	0.676
	RF	0.905	0.946	0.883	0.930	0.673	0.706
2015-2020	KNN	0.819	0.961	0.888	0.776	0.480	0.771
	MLP	0.543	0.961	0.917	0.776	0.550	0.625
	RF	0.922	0.961	0.922	0.825	0.682	0.784
2017-2019	KNN	0.805	0.961	0.902	0.985	0.917	0.903
	MLP	0.576	0.961	1	1	0.616	0.447
	RF	0.847	0.961	0.941	1	0.762	0.980
2017-2020	KNN	0.738	0.961	0.829	0.922	0.728	0.580
	MLP	0.527	0.961	0.861	0.891	0.734	0.590
	RF	0.839	0.961	0.897	0.859	0.732	0.859
2019-2020	KNN	0.761	0.980	0.897	0.985	0.910	0.853
	MLP	0.522	0.980	0.965	0.946	0.625	0.654
	RF	0.917	0.980	0.917	1	0.701	0.876

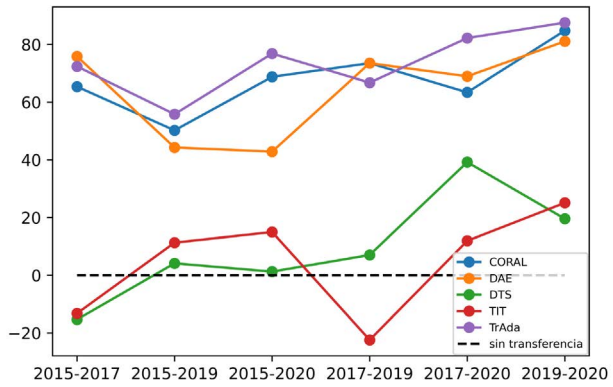
alcanzado en los dominios 2017-2019 con el modelo RF hace pensar que con una configuración adecuada, también podría ser competitivo en el resto de casos.

En cuanto a los modelos, la mejora es más marcada para MLP, seguido de KNN y finalmente RF. Obviamente, cuanto mayor sea el MCC del modelo sin aplicar transfer learning, menor será el posible margen de mejora, pero la aplicación de transfer learning hace que el rendimiento de MLP, que tiene los peores resultados sin transfer learning, sea competitivo con el de RF, que tiene los mejores resultados sin transfer learning. Esto parece indicar que los modelos lineales podrían beneficiarse más de la aplicación de algoritmos de transfer learning.

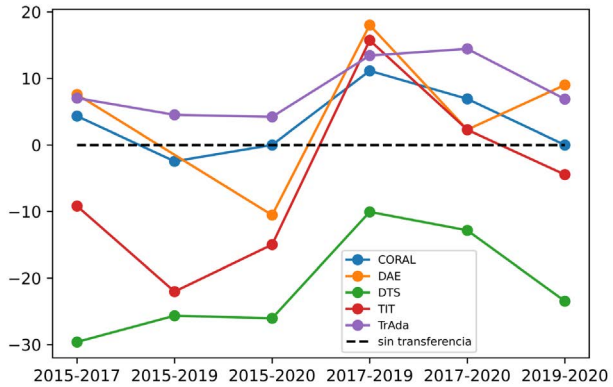
Para visualizar mejor el rendimiento de los algoritmos de transfer learning, en la *Fig. 1* se muestra, para cada modelo, el porcentaje de mejora de MCC de los algoritmos de transfer learning respecto al caso en el que no se usa transfer learning. Existen diferencias entre los diferentes dominios, quizás debido a una configuración de hiperparámetros más o menos favorable, pero los niveles de mejora suelen ser similares para cada algoritmo de transfer learning. En cualquier caso,



(a) KNN



(b) MLP



(c) RF

Figura 1: Porcentaje de incremento de MCC en cada modelo

los niveles de mejora son considerables. Con los mejores algoritmos de transfer learning se alcanzan mejoras de en torno a un 30 % con KNN, 60 % con MLP y un 10 % con RF, salvo en algunos casos que se comentan a continuación.

En la *Fig. 1c* se observa que en ciertos casos algoritmos de transfer learning como CORAL y DAE, que son de los que mejores resultados obtienen en general, empeoran a RF sin aplicación de transfer learning. Este hecho plantea algunas cuestiones. Por un lado es posible que la aplicación de transfer learning sea menos eficaz en modelos de tipo árbol. Estos modelos realizan predicciones estableciendo umbrales sobre los valores de las features y los algoritmos de transfer learning modifican la distribución original de los datos y, en algunos casos, proyectan los datos a un espacio de features distinto. Esto podría eliminar algunas features particularmente informativas para los modelos de tipo árbol. Por otro lado, y siguiendo la misma lógica, podría ser un caso de transferencia negativa. RF alcanza muy buenos ajustes sin transfer learning, así que cabe la posibilidad de que los algoritmos de transfer learning, al reducir las diferencias en las distribuciones entre fuente y objetivo, también eliminen información de los datos originales. Si el modelo ya es lo suficientemente competitivo sin transfer learning, su aplicación podría no tener ningún efecto beneficioso o incluso producir transferencia negativa.

VI. Conclusión y trabajo futuro

En este trabajo se ha llevado a cabo una evaluación de varios algoritmos de transfer learning, que han sido implementados en Python, al problema de detección de malware. Los experimentos se han llevado a cabo con datos desbalanceados y con muestras procedentes de diferentes horizontes temporales para determinar si la aplicación de técnicas de transfer learning permiten mejorar las predicciones sobre nuevas muestras aprovechando el conocimiento extraído de muestras antiguas.

Los resultados indican que en general la aplicación de algoritmos de transfer learning puede proporcionar una mejora significativa en la capacidad predictiva de los modelos en casos en los que se dispone de insuficientes datos para construir un modelo eficaz en un dominio objetivo. No obstante, no todos los algoritmos de transfer learning son igual de eficaces ni las mejoras las mismas en todos los modelos. Por ejemplo, con Random Forest no hay demasiada mejora en algunos casos. Incluso en este caso algunos algoritmos incrementan el MCC en torno a un 10 %, un 30 % para KNN y 60 % para MLP. En términos absolutos esto implica que usando los mejores algoritmos como TrAda o DAE se puede pasar de un MCC de en torno a 0.74 o 0.56 con KNN y MLP respectivamente a un MCC de 0.93, lo cual es una mejora sustancial.

Como trabajo futuro se pretende extender el estudio para evaluar el rendimiento de las técnicas de transfer learning en diferentes circunstancias para determinar cuáles son las condiciones para su aplicación eficaz. La definición del problema plantea una situación en la que los datos del dominio objetivo son insuficientes para entrenar un modelo eficaz, pero podría darse el caso de que la aplicación de técnicas de transfer learning permita mejorar el modelo aún cuando se disponga de un volumen de datos suficiente en el dominio objetivo. Por otro lado, también se pretende estudiar empíricamente su rendimiento computacional sobre un conjunto de datos más amplio para determinar la escalabilidad de estas técnicas.

Agradecimientos

Este trabajo se ha llevado a cabo con el apoyo del Instituto Nacional de Ciberseguridad (INCIBE) bajo el contrato Art. 83 con clave X54.

Referencias

[1] D. Gibert, C. Mateu, and J. Planes, "The rise of machine learning for detection and classification of malware: Research developments, trends and challenges," *J. of Netw. and Comput. Appl.*, vol. 153, p. 102526, Mar. 2020.

[2] Malwarebytes Labs, "2020 state of malware report," 2020. [Online]. Available: https://resources.malwarebytes.com/files/2020/02/2020_State-of-Malware-Report.pdf

[3] J. a. Gama, I. Žliobaitundefined, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, 2014. [Online]. Available: https://doi.org/10.1145/2523813

[4] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct 2010.

[5] M. Rhode, P. Burnap, and K. Jones, "Early-stage malware prediction using recurrent neural networks," *Comput. & Secur.*, vol. 77, pp. 578–594, Aug. 2018.

[6] W. Han, J. Xue, Y. Wang, Z. Liu, and Z. Kong, "Malinsight: A systematic profiling based malware detection framework," *J. of Netw. and Comput. Appl.*, vol. 125, pp. 236–250, Jan. 2019.

[7] J. Stiborek, T. Pevný, and M. Rehák, "Multiple instance learning for malware classification," *Expert. Syst. with Appl.*, vol. 93, pp. 346–357, Mar. 2018.

[8] Z. Yuan, Y. Lu, and Y. Xue, "Droiddetector: Android malware characterization and detection using deep learning," *Tsinghua Sci. and Technol.*, vol. 21, no. 1, pp. 114–123, Feb. 2016.

[9] K. Tian, D. Yao, B. G. Ryder, G. Tan, and G. Peng, "Detection of repackaged android malware with code-heterogeneity features," *IEEE Trans. on Dependable and Secur. Comput.*, vol. 17, no. 1, pp. 64–77, 2020.

[10] Verizon, "2019 data breach investigations report," 2020. [Online]. Available: https://enterprise.verizon.com/resources/reports/2019-data-breach-investigations-report.pdf

[11] R. Jordaney, K. Sharad, S. Dash, Z. Wang, D. Papini, I. Nouretdinov, and L. Cavallaro, "Transcend: Detecting concept drift in malware classification models," in *26th USENIX Security Symposium*, 2017.

[12] A. Pektas, E. N. Pektas, and T. Acarman, "Mining patterns of sequential malicious apis to detect malware," *Int. J. of Netw. Secur. & Its Appl.*, vol. 10, no. 4, pp. 1–9, Jul. 2018.

[13] M. Conti, G. Rigoni, and F. Toffalini, "Asaint: A spy app identification system based on network traffic," in *Proc. of the 15th Int. Conf. on Availab., Reliab. and Secur.* New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–8.

[14] N. Kumar, S. Mukhopadhyay, M. Gupta, A. Handa, and S. K. Shukla, "Malware classification using early stage behavioral analysis," in *Proc. of 14th Asia Jt. Conf. on Inf. Secur.*, Aug. 2019, pp. 16–23.

[15] A. Narayanan, M. Chandramohan, L. Chen, and Y. Liu, "Context-aware, adaptive, and scalable android malware detection through online learning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 1, no. 3, pp. 157–175, 2017.

[16] Z. Ma, X. Zhang, P. Li, D. Ye, and B. Ling, "The concept drift problem in android malware detection and its solution," *Security and Communication Networks*, 2017.

[17] W. P. Kegelmeyer, K. Chiang, and J. Ingram, "Streaming malware classification in the presence of concept drift and class imbalance," in *2013 12th International Conference on Machine Learning and Applications*, vol. 2, 2013, pp. 48–53.

[18] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, vol. 43, pp. 635–640, 2020.

[19] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 9, pp. 7316–7325, 2019.

[20] D. Vasan, M. Alazab, S. Wassan, H. Naeem, B. Safaei, and Q. Zheng, "Imcfn: Image-based malware classification using fine-tuned convolutional neural network architecture," *Computer Networks*, vol. 171, p. 107138, 2020.

[21] E. Rezende, G. Ruppert, T. Carvalho, F. Ramos, and P. de Geus, "Malicious software classification using transfer learning of resnet-50 deep neural network," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017, pp. 1011–1014.

[22] E. Rezende, G. Ruppert, T. Carvalho, A. Theophilo, F. Ramos, and P. d. Geus, "Malicious software classification using vgg16 deep neural network's bottleneck features," in *Information Technology - New Generations*, S. Latifi, Ed. Cham: Springer International Publishing, 2018, pp. 51–59.

[23] D. Nahmias, A. Cohen, N. Nissim, and Y. Elovici, "Deep feature transfer learning for trusted and automated malware signature generation in private cloud environments," *Neural Networks*, vol. 124, pp. 243–257, 2020.

[24] K. Simonyan and A. Zisserman, "Vgg-16," *arXiv Preprint*, 2014.

[25] K. Bartos and M. Sofka, "Robust representation for domain adaptation in network security," in *Machine Learning and Knowledge Discovery in Databases*. Cham: Springer International Publishing, 2015, pp. 116–132.

[26] Corvus Forensics, "Virusshare," 2011, accessed: Dec. 2019. [Online]. Available: https://virusshare.com/

[27] S. Garfinkel, P. Farrell, V. Roussev, and G. Dinolt, "Bringing science to digital forensics with standardized forensic corpora," *Digit. Investig.*, vol. 6, pp. S2–S11, Sep. 2009.

[28] Digital Corpora, "Govdocs1 — (nearly) 1 million freely-redistributable files," 2018. [Online]. Available: https://digitalcorpora.org/corpora/files

[29] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *Proc. of 5th Conf. in Learn. and Intell. Optim.*, C. A. C. Coello, Ed., Jan. 2011, pp. 507–523.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[31] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 1855–1862.

[32] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, pp. 2058–2065. [Online]. Available: http://dl.acm.org/citation.cfm?id=3016100.3016186

[33] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, "Discriminative transfer subspace learning via low-rank and sparse representation," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 850–863, Feb 2016.

[34] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[35] F. Zhuang, X. Cheng, P. Luo, S. J. Pan, and Q. He, "Supervised representation learning: Transfer learning with deep autoencoders," in *Proceedings of the 24th International Conference on Artificial Intelligence*, ser. IJCAI'15. AAAI Press, 2015, pp. 4119–4125. [Online]. Available: http://dl.acm.org/citation.cfm?id=2832747.2832823

[36] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, no. 2, pp. 233–243, 1991.

[37] J. Li, K. Lu, Z. Huang, L. Zhu, and H. Tao Shen, "Transfer independently together: A generalized framework for domain adaptation," *IEEE Transactions on Cybernetics*, vol. PP, pp. 1–12, 04 2018.

[38] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998. [Online]. Available: https://doi.org/10.1162/089976698300017467

[39] B. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophys. Acta (BBA) - Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.

[40] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC Genom.*, vol. 21, no. 1, Jan. 2020.

[41] D. Escudero-García and N. deCastro García, "Optimal feature configuration for dynamic malware detection," *Computers & Security*, 2021.

[42] Cuckoo Foundation, "Cuckoo sandbox," 2012. [Online]. Available: https://cuckoosandbox.org/

[43] D. Bekerman, B. Shapira, L. Rokach, and A. Bar, "Unknown malware detection using network traffic classification," in *Proc. of 2015 IEEE Conf. on Commun. and Netw. Secur.*, Sep. 2015, pp. 134–142.

[44] A. Mohaisen, O. Alrawi, and M. Mohaisen, "Amal: High-fidelity, behavior-based automated malware analysis and classification," *Comput. & Secur.*, vol. 52, pp. 251–266, Jul. 2015.

[45] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, May 2016. [Online]. Available: https://doi.org/10.1186/s40537-016-0043-6

Sesión de Investigación A2: Detección de intrusiones y gestión de anomalías II
http://doi.org/10.18239/jornadas_2021.34.12

Temporal graph-based approach for behavioural entity classification

Francesco Zola *†‡ , Lander Segurola* , Jan Lukas Bruse *§, Mikel Galar †¶
* Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Paseo Mikeletegi 57, 20009 Donostia/San Sebastian, Spain
{fzola, lsegurola, jbruse}@vicomtech.org
†Institute of Smart Cities, Public University of Navarre, 31006 Pamplona, Spain
mikel.galar@unavarra.es
ORCID: ‡ 0000-0002-1733-5515, §0000-0002-5774-1593, ¶0000-0003-2865-6549

***Abstract*—Graph-based analyses have gained a lot of relevance in the past years due to their high potential in describing complex systems by detailing the actors involved, their relations and their behaviours. Nevertheless, in scenarios where these aspects are evolving over time, it is not easy to extract valuable information or to characterize correctly all the actors.**

In this study, a two phased approach for exploiting the potential of graph structures in the cybersecurity domain is presented. The main idea is to convert a network classification problem into a graph-based behavioural one. We extract these graph structures that can represent the evolution of both normal and attack entities and apply a temporal dissection approach in order to highlight their micro-dynamics. Further, three clustering techniques are applied to the normal entities in order to aggregate similar behaviours, mitigate the imbalance problem and reduce noisy data. Our approach suggests the implementation of two promising deep learning paradigms for entity classification based on Graph Convolutional Networks.

***Index Terms*—Cybersecurity analysis, Behavioural classification, Temporal graph analysis, Clustering, Graph-based structure**

Tipo de contribución: *Investigación en desarollo*

I. Introduction

Cybersecurity has become a critical aspect for many companies as a vulnerability or a breach in its infrastructure can generate a considerable loss of value, in economical, reputational, digital, psychological and societal terms [1]. The increasing amount of cyber threats and their significant impacts, has led to increased efforts for preventing and reducing their risks [2]. In this sense, machine learning and derived techniques such as deep learning are promising means to obtain new insights from available data in order to quantify cyber risks and to optimize cybersecurity operations [3].

These techniques are principally used for performing anomaly detection [4], an operation that can be used for risk detection as well as a preventive approach. In particular, anomaly detection has shown very promising results analyzing data as graphs [5]. Graph structures help integrate both structured and unstructured data in a representation of entities and the relationships among them.

Nevertheless, in real use cases, entity behaviour can evolve such as entities disappearing, emerging or simply changing their dynamics. In these cases, considering a static graph is not sufficient to describe a complex system, not least as with this monolithic approach small interactions between entities can be obscured by more frequent ones - generating a loss of classification quality. In such a skewed scenario, machine learning models tend to focus on macro-dynamics and may falsely assign one static behaviour for each entity as generated by the overall status.

In this work, we propose a two phased approach for addressing these issues. The first phase, called *Manipulation phase*, is aligned with the Extract-Transform-Load (ETL) process [6], in particular, in this phase graph-based structures are extracted from the initial network traffic dataset defining entities, edges and entity behaviours using a temporal dissection approach. Then, still in the first phase, a clustering operation is applied for reducing noisy data and addressing the class imbalance problem. In the second phase, called *Learning phase*, deep learning models are trained with these graph data in order to detect attack behavioural entities. However, due to the size limitation, this study is focused in presenting just the first phase, i.e. it introduces the process for extracting graph information and for reducing the problem-space within, applying and comparing three clustering techniques.

II. Methodology

A. Problem description

In several cases, when the aim is to perform a network traffic classification, the most common and straightforward idea is to consider the information as a time series and then apply Artificial Intelligence paradigms for classifying these flows in order to discover attack patterns. In this study, a two phased approach based on converting this time series analysis into a graph-based behavioural classification is proposed.

During the first phase, the initial dataset is manipulated applying multiple operations that allow us transform it and extract graph information. In common machine learning problem, this phase is also known as Extract-Transform-Load (ETL) process [6], however, in our methodology, we have renamed it as *Manipulation phase*, since it substitute Load operations with Temporal dissection. In fact, the main idea is to split the initial dataset in fixed time interval called ”snapshot”, and, in each of them, extract graph-based structures defining entities and edges. In this sense, it is important to perform this temporal dissection operation in order to uncover micro-dynamics and highlight how the entities relations change within the time. Once the graph structures are ready, a clustering algorithm is applied in order to aggregate similar behaviours, i.e. similar entities, for reducing the overall amount of information. This operation is very important, especially when the initial dataset is characterized by noisy

data or it is affected by a strong imbalanced problem. In these cases, applying the clustering only over the majority class helps to reduce those problems. In this research, 3 clustering techniques are compared: Density-Based Spatial Clustering of Applications with Noise (DBSCAN [7]), Ordering Points To Identify the Cluster Structure (OPTICS [8]) and Hierarchical DBSCAN (HDBSCAN [9]) All these operations allow to create the appropriate input for the next phase (*Learning phase*), which is focused on training 2 graph deep learning models, both belonged to the Graph Convolutional Network (GCN) family, for classifying the node behaviours and detect the attack ones.

It is important to remark that this work is not focused on classifying or predicting directly the node's communication as presented in [10], neither on developing another framework based on graph deep learning models [11], but the key idea is to convert network flow classification into a behavioural entity task, and to investigate how to manipulate the initial data for generating new insights and improve the final classification. To the best of our knowledge, this is the first work that combine temporal graph dissection, clustering operation and planning the usage of graph deep learning for detecting attack entities in cybersecurity domain.

B. Behavioral Node Identification

In order to extract and define behavioural entities within the traffic dataset a pre-processing operation called Behavioral Node Identification is introduced. In particular, knowing that each record of the dataset represents a communication between two nodes, we map that communication as an edge in the graph. These edges connect an entity source and an entity destination defined respectively by the combination of the source IP (srcip) and source port (srcport), and the destination IP (dstip) and the destination port (dstport) given as features for the considered record.

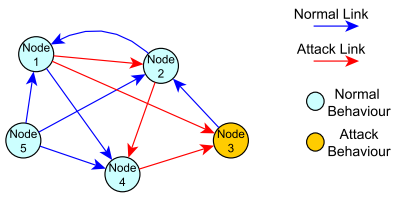


Fig. 1: Behavioral Node Identification example

Once obtained the graph, it is important to assign a label to each node in order to define its behavior in the network. In this sense, the information provided by the labelled records of the network traffic dataset are used to label the nodes' behaviour in the network. The possible nodes' behaviours are two: a normal behaviour identified with the labelled 0 and a attack behaviour identified with the label 1. Each node behaviour is defined by analyzing the nodes' connections, both the incoming and outgoing (Figure 1). If the attack connections represent the majority of the node's connections, the node takes the label of attack behaviour (1), as for example node 3 in the Figure 1. On the other hand, the node takes the label of normal behaviour (0), as for example node 1 in Figure 1. In case of draw, the node takes the label of normal behaviour (0), as for example node 4.

C. Temporal graph

The Node Behavioural Identification can be applied considering the whole network traffic dataset for creating an unique static and monolithic graph, however, with this approach all the dynamics and small interaction between the entities will be overwhelmed by the more frequently and more heavy - in terms of exchanged data - ones. Furthermore, this approach increases the amount of data in the graph requiring more computational resources, making hard the real application of the solution. For example, if a model is trained with a monolithic graph created with ~ 10 hours data, for the testing phase, it needs of a comparable structure, i.e. a graph obtained in other ~ 10 hours data. This can compromise the usability of the trained model, since in attack detection the timing is fundamental in order to mitigate them promptly and applying countermeasures and avoid the worsening of the situation.

To address these issues, and learn the micro behaviours in the network and detecting the small dynamics within, in this research, the initial dataset is divided in fixed temporal intervals (or temporal snapshot), and in each one of it, it is performed an operation of Node Behaviour Identification, as described in the Section II-B. Basically, all the communications in a temporal snapshot are analyzed in order to create a photograph of the network status through a relational graph where the behaviours' entities and their relations are drawn.

D. Clustering Algorithms

In cybersecurity domain, the network traffic datasets are usually characterized by class imbalanced problem. In fact, it is easier to find and generate traffic related with normal activities rather than attack connections. This imbalance problem in the population, between normal and attack entities, can generate loss of classification quality when supervised classification algorithms are used, since the model will be trained with a skewed dataset favoring the classification of a the major-represented class over the under-represented one [12]. In this scenario, in order to reduce the amount of information in each temporal snapshot, a clustering algorithm is used. The idea is to apply this operation in order to aggregate entities with similar behaviours and homogenize the final population. In particular, we propose to apply this operation just over the normal behavioural entities in order to reduce their population.

A variety of clustering algorithms can be separated based on the criteria used to create the groups [13], the more common are the hierarchical clustering, the distribution-based clustering, the density-based clustering and the centroid-based clustering. In this work, the analysis is focused over the density-based clustering which represents a group of algorithms where the main parameter to define the clusters is the density difference between zones of the space [14]. This cluster's category uses unsupervised learning methods that identify distinctive groups in the data, based on the idea that a cluster in a data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density [14]. The data points in the separating regions of low point density are typically considered noise/outliers [15].

In particular, in this analysis, DBSCAN, OPTICS and HDBSCAN techniques are used. The first 2 algorithms define

dense regions using a minimum number of points that must belong to the cluster (*minPts*) and the maximum distance from one point to another for both to be considered neighbours (ϵ). On the other hand, DBSCAN paradigm intends to automatically find a clustering that gives the best stability over ϵ.

After the application of the cluster algorithm, the shape of each temporal graph changes, and so, the corresponding adjacency matrix changes. This aggregation promotes the creation of new clustered entities in which their behaviours are computed by combining (averaging) the single behaviour of the aggregated entities.

III. Validation

A. Experiment

In this work, UNSW-NB15 dataset[1] is used as input for the *Manipulation phase*. As described in [16], this dataset was created with the aim to improve the existing benchmark datasets, which are not able to describe the comprehensive representation of network traffic and attack scenarios. The UNSW-NB15 dataset contains real normal and synthetic abnormal network traffic generated in the synthetic environment of the University of New South Wales (UNSW) cyber security lab. The UNSW-NB15 dataset is characterized by 9 attack families and normal traffic generated in two distinct capture-day. In particular, in this study, the connections are labelled as 0 if they are normal traffic or 1 if they are attack traffic.

The *Manipulation phase* starts fixing the value for the temporal dissection operation, in particular, for this work, 600 seconds (10 minutes) is chosen. Then, from each of the extracted temporal snapshots a behavioural graph is created in order to detect the entities and their relations within the network. For each node, a feature vector that describes its behaviour, is extracted and it is used as input for the clustering operation. This clustering process allows us to reduce the amount of data and to mitigate the imbalance problem as presented in Section II-C. In particular, 3 clustering algorithms - OPTICS, DBSCAN and HBDSCAN - are applied separately, studying how their parameters affect their effectiveness. In particular, the OPTICS and the DBSCAN algorithm are implemented both with *minPts* value of 2 and with 3 different ϵ values: 0.2, 0.5 and 0.8. The HDBSCAN instead is applied again with *minPts* equals 2 and with a minimum size of clusters of 5, and a euclidean metric to compute the distance between the instances is used.

B. Results

Choosing fixed temporal intervals of 600 seconds, 147 temporal snapshots are generated, as shown in Figure 2. Each temporal snapshot is characterized by a very large number of nodes, where the amount of normal entities overwhelm the amount of attack one. In fact, in all the extracted 147 temporal snapshots, the average number of normal nodes in a snapshot is more than $17,000$, for an overall amount of $2,583,605$, while the average attack population is about 545, for a global value of $80,174$. Furthermore, in several temporal snapshots there are not attack nodes, as for example between the 17th and the 77th snapshots.

[1]https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/

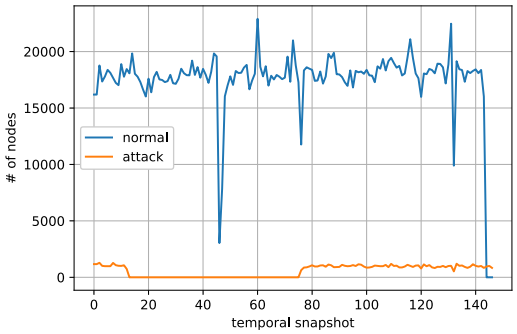


Fig. 2: Initial population in each temporal size

Method	ϵ	normal entities after clustering	clustered normal entities over the whole dataset
OPTICS	0.2	369,931	82.19 %
	0.5	447,026	84.79 %
	0.8	479,938	85.69 %
DBSCAN	0.2	141,935	63.90 %
	0.5	137,903	63.24 %
	0.8	127,346	61.37 %
HDBSCAN	-	176,645	68.78 %

TABLE I: Clustering effects

Table I shows the overall clustered population generated by each algorithm. This overall value is computed by summing the number of normal nodes in all temporal snapshots after the execution of the cluster algorithms, where the low point density regions are discarded, as they are considered as noise/outliers. It should be noted that from an initial overall amount of $2,583,605$ normal nodes, the cluster algorithms reduce them in a range between $479,938$ and $127,346$, i.e. reducing the class of about 81% - 95% (Figure 3). Nevertheless, for the OPTICS datasets, the new normal population still represent more than 80% of the whole population. These cluster methods are not applied over the attack class, as explained in Section II-C, so the final attack population is invariant ($80,174$ nodes)

Table I shows also that a fixed value of ϵ parameter generates different results according to the considered clustering algorithm. In particular, increasing the ϵ value increases the ability of the OPTICS algorithm to aggregate the samples in dense regions, reducing the number of noise/outliers. The same highest value of ϵ produces a reverse effect in the DBSCAN, in fact, it decreases its ability to create dense regions, increasing the number of noise/outliers and reducing the final population. This divergence in their behaviours is due to their structures, in fact, whereas the DBSCAN uses a fixed radius for detecting the density regions, OPTICS allows the search radius to dynamically expand itself [8].

C. Learning guidelines

Once the *Manipulation phase* is ended, the information is ready for training learning models and achieve the attack behavioural classification task related with the *Learning phase*. In particular, in order to fully exploit the extracted graph

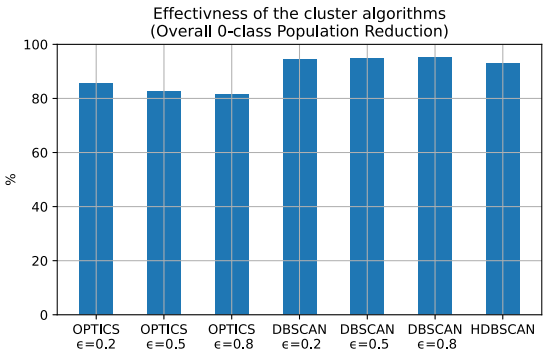


Fig. 3: Clustered Population

relations highlighted in the first phase, it is planned to use the Graph Convolutional Network. These models, introduced in [17], are able to learn the local and global structural patterns of a graph through a convolution operation as well as happens in the Convolution Neural Network (CNN).

The convolution operation can be defined through the Equation 1, where $x \in \mathbb{R}^n$ represents a scalar vector (the scalar vector for every node in the graph), U is the matrix of eigenvectors of the Laplacian of the graph and $g_\theta(\Lambda)$ represents the filter in the Fourier domain. Nevertheless, solving this equation can be computationally complex and unreachable, even more prohibitively expensive for large graphs.

$$y = g_\theta * x = U g_\theta(\Lambda) U^T x \tag{1}$$

In this sense, in [18] a promising solution to the Equation 1 is obtained parametrizing the term $g_\theta(\Lambda)$ as a polynomial function that can be computed recursively. In particular, Chebyshev polynomials with k degree are used. Kipf et al. [17] demonstrated that a good approximation can be reached by truncating the Chebichev polynomial to get a linear polynomial (k = 1) and by performing a renormalization trick in order to avoid numerical instabilities and vanishing gradients.

For the validation of the *Learning phase*, the idea is to exploit the highlighted graph-based relations applying 2 GCNs models, the simple one introduced in [17], and its higher version based on a different approximation of the Chebychev polynomials ($k > 1$), in order to highlight how the relational information affect the final classification.

IV. Conclusions

In this research, a two phased approach for converting a network traffic analysis into a behavioural entity classification task is presented. In particular, this work introduces and validates the first phase of this approach called *Manipulation phase*, and draws several guidelines for the next phase, called *Learning phase*

The main idea of the first phase is to understand and manipulate correctly the input data regardless the chosen classification model. In particular, a method for converting the network traffic time series into graph-based structures is presented, as well as a temporal dissection operation used to uncover micro-dynamics that in a complex system can be overwhelmed by the most recurrent ones. Furthermore, 3 clustering techniques - OPTICS, DBSCAN and HDBSCAN - for reducing noisy data and aggregate similar behaviour, are tested and compared using a relevant cybersecurity dataset (UNSW15). A preliminary study shows how their effectiveness changes according to several parameters, as well as the DBSCAN is more proned to create regions with lower density that are considered as noise/outliers. However, it should be noted that these results are strictly related with the chosen time interval ($600s.$), and in future works will be interesting apply other temporal values - for example $300s.$ and $900s.$ - for evaluating their impact in the *Manipulation phase*.

Acknowledgement

This work has been partially supported by the Basque Country Government under the ELKARTEK program, project TRUSTIND (KK-2020/00054).

References

[1] I. Agrafiotis, J. R. Nurse, M. Goldsmith, S. Creese, and D. Upton, "A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate," *Journal of Cybersecurity*, vol. 4, no. 1, p. tyy006, 2018.

[2] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, p. 20, 2019.

[3] I. H. Sarker, A. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *Journal of Big Data*, vol. 7, no. 1, pp. 1–29, 2020.

[4] S. Omar, A. Ngadi, and H. H. Jebur, "Machine learning techniques for anomaly detection: an overview," *International Journal of Computer Applications*, vol. 79, no. 2, 2013.

[5] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data mining and knowledge discovery*, vol. 29, no. 3, pp. 626–688, 2015.

[6] S. K. Bansal and S. Kagemann, "Integrating big data: A semantic extract-transform-load framework," *Computer*, vol. 48, no. 3, pp. 42–50, 2015.

[7] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[8] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: ordering points to identify the clustering structure," *ACM Sigmod record*, vol. 28, no. 2, pp. 49–60, 1999.

[9] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2013, pp. 160–172.

[10] T. Oba and T. Taniguchi, "Graph convolutional network-based suspicious communication pair estimation for industrial control systems," *arXiv preprint arXiv:2007.10204*, 2020.

[11] J. Jiang, J. Chen, T. Gu, K.-K. R. Choo, C. Liu, M. Yu, W. Huang, and P. Mohapatra, "Anomaly detection with graph convolutional networks for insider threat and fraud detection," in *MILCOM 2019-2019 IEEE Military Communications Conference (MILCOM)*. IEEE, 2019, pp. 109–114.

[12] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *2008 Fourth international conference on natural computation*, vol. 4. IEEE, 2008, pp. 192–201.

[13] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, 2015.

[14] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011.

[15] J. Sander, *Density-Based Clustering*. Boston, MA: Springer US, 2010, pp. 270–273. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_211

[16] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 military communications and information systems conference (MilCIS)*. IEEE, 2015, pp. 1–6.

[17] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[18] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in neural information processing systems*, 2016, pp. 3844–3852.

http://doi.org/10.18239/jornadas_2021.34.13

Diseño de un sistema de correlación basado en software libre para la detección de anomalías en el campo de la ciberseguridad

Beatriz Esteban-Navarro, Xavier A. Larriva-Novo, Víctor A. Villagrá
Universidad Politécnica de Madrid (UPM). DIT, ETSI Telecomunicación. Avda. Complutense 30. 28040 Madrid
beatriz.esteban.navarro@alumnos.upm.es, xavier.larriva.novo@upm.es, victor.villagra@upm.es

***Resumen*- A la hora de desplegar un sistema de gestión de información y eventos de seguridad (SIEM) hay que tener en cuenta diferentes factores como, por ejemplo, el número de equipos a monitorizar o la granularidad que se pretende conseguir. Sin embargo, el aspecto más importante a considerar es el presupuesto disponible para invertir en herramientas comerciales o para adecuar una solución existente aprovechando plataformas *open source*. Este artículo propone un sistema para la correlación de eventos de bajo coste, basado en ElasticSearch y ElastAlert, marco que permite generar, mediante el uso de intervalos temporales, diferentes reglas, alertas y acciones en tiempo real. Gracias a esta solución *open source*, altamente personalizable, es posible identificar la correlación de anomalías a partir de eventos con origen diverso, incrementando así la capacidad para detectar amenazas en cualquier red en la que se desee implementar.**

***Index Terms*- Ciberseguridad, Open source, SIEM, ElasticSearch, ElastAlert, Correlación, Anomalía**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

En la actualidad, contar con un sistema de gestión de información y eventos de seguridad (SIEM en inglés) es definitivo a la hora de detectar amenazas en una red. Esta tecnología nace de la combinación de funciones SEM (*Security Event Management)* para monitorizar, almacenar, identificar y notificar sobre amenazas en tiempo real, y SIM (*Security Information Management)* para recopilar datos a largo plazo haciendo posible realizar análisis y generar reportes. Esta combinación permite además la correlación de eventos, que aumenta las capacidades de detección permitiendo una rápida y eficaz actuación.

A la hora de desplegar un sistema SIEM, es necesario tener en cuenta múltiples factores, como puede ser el número de equipos a monitorizar, la granularidad que se pretende obtener, la heterogeneidad de la infraestructura, el tipo de alertas que se buscan o el tiempo de respuesta esperado. Sin embargo, es muy importante considerar el presupuesto para herramientas comerciales o el tiempo disponible para mejorar o adecuar una solución ya desarrollada, aprovechando así las múltiples herramientas *open source* existentes.

Este artículo propone el diseño de un sistema de correlación de bajo coste, basado en plataformas *open source*, para la detección de anomalías en el campo de la ciberseguridad. Para poder construir esta infraestructura es necesario tener en cuenta los principales sistemas que componen el entorno de desarrollo (ver Fig. 1), que se enmarca en el proyecto PLICA (Plataforma Integrada de Conciencia Cibersituacional). En primer lugar, un ingestor de eventos a través del cual los *logs* procedentes de múltiples sensores son parseados y enviados al siguiente bloque, un sistema de aprendizaje automático que, a partir del conjunto de datos brindado determina anomalías y las envía al sistema de almacenamiento, en este caso, ElasticSearch (ES). Los eventos se almacenan para su monitorización, alerta y notificación, gracias al *framework* ElastAlert, que realiza consultas periódicas a ES en base a reglas personalizadas que permiten correlar sucesos con múltiples orígenes y envía notificaciones cuando se dispara una alerta. Por último, el módulo de visualización mediante Kibana, permite analizar los eventos ocurridos de manera sencilla, y realizar así los análisis e informes pertinentes.

Por tanto, el sistema de correlación de eventos propuesto está basado en correlaciones temporales de anomalías heterogéneas a través de reglas de ElastAlert y alertas que se almacenan en ES.

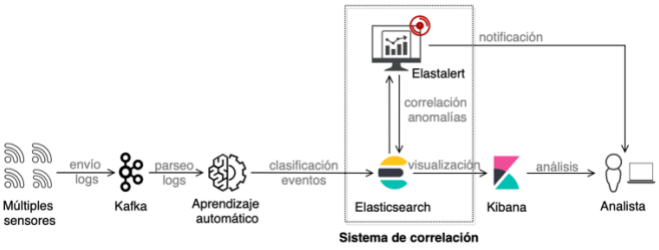


Figura 1. Arquitectura modular del entorno de desarrollo

II. Análisis de tecnologías

La correlación de eventos es una técnica esencial de monitorización que, según múltiples investigadores, puede plantearse a través de métodos basados en gráficos, en *codebooks*, en redes neuronales, en relaciones estadísticas y en reglas, siendo el último el más común [1]. Esta correlación forma parte de las prestaciones esperadas de los sistemas de gestión de redes y SIEM propietarios, sin embargo, en el ámbito del software libre, las soluciones disponibles para monitorización presentan múltiples inconvenientes a la hora de correlar eventos con orígenes heterogéneos. Esto se debe principalmente a que las herramientas existentes son soluciones robustas, con una arquitectura compleja que requieren mucho tiempo para su despliegue y mantenimiento, así como una amplía formación para los usuarios [2]. Actualmente, los escasos enfoques que abordan el diseño de un sistema de correlación de eventos gracias a herramientas open source se basan en la detección mediante reglas a través de SEC (*Simple Event Correlator)*, una plataforma ligera e independiente para la correlación en tiempo real. Tanto [1]

como [2] revisan esta herramienta a través de su descripción, diferentes casos de uso y recomendaciones, concluyendo que su rendimiento es bueno y puede soportar cientos de eventos por segundo, teniendo en cuenta que solo se exploran de manera práctica algunas de las prestaciones. Por otra parte, [3] se hace uso de SEC para correlar las alertas generadas por un IDS, con el fin de eliminar las alertas duplicadas. Sin embargo, ninguno de los planteamientos hace un análisis profundo de las capacidades de esta herramienta para correlar eventos con orígenes diversos, siendo la situación más habitual la correlación de datos homogéneos.

Tras la revisión realizada y con el fin de seleccionar la solución *open source* que más se ajuste a las necesidades del sistema propuesto, se ha llevado a cabo el siguiente análisis de tecnologías, por una parte de herramientas robustas y que ofrecen soluciones más completas, así como de aquellas más ligeras cuyo fin es la correlación y alerta de eventos, concluyendo que no existe ninguna plataforma que, de manera pública, proporcione todas las capacidades que un SIEM comercial ofrece, siendo común la combinación de diferentes soluciones con el fin de desarrollar una herramienta personalizada.

En primer lugar, la pila ELK [4] consiste en ES, Logstash, Kibana y Beats, productos *open source* a través de los cuales es posible recolectar y procesar datos, almacenarlos, filtrarlos y visualizarlos. Sin embargo, la correlación de eventos y alerta no están contemplados en la versión libre, siendo posible utilizar extensiones comerciales propias de ES como Elastic Security [5] o Watcher [6]. Elastic Security combina la detección de amenazas a través de un SIEM propio con prevenciones en *endpoints* y una alta capacidad de respuesta, múltiples herramientas para la identificación de ataques, además de entornos de gestión y visualización, así como la posibilidad de configurar modelos de *machine learning*.

Por otra parte, Watcher permite crear acciones basadas en determinadas condiciones, que son periódicamente evaluadas lanzando consultas a la base de datos, siendo muy útil a la hora de analizar flujos de datos en situaciones críticas. Esta solución cuenta con una API para crear, gestionar y testear *watches*, que describen una alerta y pueden contener múltiples acciones de notificación. Tanto Elastic Security como Watcher cuentan como casos de usos predefinidos, sin embargo, entre ellos no se contempla la correlación de eventos heterogéneos.

Security Onion [7] es una solución *open source* diseñada para la detección de amenazas, monitorización en empresas y gestión de *logs*, que incluye, además, múltiples herramientas de seguridad adicionales como, por ejemplo, ELK. Security Onion tiene tres funciones principales: captura de paquetes, sistemas de detección de intrusiones basados en red y en puntos finales (NIDS y HIDS por sus siglas en inglés) apoyándose en otras soluciones, tanto NIDS como HIDS, como, por ejemplo, Suricata o Wazuh respectivamente, y, por último, herramientas para el análisis, entre las que se incluye Kibana o Sguil. En lo referente a la correlación de eventos, Security Onion no cuenta con una funcionalidad específica para este fin puesto que proporciona una solución muy robusta orientada a cubrir desde la detección de eventos y su inyección hasta su análisis, llegando a una arquitectura final realmente personalizada debido a la integración de esta solución con un gran número de plataformas *open source*.

En cuanto a las plataformas ligeras, *Simple Event Correlator* (SEC) [8] se trata de una herramienta de correlación de eventos en tiempo real que puede ser aprovechada para la monitorización de *logs,* gestión de redes y cualquier otra tarea que implique correlar sucesos. A diferencia de las soluciones previamente analizadas, SEC es una solución ligera e independiente que se ejecuta en un solo proceso, pudiendo el usuario emplearla de manera iterativa en un terminal o ejecutar varios procesos SEC simultáneamente, por ejemplo. Si bien esta herramienta se basa en un conjunto de reglas mayoritariamente diseñadas para ser usadas con datos de entrada homogéneos, como por ejemplo, realizar acciones si se encuentran *x* coincidencias en un tiempo determinado, existe la opción de establecer correlaciones entre múltiples ficheros que contengan sucesos heterogéneos, como puede ser la generación de una alarma si durante 1 minuto ocurren diez eventos provenientes del *firewall* y cinco del IDS, todos con la misma dirección IP asociada [9].

Por último, ElastAlert [10] es un *framework* simple de Yelp, que, junto a ELK, está diseñado para alertar en tiempo real sobre anomalías u otros patrones de interés haciendo uso de los datos de ES. ElastAlert funciona combinando tres componentes: tipos de regla, alertas y mejoras. Se realizan consultas periódicas a ES y los datos se pasan a la regla, que determina si hay un *match* (un evento que coincide con las características definidas en la regla). Cuando esto ocurre, ElastAlert genera una o mas alertas, que ejecutan diferentes acciones definidas en la regla, almacenando periódicamente su estado en ES. A pesar de que en ElastAlert existe un conjunto de reglas por defecto, entre ellas no se encuentra la opción de correlación de eventos entre diferentes índices de ES, siendo posible configurar nuevas reglas o realizar adaptaciones o mejoras a las existentes. Adicionalmente, esta solución, cuenta con la posibilidad de añadir un *plugin* para Kibana con el fin de crear, editar y probar las reglas desde la consola de visualización, de manera más sencilla. Otra opción interesante es la integrabilidad de ElastAlert con las reglas SIGMA, un formato estándar de reglas para sistemas SIEM que permite definir consultas a la base de datos que pueden ser convertidas a múltiples formatos.

Tras el análisis realizado, se puede comprobar que tanto ELK como Security Onion están diseñadas para cubrir e implementar todas las etapas para la monitorización de una red, y aunque la correlación de eventos forma parte de este proceso, las opciones que ofrece ELK no están disponibles de manera abierta y Security Onion no cuenta con una funcionalidad especifica para este fin, siendo ambas son soluciones muy robustas, quedándose fuera del alcance de la investigación. Por otra parte, SEC y ElastAlert proporcionan independencia y correlación de eventos basada en reglas. Sin embargo, SEC cuenta con una regla específica para la correlación entre datos heterogéneos, al contrario que ElastAlert, cuyas reglas están enfocadas a datos de entrada homogéneos, siendo posible generar nuevas reglas que permitan el uso de eventos con origen y características diversas, teniendo también presente que está diseñado para funcionar sobre ES, al contrario que la primera herramienta ligera mencionada. Otros factores diferenciadores son, por una parte, la sintaxis de las reglas, siendo la de SEC más compleja, y por otra la amplia comunidad con la que ElastAlert cuenta.

Por tanto, la tecnología que más se adapta al sistema propuesto es ElastAlert ya que a través de sus reglas personalizadas, en primer lugar, permite plantear una solución ligera con una funcionalidad específica para correlar eventos heterogéneos que pueda ser implementada en cualquier sistema, además de ofrecer capacidades de detección en otros

escenarios gracias al amplio portfolio de reglas para seguridad disponibles, como establecer umbrales, generar métricas, *blacklist*, etc, y la posibilidad de filtrar por un valor de parámetro específico en un evento como puede ser el *payload* o el número de bytes transmitidos. Por otra parte, ElastAlert está diseñado para alertar directamente sobre ES, siendo esta una de las plataformas de almacenamiento más comunes y la empleada en el entorno de desarrollo de esta investigación, permitiendo así correlar alertas generadas por cualquier tipo de sensores de ciberseguridad como: IDS, SIEM, Firwall, etc. Como consecuencia permite, únicamente con el requisito de ingestar las alertas en ES, realizar la correlación a través de los eventos indexados en diversos índices propios de ES.

III. Sistema propuesto

El sistema propuesto, mostrado en la Fig. 3, tiene como objetivo identificar la correlación entre anomalías a partir de un conjunto de anomalías previamente clasificadas con origen en múltiples sensores. El sistema está compuesto por un subsistema de correlación individual y un subsistema de correlación conjunta, sirviendo la salida del primero como entrada del segundo.

A. Subsistema de correlación individual

El subsistema de correlación individual es el encargado de identificar las anomalías procedentes de los múltiples orígenes, generar alertas de correlación individual y almacenarlas de forma conjunta en un nuevo índice de ES, que servirá como entrada para el siguiente subsistema.

Los datos de entrada de este subsistema son los índices individuales de la base de datos, para cada tipo de sensor, en los que se ha ingestado los eventos (anómalos y no anómalos) procesados por el sistema de aprendizaje automático, como muestra la Fig. 1. Un evento, que puede haber sido generado por cualquier tipo de sensor, está compuesto por un conjunto de características en un tiempo determinado, como se muestra en la Fig. 2, etiquetado por un valor de "*False*" o "*True*" en caso de ser clasificado como anomalía. Por otra parte, los datos de salida son similares a los de entrada, siendo las diferencias principales, por una parte, el filtrado de anomalías, y por otra, el almacenamiento conjunto de todas ellas gracias a las correlaciones individuales, además de añadir algunos metacampos propios de ElastAlert a las características de un evento.

```
"versión": "1.0",      // Versión del sensor
"time": "<EPOCH>",     // Timestamp en formato EPOCH
"id": "<ID>",          // Identificador único del sensor (UUID/GUID)
"type": "<TIPO>",      // Tipo de sensor: IDS, RF, RM, WF, BT, etc.
"event": "<EVENT>",    // Tipo de mensaje: DATA o ALARM
"data": [
{
    "dato1": xxx,
    "dato2": xxx,
    "dato3": xxx,
    …
    "datoM": xxx,
}]
```

Figura 2. Formato de los eventos generados por los diversos tipos de sensores, en el que el campo *data* incluye los parámetros relevantes captados por el mismo, así como el campo *anomalía*

B. Subsistema de correlación conjunta

El subsistema de correlación conjunta es el encargado de recibir un conjunto de anomalías o correlaciones individuales e identificar una correlación entre ellas, generando una anomalía correlada conjunta en base a la procedencia de las anomalías que se quieren correlar y el intervalo de tiempo en el que deben ocurrir.

Los datos de entrada del segundo subsistema coinciden con los de salida del primero, es decir, son las correlaciones individuales almacenadas en el nuevo índice generado por el subsistema de correlación individual. Estos datos están compuestos por el conjunto de características previamente mencionadas en la descripción de los datos de salida del subsistema anterior.

Como salida, el subsistema genera las anomalías correladas conjuntas y las almacena en un nuevo índice final de la base de datos. Estas nuevas anomalías están compuestas por el conjunto de características respectivas a un evento en un tiempo determinado, en concreto, del último evento detectado dentro del rango temporal especificado. Además, se añade el campo *related_events* en el que se incluyen las características de las diferentes anomalías involucradas en correlación conjunta final.

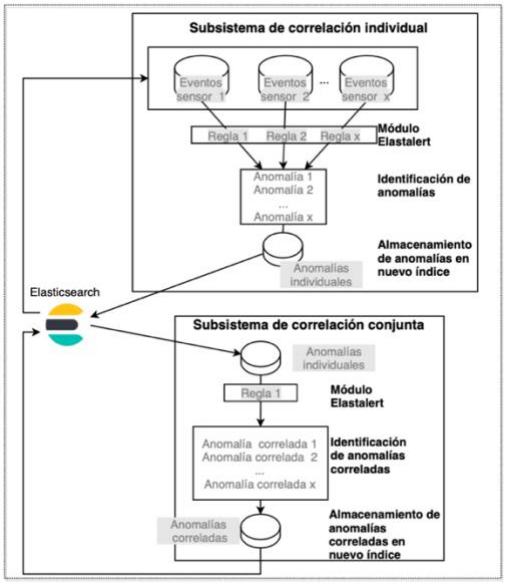


Figura 3. Arquitectura modular del sistema de correlación

IV. Implementación

Como se ha descrito anteriormente, el sistema de correlación propuesto está compuesto por dos subsistemas que hacen uso de ES, tanto para realizar consultas como para almacenar su estado. Es necesario por lo tanto definir una configuración de ElastAlert específica para cada subsistema, en la cual se definen desde los parámetros respectivos a ES, como la dirección IP o el índice de almacenamiento, además del directorio en el que se encuentran las reglas que deben seguir la configuración. Por otra parte, se generan las reglas de correlación individual, una por tipo de sensor, y las reglas de correlación conjunta, una por conjunto de anomalías que se quiera correlar.

Para poner en marcha el sistema de correlación, en primer lugar, se comprueba el correcto estado y funcionamiento del entorno de desarrollo: sensores desplegados, instalados y enviando datos de actividad; sistema de aprendizaje automático procesando eventos y enviándolos a ES; y base de datos levantada y recibiendo eventos. Como se muestra en el Algoritmo I, se ejecuta el subsistema de correlación individual, el cual ejecuta *n* reglas para *n* tipo de sensores diferentes, las cuales consultan los índices correspondientes a cada uno, filtrando las anomalías, generando una alerta y almacenándola de nuevo en la base de datos si estas ocurren en el rango

temporal determinado. Al mismo tiempo, se levanta el subsistema de correlación conjunta, que, con consultas al nuevo índice creado, genera una alerta de correlación conjunta, y su seguido almacenamiento, si en el rango temporal determinado ocurren *n* correlaciones individuales.

Algoritmo I
ALGORITMO DE LANZAMIENTO DEL SISTEMA DE CORRELACIÓN

Lanzamiento del subsistema de correlación individual
correlacion_individual_sensor_1.yaml
Consulta *index_sensor_1*
    Si timeframe: *x* AND anomalia: true
    Alerta de correlación individual sensor_1
    Almacenamiento de alerta 1 en *index_conjunto*
Ejecución de *n* reglas de correlación individual para *n* sensores
Lanzamiento del subsistema de correlación conjunta
correlacion_conjunta.yaml
Ejecución en paralelo del subsistema
Consulta *index_conjunto*
    Si timeframe: *x* AND num_correlaciones_individuales: *n*
    Alerta de correlación conjunta
    Almacenamiento de alerta de correlación conjunta en *index_final*

V. PRUEBAS INICIALES Y RESULTADOS

El entorno de validación y evaluación inicial del sistema de correlación está basado en una primera versión en la que la inyección de datos al sistema es simulada, teniendo en cuenta que estos eventos han sido recopilados previamente por diversos sensores en un determinado entorno real.

Para esta primera versión se propone el siguiente caso de uso: identificación de una anomalía correlada conjunta con origen en sensores de actividad en señales Wifi, Bluetooth, telefonía móvil y radio frecuencia, dividiendo el proceso en dos fases.

La primera fase implica la verificación del funcionamiento del subsistema de correlación individual, por una parte, con datos posiblemente anómalos, y por otra, con datos no anómalos. Los resultados obtenidos son los esperados, es decir, la generación de alertas y su correspondiente almacenamiento cuando se reciben anomalías y el caso contrario cuando los eventos son normales, es decir, no anómalos.

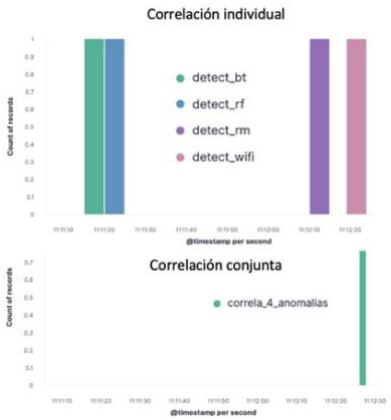


Figura 4. Panel de visualización del sistema de correlación para el caso de uso propuesto. Correlación individual por cada tipo de sensor (arriba), y correlación conjunta (abajo)

La segunda fase implica la verificación del funcionamiento del subsistema de correlación conjunta, por una parte, en un entorno en el que existen suficientes correlaciones individuales, generadas por el subsistema de correlación individual para cada uno de los sensores heterogéneos, para poder identificar una correlación conjunta como se observa en la Fig. 4, como la validación en un entorno en el que las correlaciones individuales no son suficientes para obtener una correlación conjunta final. Adicionalmente se verifica la funcionalidad de adjuntar en una misma alerta los eventos implicados en la generación de esta correlación heterogénea.

Los resultados obtenidos en la validación del segundo subsistema son los esperados: la obtención de una correlación conjunta, en la que se incluyen todos los sucesos involucrados, y su almacenamiento en la base de datos cuando el número de correlaciones individuales es el indicado. Por el contrario, si estas correlaciones previas son insuficientes, no se produce la correlación final.

VI. CONCLUSIONES Y LÍNEAS FUTURAS

El presente trabajo demuestra la posibilidad de desplegar un sistema de correlación de eventos en tiempo real de bajo coste en un entorno de ciberseguridad, basado en plataformas *open source* como ES y ElastAlert, a través de múltiples reglas y alertas. Las pruebas desarrolladas para los subsistemas que componen el sistema propuesto verifican su correcto funcionamiento con actividad recopilada por múltiples sensores en un entorno real, siendo posible identificar anomalías correladas conjuntas con origen en múltiples sensores o sistemas.

Como líneas futuras se propone en primer lugar, la mejora de la precisión temporal, así como la incorporación de nuevas funcionalidades de *threat intelligence* mediante la correlación a través de *blacklist* y/o *whitelist*, o la generación de un subsistema para mantenimiento. Por otra parte, se plantea la integración en el sistema de otras plataformas *open source*, con el fin de aumentar las capacidades de detección, como puede ser SIGMA Rules.

AGRADECIMIENTOS

La investigación presentada en este artículo ha sido parcialmente financiada por el proyecto PLICA dentro del programa COINCIDENTE del Ministerio de Defensa del Gobierno de España.

REFERENCIAS

[1] R. Vaarandi, B. Blumbergs y E. Çalışkan, «Simple Event Correlator - Best Practices for Creating Scalable Configurations,» *IEEE CogSIMA Conference,* 2015.

[2] R. Vaarandi, «Simple Event Correlator for real-time security log monitoring,» *Hakin9 Magazine,* 2006.

[3] N. Dwivedi y A. Tripathi, «Event Correlation for Intrusion Detection Systems,» *IEEE International Conference on Computational Intelligence & Communication Technology,* 2015.

[4] «ELK: Elasticsearch, Kibana, Beats y Logstash,» Elasticsearch, [En línea]. Available: https://www.elastic.co/es/elastic-stack.

[5] «Elastic SIEM,» Elasticsearch , [En línea]. Available: https://www.elastic.co/es/siem.

[6] «How Watcher works,» Elasticsearch, [En línea]. Available: https://www.elastic.co/guide/en/elasticsearch/reference/current/how-watcher-works.html.

[7] «Security Onion Documentation,» Security Onion , 2021. [En línea]. Available: https://docs.securityonion.net/en/2.3/.

[8] «Simple Event Correlator v2.9.apha2,» [En línea]. Available: https://simple-evcorr.github.io.

[9] D. Lang, «Using SEC,» *USENIX ;login: Magazine,* vol. 38, nº 6, 2013.

[10] «Elastalert, Easy & Flexible Alerting With Elasticsearch,» Yelp, [En línea]. Available: https://elastalert.readthedocs.io/en/latest/.

http://doi.org/10.18239/jornadas_2021.34.14

Un resumen de: *Present and future of network security monitoring*

Marta Fuentes-García
Fundación I+D del Software Libre (Fidesol)
mfuentes@fidesol.org

https://orcid.org/0000-0002-7428-1277

José Camacho (1), Gabriel Maciá-Fernández (2)
Universidad de Granada - *NESG* - CITIC
{josecamacho, gmacia}@ugr.es

(1) https://orcid.org/0000-0001-9804-8122
(2) https://orcid.org/0000-0001-9256-453X

***Resumen*—NSM (*Network Security Monitoring*) es un término utilizado para referirse a la detección de incidentes de seguridad mediante la monitorización de eventos de red. Los sistemas NSM son esenciales para las seguridad en las redes actuales, dada la escalada en la sofisticación del cibercrimen. En este artículo se revisa el estado del arte de NSM y se deriva una nueva taxonomía de las funcionalidades y módulos de un sistema NSM. Esta taxonomía es útil para evaluar los desarrollos y herramientas actuales, tanto para investigadores como profesionales en ciberseguridad. Además, identificamos los retos de la aplicación de NSM en redes modernas, como son SDN (*Software Defined Networks*) e IoT (*Internet of Things*).**

***Index Terms*—seguridad en redes, NSM, monitorización de seguridad, detección y respuesta a incidentes, SDN, IoT**

Tipo de contribución: *Investigación ya publicada: "Present and Future of Network Security Monitoring," in IEEE Access, doi: 10.1109/ACCESS.2021.3067106.*

I. Introducción

Aunque la mayor parte de los esfuerzos en la seguridad en redes aún se basan en la prevención de ataques, las técnicas y soluciones basadas en la detección y respuesta están ganando cada vez más relevancia [1]. NSM (*Network Security Monitoring*) es uno de los enfoques más relevantes para la seguridad en redes [2]. El ciclo NSM tiene cuatro fases [2, 3]: *1) Monitorización*, *2) Detección*, *3) Análisis forense*, y *4) Respuesta*. El objetivo es monitorizar el estado de una red para detectar eventos anormales y, una vez detectados, gestionarlos a tiempo y de manera adecuada.

En este trabajo presentamos un resumen del artículo [4], donde revisamos el estado del arte de NSM, con el objetivo de proporcionar una taxonomía y una descripción unificada de sus componentes. Evaluamos y clasificamos algunas de las soluciones existentes siguiendo la taxonomía propuesta. Las soluciones revisadas incluyen: IDS/IPS (*Intrusion Detection System* / *Intrusion Prevention System*), SEM/SIEM (*Security Event Management* / *Security Information and Event Management*), y UTM (Universal Threat Management). Finalmente, analizamos las tendencias más relevantes en redes modernas, los (nuevos) retos que plantean, y cómo se afrontan desde la perspectiva NSM[1].

II. Taxonomía modular de un sistema NSM

Un sistema NSM debería ser capaz de proporcionar trazabilidad de las actividades y procesos que tienen lugar en la red y los subsistemas bajo monitorización. Para lograr este objetivo, una arquitectura NSM típica se compone de distintos elementos (hardware y software) que se distribuyen a lo largo de la red. Estos elementos envían información sobre los eventos de red a un punto centralizado, donde son almacenados y analizados.

Una revisión exhaustiva sobre los sistemas NSM más utilizados nos ha llevado a proponer una taxonomía de las funcionalidades NSM. La mayoría de estas soluciones implementan al menos una de las siguientes funcionalidades: *sensor*, *parser*, *integrador*, *detector*, *inspector*, y *actuador*.

- El **sensor** captura datos de un subsistema/red. Los datos se pueden obtenen en forma de registros o *logs*, entre otros.
- El **parser** transforma el formato de los datos.
- El **integrador** combina múltiples fuentes de datos en un único flujo de datos.
- El **detector** identifica eventos/registros anómalos en el flujo de datos.
- El **inspector** permite la exploración de los datos.
- El **actuador** lleva a cabo acciones automáticas en la configuración del subsistema/red.

Se trata de sistemas que son inherentemente modulares, lo que facilita la escalabilidad para construir otros sistemas más complejos combinando las salidas de distintos módulos. El sensor, el parser y el integrador se suelen incluir dentro de la fase de monitorización, mientras que el resto de los módulos tienen una relación uno a uno con los demás pasos del ciclo NSM (detección, análisis forense y respuesta).

III. Retos actuales de investigación para NSM

Creemos que NSM puede beneficiar a los nuevos paradigmas de comunicación: SDN (*Software Defined Networks*), IoT (*Internet of Things*), e IIoT (*Industrial IoT*). En el estudio realizado en [4], identificamos los retos de investigación para cada uno de los módulos propuestos en la taxonomía de la Sección II. Dichos retos se resumen a continuación:

- **Sensor.** Este es un componente bien establecido, implementado por la mayoría de soluciones investigadas. La mayoría de los sensores individuales estudiados se pueden utilizar en redes modernas, junto con los sensores que son intrínsecos a la naturaleza de IoT/IIoT.
- **Parser.** Este componente se considera en algunos de los trabajos estudiados desde el punto de vista de la extracción automática de características. En algunos casos, también se lleva a cabo una unificación y reducción

[1]La clasificación detallada y las tablas que resumen las soluciones y trabajos evaluados (tanto para redes tradicionales como redes modernas) se pueden encontrar en el trabajo original [4].

de redundancia. Sin embargo, aún existen posibilidades de investigación, que se deberían enfocar hacia el establecimiento de un formato unificado para el registro de eventos y la extracción de características en datos masivos.

- **Integrador.** Este es uno de los módulos más importantes y útiles de NSM, ya que permite agregar datos de distintas fuentes. La integración es aún más relevante cuando se trata de paradigmas de comunicación modernos, donde varios dispositivos heterogéneos envían y reciben información que debe ser unificada para su monitorización y detección de incidentes. Sin embargo, sólo cuatro de los trabajos estudiados consideran este módulo de forma explícita. Por esta razón, encontrar una estrategia para integrar, agregar y correlar distintas fuentes de datos es todavía uno de los principales retos para los investigadores.
- **Detector.** La mayoría de los trabajos de investigación actuales se centran en este módulo, con el objetivo de encontrar nuevas formas de aplicar algoritmos de aprendizaje automático y mejorar la capacidad de detección de ataques y/o anomalías. Uno de los principales retos de este componente es el procesamiento de cantidades masivas de datos para crear y aplicar los modelos de detección. Además, la priorización de alarmas y la reducción del número de falsos positivos son aún problemas abiertos.
- **Inspector.** Este componente pretende localizar un incidente, una vez se ha detectado por el módulo detector, tanto en el espacio como en el tiempo. Esto es incluso más importante cuando hablamos de redes descentralizadas (p.e. IoT), debido a los siguientes factores: *i)* la fuente del evento podría no estar ubicada en la misma localización que el evento, y *ii)* varios dispositivos distintos probablemente estén intercambiando información a través de la red. Sin embargo, sólo dos de los trabajos revisados incluyen un módulo de inspección. Este componente necesita investigación adicional, no sólo para proporcionar registros y almacenar información del estado cuando tiene lugar un incidente, sino para hacerlos interpretables. Esto ayudará a los operadores de seguridad a comprender los hechos y a hacer más eficientes las tareas forenses.
- **Actuador.** Este módulo se considera en varios de los trabajos estudiados, que definen algunas acciones, como el bloqueo de acciones maliciosas o de los equipos afectados. Los principales retos de este componente son la definición e implementación de mecanismos de autorecuperación que hagan las redes de comunicaciones resilientes. Esto es especialmente importante en sistemas críticos, típicamente relacionados con entornos IIoT.

Finalmente, la escalabilidad y compatibilidad con los recursos restringidos (estos últimos especialmente para IoT/IIoT) son problemas comunes que aún permanecen abiertos para estos módulos. Además, las soluciones disponibles en el mercado necesitan ser actualizadas para poder superar estos retos y proporcionar soluciones modernas que cubran la mayoría de los módulos NSM. Si estas soluciones se diseñan siguiendo la filosofía propuesta, serán más escalables y será más fácil completarlas y mejorarlas.

IV. Conclusiones

En este artículo revisamos el estado del arte de NSM, proporcionando una visión global y una clasificación unificada de sus componentes. Nuestra taxonomía clasifica dichos componentes como sensores, parseadores, integradores, detectores, inspectores y actuadores. Estos módulos pueden combinarse de distintas formas, proporcionando una arquitectura escalable y de alto potencial para la detección de intrusiones. Este trabajo resalta los puntos fuertes y débiles de los módulos identificados en las herramientas NSM disponibles, así como en las propuestas de la literatura.

Revisamos las soluciones existentes para soluciones multimódulo que siguen la filosofía NSM. Los mejores ejemplos de estas combinaciones son los IDS/IPS, SEM/SIEM y UTM.

Finalmente, evaluamos la aplicabilidad del enfoque NSM en redes de comunicaciones modernas. Centramos esta evaluación en SDN y redes IoT/IIoT. Además, se resumen los problemas abiertos y futuros intereses de investigación para cada uno de los módulos NSM en relación con los nuevos paradigmas de comunicación.

Creemos que este artículo es de interés tanto para la comunidad investigadora como para los profesionales de la ciberseguridad, ya que ayuda a centrar el esfuerzo de investigación y las soluciones de mercado de forma efectiva. Además, permite la identificación de herramientas y métodos que están disponibles para recopilar y procesar datos de seguridad en redes para detección de incidentes.

Como conclusión final, creemos que el panorama de seguridad en redes (tanto modernas como tradicionales) se beneficiaría de *i)* la investigación y desarrollo de módulos inspectores y actuadores, que son las soluciones menos desarrolladas hasta la fecha; y *ii)* el diseño de sistemas que incluyan todos los componentes identificados. Además, aún es necesario proporcionar soluciones eficientes que consideren la restricción de recursos existente en IoT/IIoT, así como mejorar la resiliencia en infraestructuras críticas.

Agradecimientos

Este trabajo está financiado en parte por las Ayudas Cervera para Centros Tecnológicos del Centro Español para el Desarrollo de Tecnología Industrial (CDTI) en el marco del proyecto EGIDA (CER-20191012) y por el Ministerio de Economía y Competitividad de España y fondos FEDER (Fondo Europeo de Desarrollo Regional) a través del proyecto TIN2017-83494-R.

Referencias

[1] R. Samson, "Prevention vs Detection-Based Security Approach," Clearnetwork, https://www.clearnetwork.com/prevention-vs-detection-cybersecurity-approach/, Tech. Rep., 2020, [Online, accessed on 16/07/2020].

[2] R. Bejtlich, *The TAO of the Network Security Monitoring. Beyond Intrusion Detection*. Addison-Wesley, 2005.

[3] R. G. Bace, *Intrusion Detection*. Macmillan Technical Publishing (Technology Series), 2000.

[4] M. Fuentes-García, J. Camacho, and G. Maciá-Fernández, "Present and Future of Network Security Monitoring," *IEEE Access*, pp. 1–1, 2021.

http://doi.org/10.18239/jornadas_2021.34.15

Secure Crowdsensing Platforms Through Device Behavior Fingerprinting

Pedro Miguel Sánchez Sánchez
University of Murcia
pedromiguel.sanchez@um.es

Alberto Huertas Celdrán
University of Zurich
huertas@ifi.uzh.ch

Gérôme Bovet
armasuisse Science & Technology
gerome.bovet@armasuisse.ch

Gregorio Martínez Pérez
University of Murcia
gregorio@um.es

Burkhard Stiller
University of Zurich
stiller@ifi.uzh.ch

***Abstract*—Crowdsensing platforms allow sharing data, collected by devices of individuals, to achieve common objectives based on the analysis of shared information. Despite their benefits, these platforms also bring security threats that must be addressed to provide secure data and reliable services. In this context, device behavior fingerprinting becomes a key technique to detect and mitigate possible cyberattacks affecting resource-constrained devices. This work presents the most relevant research questions in the field of behavior fingerprinting to identify devices and detect anomalies produced by cyberattacks. In addition, it also introduces the main goals and current status of two research projects dealing with such research questions.**

***Index Terms*—Crowdsensing, Device Behavior, Cybersecurity, Identification, Attack Detection**

Tipo de contribución: *Investigación en desarrollo*

I. Introduction

The evolution of wireless communications, technologies, and devices is bringing novel platforms that integrate IoT (Internet-of-Things) sensors and actuators into heterogeneous environments such as industry, agriculture, health, or smart homes to provide innovative and real-time functionalities. One of the most promising and recent types of Web platforms is based on crowdsensing. Crowdsensing platforms allow users to collaborate, altruistically or not, to achieve a particular common goal by sharing data collected in local environments and analyze them, from a global perspective, to draw common conclusions. As an example, there exist successful platforms, such as Flightradar24 and OpenSky24 in the field of air traffic or ElectroSense [1] focused on analyzing the electromagnetic spectrum usage.

Crowdsensing-based platforms show benefits, such as flexibility, low deployment, and maintenance costs, or wide data variety. However, they also open the door to critical cybersecurity concerns to be addressed in order to provide secure and reliable platforms, services, and data. In this regard, one of the most critical cybersecurity concerns is on exploiting well-known vulnerabilities of resource-constrained sensors and actuators used in crowdsensing platforms. It is complemented by a second one focused on the replacement and/or modification of devices with malicious functionality. Finally, it is also important to note that sensors and actuators are permanently connected to the Internet and can be reached by cybercriminals. It is, therefore, necessary to devise solutions capable of detecting and mitigating cyberattacks before a crowdsensing platform is compromised.

One of the most promising solutions to improve such concerns is the creation of fingerprints that model the internal behavior of resource-constrained devices acting as sensors or actuators. A multitude of behavioral data coming from a wide variety of heterogeneous sources, such as hardware events, system logs, or clock skew, can be leveraged in order to model the "normal" behavior of devices and detect deviations produced by cyberattacks. In this context, the literature proposes the usage of device behavior fingerprinting in two main cybersecurity scenarios [2]. The first identifies devices with different granularity levels, such as type, model, or individual device. Depending on the granularity desired, the resources used will be different. Network, resource consumption, or performance are related to the type and model identification, while resources, such as clock skew and oscillator variations, are used to distinguish identical devices. The second scenario focuses on detecting misbehavior caused by cyberattacks or device malfunctioning. In this sense, resources, such as network communications, resource usage, system calls, and logs, are monitored to deploy a Host-based Intrusion Detection System (HIDS).

Despite the progress of related work, more efforts are required to improve device identification and detection of cyberattacks in crowdsensing platforms. Among them, the necessity for solutions considering unique characteristics of novel devices and modern crowdsensing platforms is highlighted. In this context, besides hardware and software limitations, resources of dedicated sensors and actuators are critical when developing new security solutions. Moreover, existing solutions identifying devices in an individual way raise scalability concerns that need to be analyzed in detail and solved, if possible. Finally, privacy issues related to monitored data and how it is handled outside the device are appearing.

With the goal of improving these challenges, this paper introduces two research lines aligned with device behavior fingerprinting. The first one focuses on the identification of

identical devices to detect spoofing cyberattacks, while the second deals with detecting behavior anomalies caused by cyberattacks affecting devices of crowdsensing platforms. In this sense, the main contributions of this paper include:

- A common threat model of crowdsensing platforms, identifying the most common cyberattacks affecting device identification and behavior anomalies.
- A set of open research questions on individual device identification as well as attack and misbehavior detection.
- The objectives and current status of two research projects, TREASURE and CyberSpec, dealing with identical device identification and detection of behavior anomalies produced by cyberattacks, respectively.
- An Artificial Intelligence (AI)-based security framework, common for TREASURE and CyberSpec, that uses device behavior fingerprinting to detect security threats present in devices belonging to a crowdsensing platform focused on radio frequency monitoring.

The remainder of this paper is structured as follows. While Section II defines the threat model identified in crowdsensing environments, Section III sketches the design and objectives of this research as well as its current status. Finally, Section IV provides insights gained from the development so far and outlines future steps.

II. Threat Model

The literature has already identified cybersecurity threats that must be considered when developing and using crowdsensing platforms [3], [4]. This section summarizes the most critical and well-known threats affecting these platforms. Regarding privacy, the main threat found is the next one:

Data disclosure. This threat refers to sensitive information that is erroneously published or accessed by an attacker.

Focusing on security aspects, the threats detected are more diverse and can affect the platform in several ways:

Spoofing. This threat occurs when an attacker replaces a legitimate sensor or actuator with a malicious device using the same identity. Usually, it is the starting point to further attacks like data injection.

Sybil. This threat happens when an attacker sends a lot of fake data with many different user's identities to alter the decisions generated by the platform that processes the data.

Malware. This threat represents the infection with malware of a vulnerable device. There are diverse malware types based on the activities and propagation methods such as Viruses, Worms, Trojans, or Spy tools.

Jamming. This threat occurs when a jammer inserts fake or repeated signals with the objective of interrupting ongoing communications between legitimate sensors and the server.

Denial of Service (DoS). This threat is related to the platform or device services degradation by exhausting the available resources. It can appear at network level or directly at application level.

Advanced Persistent Threat (APT). In this threat, attackers launch sophisticated, continuous, and targeted attacks over the platform or its devices for a large time period.

Smart attacks. This threat involves the usage of machine learning techniques and smart devices to gather information about the platform defense countermeasures and attack it.

Data poisoning. This threat interferes with the consumer decision-making process modifying data generated by crowdsensing platforms. It leads to wrong decisions or not even drawing any information from these data. Two variants are differentiated.

- *Availability Attack:* This threat maximizes the harm on the decision-making processes of the crowdsensing platform, while maximizing the error of data available on the platform.
- *Targeted Attack:* This threat interferes with the decision process of the platform, attacking sensed objects or resources, like selected frequencies.

Table I compares the main threats identified in literature, detailing affected assets, the property damaged, and existing countermeasures for these threats. Threats are focused on crowdsensing environments, such as data poisoning or jamming. However, additional threat categories are also present in other types of Internet-of-Things (IoT) environments. In this sense, crowdsensing-based solutions can also be applied to different environments with common characteristics, such as Industrial IoT, Smart Cities/Buildings, or Healthcare IoT.

TABLE I
Threat Model Summary

Threat	Affected Asset	Damage	Countermeasure
Data disclosure	Sensors	Privacy	Attack / Behavior Anomaly Detection
Spoofing	Platform and Sensors	Integrity	Individual Device Identification
Sybil	Platform	Integrity	Individual Device Identification
Malware	Platform and Sensors	Availability	Attack / Behavior Anomaly Detection
Jamming	Platform and Sensors	Availability	Attack / Behavior Anomaly Detection
DoS	Platform	Availability	Attack / Behavior Anomaly Detection
APT	Platform and Sensors	Integrity	Attack / Behavior Anomaly Detection
Smart attacks	Platform and Sensors	Integrity	Attack / Behavior Anomaly Detection
Data poisoning	Platform	Integrity	Attack / Behavior Anomaly Detection

III. Device Behavior Fingerprinting

This section introduces the main research question on the usage of device behavior fingerprinting to *(a)* identify identical devices belonging to crowdsensing platforms and *(b)* detect cyberattacks affecting those devices. In addition, it explains the goals and current status of two on-going research projects, entitled TREASURE and CyberSpec, which are oriented to address (or at least improve) the previous two challenges.

As Fig. 1 depicts, five main modules have been designed to identify identical devices and detect cyberattacks affecting devices of crowdsensing platforms. In detail, the *Monitoring*

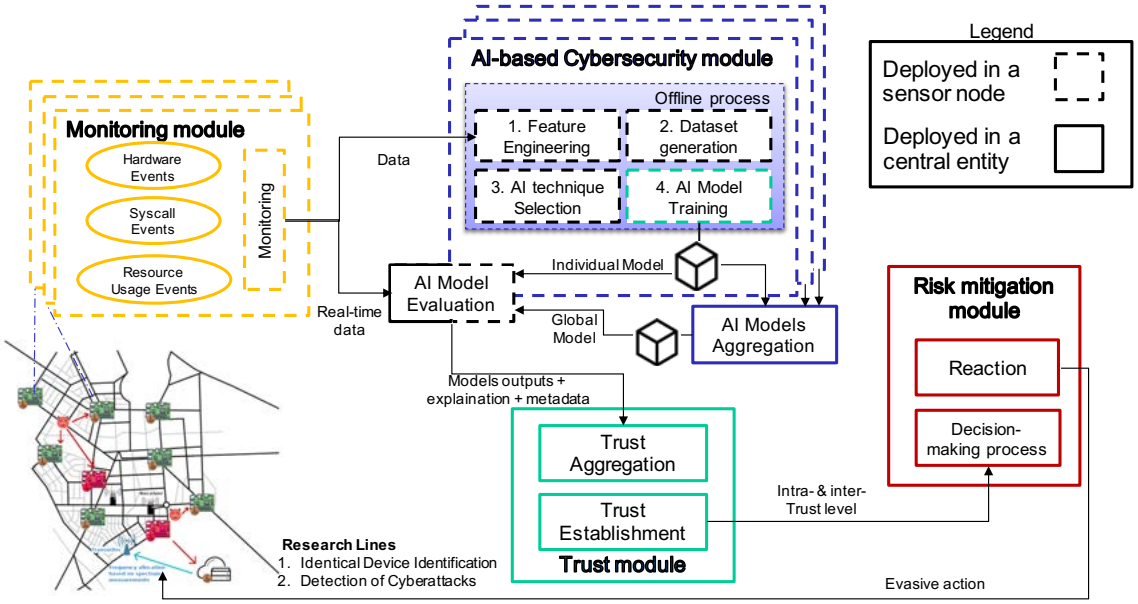


Fig. 1. Device Behavior Fingerprinting Architecture

module periodically acquires the internal behavior of each device, taking into account three resources: hardware events (such as CPU, GPU, and Hardware Performance Counters), systems calls, and the usage of resources (Memory, CPU, tasks, or network). These characteristics do not compromise device data privacy and they are shared and stored using secure encryption algorithms. After that, the data acquired is sent to the *AI-based Cybersecurity module*, which trains and evaluates individual and global AI-based models identifying devices and detecting cyberattacks. Later the *Trust module* evaluates models and their outputs to calculate a confidence score per prediction. Finally, the *Risk mitigation module* decides and enforces countermeasures according to the Trust and AI-based Cybersecurity outputs. It is important to note that this architecture is generic to be deployed in other IoT scenarios.

A. Device Identification

In the field of device identification, the following research questions are open and need major attention:

- T.RQ1: Is it feasible to build a solution to uniquely identify the sensors belonging to a crowdsensing platform in a reliable manner?
- T.RQ2: How does the solution scale when the number of devices deployed increases?
- T.RQ3: Are identification solutions resilient against possible adversarial attacks or situations affecting the identification process robustness?
- T.RQ4: Are the generated Machine and Deep Learning (ML/DL) models secure against adversarial attacks?
- T.RQ5: Can Federated Learning (FL) techniques improve existing data privacy problems of traditional ML/DL approaches identifying devices?

The main goal of TREASURE is to answer the previous questions designing and deploying an ML/DL-based and robust framework able to identify identical sensors of a crowdsensing platform, solving security threats based on sensor impersonation or malicious sensor deployment.

The proposed framework considers device behavior fingerprinting and the generation of ML and DL models to identify the sensors as well as possible malicious elements affecting the identification process robustness.

Currently, the work is focused on T.RQ1 and T.RQ2, dealing with sources of data capable of uniquely identifying a device and the most suitable ML/DL and algorithms for device identification and authentication. In terms of experimentation and solution validation, Raspberry Pis deployed in crowdsensing environments such as ElectroSense are used. The leveraged sources are the CPU and GPU performance, replicating environmental conditions, such as frequency, kernel interruptions, or temperature, in all test devices. Then, variations in performance and the correlation between components are analyzed using ML/DL techniques, mainly anomaly detection ones, to define a robust fingerprinting solution.

Fig. 2 shows results upon comparing five identical Raspberry Pi 4 devices. The evaluation is done by generating several CPU-GPU fingerprints for each device and verifying that the fingerprints generated by the same device (minimum similarity percentage in red) are recognized against the rest from other devices (maximum similarity percentage in blue) using Local Outlier Factor (LOF) algorithm. Further testing of the scalability and stability of the fingerprints is still required.

B. Cyberattacks and Behavior Anomaly Detection

In the context of detecting anomalies produced by cyberattacks affecting resource-constrained devices of crowdsensing platforms, the following research questions (C.RQ) are key:

- C.RQ1: What are the most suitable data sources of resource-constraint devices and ML/DL algorithms to

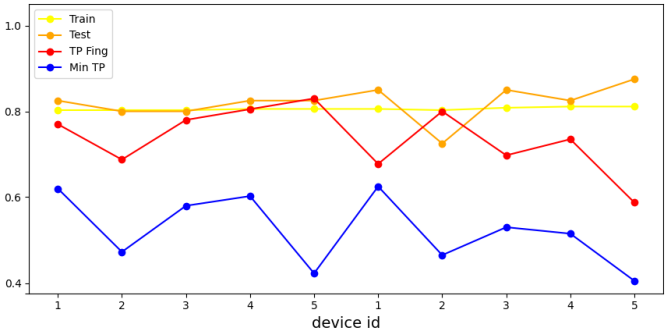


Fig. 2. Device Identification Results

create behavior fingerprints detecting anomalies produced by threats as of Table I?

- C.RQ2: Is it possible to build a common ML/DL-based system that uses device behavior fingerprinting to detect anomalies produced by heterogeneous cyberattacks affecting different resource-constrained devices of crowdsensing platforms?
- C.RQ3: Can Federated Learning techniques improve the performance, privacy, and robustness issues of traditional ML/DL techniques detecting cyberattacks affecting devices of crowdsensing platforms?
- C.RQ4: What are the key pillars and dimensions to build a trust algorithm able to calculate the trustworthiness level of AI-based predictions?
- C.RQ5: Is it possible to detect and mitigate in real-time those cyberattacks as of Table I in heterogeneous resource-constrained devices of crowdsensing platforms?

To answer these research questions, the main goal of the CyberSpec project is to research, design, and implement an intelligent and automatic framework providing secure and trusted resource-constrained devices, such as Raspberries Pi, used by the ElectroSense platform.

As a starting point, a systematic literature review has been performed to study and analyze internal dimensions and events available in Raspberry Pis (related to C.RQ1). As a consequence, a monitoring module has been designed and implemented to periodically acquire around 100 events belonging to the usage of resources, hardware, and software events produced in Raspberries Pi. To evaluate the suitability of these dimensions and events selected by the monitoring module, two versions of data poisoning and one privacy leakage attack (*cf.* Section II) have been analyzed and executed. After that, four datasets, one with "normal" behavior of the Raspberry Pi running ElectroSense and three with anomalies produced by each one of the previous cyberattacks, have been created. After collecting these datasets aligned with C.RQ2, a suitable methodology was followed to clean the data, scale features, select and train unsupervised ML/DL models, and evaluate them. After performing these steps, Table II shows results obtained by an AutoEncoder trained with the normal behavior and evaluated with the remaining datasets. The selected Autoencoder shows three layers of 32, 16, 32 nodes per layer, 20% of the contamination factor, a *relu* activation function, and 50 epochs.

TABLE II
ANOMALY DETECTION RESULTS

Behavior	Normal	Privacy leakage	Data Poisoning 1: Data Injection	Data Poisoning 2: Random Noise
Accuracy	92.9%	100%	100%	9.3%

As of Table II the current model is able to detect three of the four existing behaviors, providing low performance in the detection of the second data poisoning attack due to its similarity with the "normal" behavior.

IV. SUMMARY AND NEXT STEPS

This paper derived relevant research questions concerning device behavior fingerprinting to identify identical devices and detect anomalies produced by cyberattacks. The goals and current status of the two ongoing research projects, TREASURE and CyberSpec, aligning with these research lines, were introduced. In preliminary conclusions, both projects show promising results, since five identical Raspberry Pis are correctly recognized and anomalies generated by two different cyberattacks affecting the data integrity and confidentiality of ElectroSense devices are also detected.

As future work, the validation of current results is key as well as improving the performance of the anomaly detection mechanism with new Machine/Deep Learning (ML/DL) algorithms. Besides, since the proposed platform is independent of the IoT scenario, it is planned to deploy it in other use cases, too. Additionally, further objectives and research questions related to the robustness and privacy management of security solutions are going to be addressed.

ACKNOWLEDGMENTS

This work has been partially supported by *(a)* the Swiss Federal Office for Defense Procurement (armasuisse) with the TREASURE (R-3210/047-31) and CyberSpec (CYD-C-2020003) projects and *(b)* the University of Zürich UZH.

REFERENCES

[1] S. Rajendran, R. Calvo-Palomino, M. Fuchs, B. V. den Bergh, H. Cordobés, D. Giustiniano, S. Pollin, and V. Lenders, "Electrosense: Open and Big Spectrum Data," *IEEE Communications Magazine*, vol. 56, no. 1, pp. 210–217, January 2018.

[2] P. M. S. Sánchez, J. M. J. Valero, A. H. Celdrán, G. Bovet, , M. G. Pérez, and G. M. Pérez, "A Survey on Device Behavior Fingerprinting: Data Sources, Techniques, Application Scenarios, and Datasets," *IEEE Communications Surveys & Tutorials*, In press.

[3] L. Xiao, D. Jiang, D. Xu, W. Su, N. An, and D. Wang, "Secure Mobile Crowdsensing based on Deep Learning," *China Communications*, vol. 15, no. 10, pp. 1–11, October 2018.

[4] C. Miao, Q. Li, H. Xiao, W. Jiang, M. Huai, and L. Su, "Towards Data Poisoning Attacks in Crowd Sensing Systems," in *18th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2018)*, Los Angeles, California, U.S.A., 2018, pp. 111–120.

Sesión de Investigación A3:
Ciberataques e inteligencia de amenazas

Sesión de Investigación A3: Ciberataques e inteligencia de amenazas
http://doi.org/10.18239/jornadas_2021.34.16

A Review of Cyberattacks on Miniature Brain Implants to Disrupt Spontaneous Neural Signaling

Sergio López Bernal [1], Alberto Huertas Celdrán[2], Gregorio Martínez Pérez[1]
[1]*Department of Information and Communications Engineering, University of Murcia*, 30100 Murcia, Spain
{slopez, gregorio}@um.es
[2]*Communication Systems Group CSG, Department of Informatics IfI, at the University of Zurich UZH*, CH 8050 Zürich, Switzerland
huertas@ifi.uzh.ch

***Abstract*—Brain Computer-Interfaces are bidirectional systems focused on communicating computers with the brain, allowing acquisition of neural activity and neurostimulation. Considering the latter, existing vulnerabilities in stimulation devices introduce the possibility to perform neural cyberattacks to disrupt spontaneous neuronal behavior. In this regard, this paper defines two novel neural cyberattacks, Neuronal Flooding and Neuronal Scanning, and three metrics to measure their impact: number of spikes, percentage of shifted spikes, and dispersion of spikes. These cyberattacks have been implemented using a neuronal simulator, concluding that both have a considerable impact on neural activity, although their action mechanism and impact differ. Neuronal Flooding is more suitable to introduce an immediate impact, while Neuronal Scanning generates a higher impact in the long-term.**

***Index Terms*—BCI, cybersecurity, neural cyberattacks, brain**

Tipo de contribución: *Investigación ya publicada*

I. Introduction

Brain-Computer Interfaces (BCIs) allow the acquisition of neural data and the stimulation and inhibition of brain activity. These functionalities are extensively used in medical scenarios to treat neurodegenerative conditions, such as Parkinson's disease. Based on their advantages, BCIs have evolved in recent years, where nanotechnology plays an essential role in the current research field, highlighting neural dust and Neuralink as promising solutions.

Despite the advantages of BCIs, the literature has demonstrated that they can be vulnerable to multiple common cyberattacks, having a tremendous impact on the integrity, confidentiality, and availability of data and services, as well as the users' safety [1]. Moreover, these novel BCIs focused on stimulation present vulnerabilities that attackers could exploit to access their neurostimulation capabilities and damage patients' medical conditions.

This work summarizes the research published in [2], which defined and implemented two neural cyberattacks exploiting existing vulnerabilities of neurostimulation BCIs to disrupt spontaneous neural activity. In particular, Neuronal Flooding (FLO) overstimulates a given set of neurons in a determined time instant, while Neuronal Scanning (SCA) sequentially targets the overstimulation of particular neurons emulating a port scanning. To demonstrate their feasibility, this work applied them over a simulated and realistic portion of mice visual cortex, presenting three metrics to evaluate the impact of these cyberattacks: number of spikes, percentage of shifted spikes, and dispersion of spikes. A spike represents the impulse emitted from a neuron to its interconnected neurons.

II. Use Case and Experimental Setup

Since the knowledge of precise neocortical synaptic connections in mammalian is an open challenge, this work obtained a neuronal topology from training a Convolutional Neural Network (CNN) focused on solving the problem of a mouse trying to exit a particular maze. Although the complexity of biological neuronal networks from the visual cortex differs from the behavior of CNNs, the literature has demonstrated certain similarities. Based on that, this work trains a CNN with three layers, composed of 200, 72, and four nodes, respectively. The input of this system is a maze of 7x7 cells indicating the current mouse location (see Figure 1). Once trained the model, the mouse is able to find the optimal path to exit the maze.

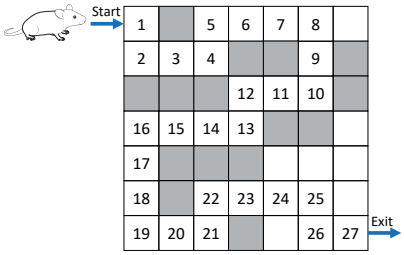


Figure 1. Maze representing the optimal path to find the exit.

The neuronal topology of mice visual cortex obtained from the trained CNN was represented in a biological neuronal simulator to measure the impact of the neural cyberattacks later. For that, the well-known Izhikevich neuronal model is used to represent neuronal activity. Equation 1 presents its behavior, where v represents the membrane potential (voltage) of a neuron, u the membrane recovery, and the parameters a, b, c and d are used to model different neuronal behaviors.

$$\begin{aligned} v' &= 0.04v^2 + 5v + 140 + u + I \\ u' &= a(bv - u) \\ if\, v \geqslant 30mV, then &\begin{cases} v \leftarrow c \\ u \leftarrow u + d \end{cases} \end{aligned} \tag{1}$$

Moreover, the mouse's movement across the maze was introduced to the model using the *I* parameter, representing an external input to the model. After that, this work implements a simulation of 27 seconds, where the mouse stays one second per position of the optimal path. Finally, Figure 2 presents the association between layers of the CNN and those corresponding to the biological visual cortex.

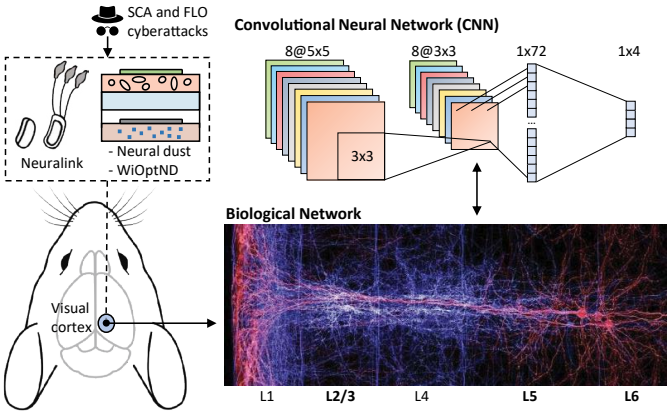


Figure 2. Experimental setup with the relationship between a biological and an artificial neural network.

III. Results Analysis

FLO cyberattacks were performed at 50ms after starting the simulation, representing in Figure 3 its temporal impact on the number of spikes per position. This work presents the total number of spikes across all neuronal layers resulting from attacking 55 and 105 randomly selected neurons within the first layer and implementing ten executions for each size. The plot indicates that FLO configurations considerably reduce the number of spikes when the mouse progresses in the maze. Moreover, these results highlight that attacking the first position propagates the disruption until the end of the simulation, not converging with the spontaneous case. In fact, a FLO cyberattack targeting 105 neurons induces a reduction of around 19% of spikes in the last position within the range of values defined in the Y-axis, being the most affected position.

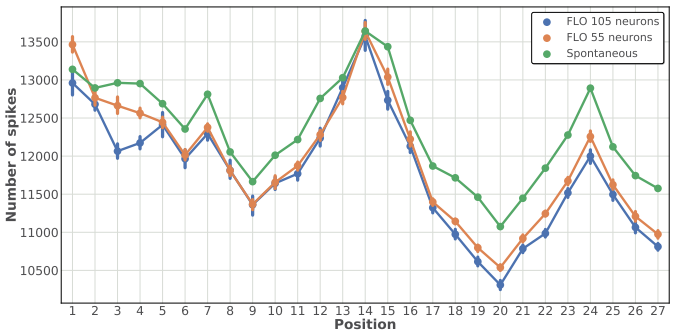


Figure 3. Impact of FLO neuronal cyberattacks on the number of spikes.

Figure 4 presents the results for SCA cyberattacks, which sequentially attack all neurons of the first layer without repetitions. SCA achieved a substantial reduction of spikes increased over time, achieving its maximum impact at position 27 by inducing a reduction of approximately a 31% of spikes. These results can be explained by the incremental behavior of the attack, which amplifies the impact when moving to the end of the maze. Comparing these results with FLO cyberattacks, SCA presented a higher impact than the most aggressive FLO configuration, mainly represented in the last position of the maze. These differences are motivated by the inner behavior of the cyberattacks, where SCA incrementally augments the number of affected neurons. Analyzing the behavior of each layer from this metric, the most relevant differences resided in the third layer, being the deeper one.

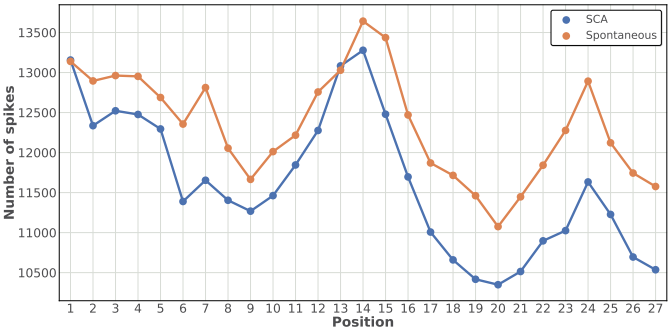


Figure 4. Impact of SCA neuronal cyberattacks on the number of spikes.

In terms of the percentage of spike shifts, FLO presented a higher impact when aggregating the spikes of the three layers. Individually, the main difference is in the first layer, where SCA duplicated its impact based on the number of attacked neurons. The rest of the individual layers did not offer substantial differences. Moving to the dispersion metric, the experiments concluded that FLO had a higher impact during the first five positions of the optimal path based on its synchronized attacking behavior. However, the trend of SCA was more damaging. Due to room restrictions, this work only indicates the most relevant findings for the last two metrics.

This comparative highlighted that the particularities of each cyberattack differently affected spontaneous neural signaling. FLO was more effective for altering neuronal behavior in a short period, while SCA is more harmful in the long term, requiring more time to target a significant number of neurons.

IV. Conclusion

This work introduces two neuronal cyberattacks, FLO and SCA, able to disrupt spontaneous neuronal activity. To test them, this paper presents a realistic simulation of a portion of a mouse visual cortex. Due to the current limitations in realistic neuronal simulations, the neuronal topology has been obtained from training a CNN to solve the problem of a mouse trying to exit a maze. The experimentation indicates that FLO is adequate for affecting a set of neurons in a particular moment, presenting the highest impact in the instants after the attack. In contrast, SCA is more damaging in the long term, based on its inner incremental behavior.

References

[1] S. López Bernal, A. Huertas Celdrán, G. Martínez Pérez, M. T. Barros, and S. Balasubramaniam, "Security in brain-computer interfaces: State-of-the-art, opportunities, and future challenges," *ACM Computing Surveys*, vol. 54, no. 1, Jan. 2021.

[2] S. López Bernal, A. Huertas Celdrán, L. Fernáandez Maimó, M. T. Barros, S. Balasubramaniam, and G. Martínez Pérez, "Cyberattacks on miniature brain implants to disrupt spontaneous neural signaling," *IEEE Access*, vol. 8, pp. 152 204–152 222, 2020.

Sesión de Investigación A3: Ciberataques e inteligencia de amenazas
http://doi.org/10.18239/jornadas_2021.34.17

Extended Abstract – AVCLASS2: Massive Malware Tag Extraction from AV Labels

Silvia Sebastián*§, Juan Caballero*
*IMDEA Software Institute §Universidad Politécnica de Madrid
0000-0001-7675-0535, 0000-0003-2962-1348
silvia.sebastian@imdea.org, juan.caballero@imdea.org

***Abstract*—Tags can be used by malware repositories and analysis services to enable searches for samples of interest across different dimensions. Automatically extracting tags from AV labels is an efficient approach to categorize and index massive amounts of samples. Recent tools like AVCLASS and EUPHONY have demonstrated that, despite their noisy nature, it is possible to extract family names from AV labels. However, beyond the family name, AV labels contain much valuable information such as malware classes, file properties, and behaviors.**

This work presents AVCLASS2, an automatic malware tagging tool that given the AV labels for a potentially massive number of samples, extracts clean tags that categorize the samples. AVCLASS2 uses, and helps building, an *open taxonomy* that organizes concepts in AV labels, but is not constrained to a predefined set of tags. To keep itself updated as AV vendors introduce new tags, it provides an update module that automatically identifies new taxonomy entries, as well as tagging and expansion rules that capture relations between tags. We have evaluated AVCLASS2 on 42M samples and showed how it enables advanced malware searches and to maintain an updated knowledge base of malware concepts in AV labels.

***Index Terms*—AV labels, Tag Malware, Taxonomy.**

Type of contribution: *Published research.*The full version of this paper appears in the Proceedings of the 2020 Annual Computer Security Applications Conference (ACSAC 36)

I. INTRODUCTION

Tags are keywords assigned to data objects (e.g., documents, videos, images) to categorize them and to enable efficient searches along different dimensions such as properties, ownership, and origin. Tags can be used by malware repositories and malware analysis services for enabling searches for samples of interest. Malware tags can be manually produced by analysts, or output by analysis tools such as sandboxes and signature-matching engines. In both cases, each analyst and tool developer may use its own *vocabulary*, i.e., their own custom set of tags. This is similar to user tagging, or *folksonomies*, in Web services [1], [2], which are known to lead to issues such as tags produced by different entities being *aliases* (or synonyms) for the same concept, some tags being highly specific to the entity producing them, and a tag from an entity corresponding to multiple tags from another entity. To address these issues, standards such as Malware Attribute Enumeration and Characterization (MAEC) [3] define a language for sharing malware analysis results. However, they have low adoption due to their use of rigid *controlled vocabularies* (i.e., predefined tags) that may not always match analyst needs, require frequent updates, and are necessarily incomplete.

Detection labels by anti-virus engines (i.e., *AV labels*) are an instance of the above problem. An AV label can be seen as a serialization of the tags an AV engine assigns to the sample. Since tags are selected by each AV vendor rather independently, inconsistencies among labels from different vendors are widespread, as frequently observed in malware family names [4]–[9]. Recent tools like AVCLASS [8] and EUPHONY [9] demonstrate that, despite their noisy nature, it is possible to extract accurate family tags from AV labels. However, beyond the family name, AV labels may also contain much valuable information such as the class of malware (e.g., *ransomware*, *downloader*, *adware*), file properties (e.g., *packed*, *themida*, *bundle*, *nsis*) and behaviors (e.g., *spam*, *ddos*, *infosteal*).

Automatically extracting malware tags from AV labels is an important, but challenging, problem. It enables to cheaply categorize and index massive amounts of samples without waiting for those samples to be statically analyzed or executed in a sandbox. And, since different AV vendors may execute a sample, the extracted tags may accumulate behaviors observed under different conditions. Furthermore, AV labels may encode domain knowledge from human analysts that is not produced by automated tools. Once obtained, the tags can be used to enable efficient search of samples of a specific class, type, family, or with a specific behavior. And, the identified samples can then be used as ground truth for machine learning approaches [10]–[13].

This work presents AVCLASS2, an automatic malware tagging tool that given the AV labels for a potentially massive number of samples, extracts for each input sample a clean set of tags that capture properties such as the malware class, family, file properties, and behaviors. AVCLASS2 ships with a default *open taxonomy* that classifies concepts in AV labels into *categories*, as well as default *tagging rules* and *expansion rules* that capture relations between tags. In contrast to closed taxonomies, AVCLASS2 does not mandate a predefined set of tags. Instead, unknown tags in AV labels, e.g., a new behavior or family name, are also considered. AVCLASS2 has an *update module* that uses tag co-occurrence to identify relations between tags. Those relations are a form of generalized knowledge that the update module uses to automatically generate taxonomy updates, tagging, and expansion rules to keep the tool updated as AV vendors introduce new tags. Thus, AVCLASS2 can maintain an updated knowledge base of malware concepts in AV labels.

AVCLASS2 builds on AVCLASS [8]. The goal is to evolve from a malware labeling tool, focused exclusively on malware

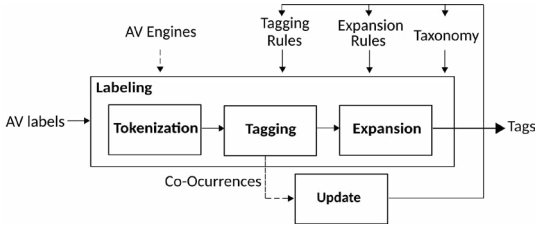


Fig. 1: AVCLASS2 architecture.

families, to a malware tag extraction tool that provides rich threat intelligence by extracting and structuring all useful information in AV labels. Thus, AVClass2 inherits AVClass major properties: scalability, AV engine independence, platform-agnostic, no access to samples required (only to their labels), and open source.

We have evaluated AVCLASS2 on 42M samples and compared it with AVCLASS and EUPHONY, the two state-of-the-art malware family labeling tools. We show how the tags AVCLASS2 extract enable rich searches on malware samples, not possibly with existing tools, how the extracted tags are complementary to those already in use by popular malware repositories such as VirusTotal [14], and how the update module can be used to maintain an updated knowledge base of malware concepts in AV labels.

The main properties of AVCLASS2 are:

- Automatically extracts tags from AV labels that categorize malware samples according to their malware class, family, behaviors, and file properties.
- Uses and builds an open taxonomy that does not use a closed set of tags, and thus can handle new tags introduced over time by AV vendors.
- Can expand the input taxonomy, tagging rules, and expansion rules, by generalizing relations found in AV labels, allowing to maintain over time an up-to-date knowledge base of malware concepts in AV labels.
- Evaluated on 42M samples and compared with the two state-of-the-art malware family labeling tools [8], [9].
- Open source[1].

II. APPROACH

The architecture of AVCLASS2 is shown in Figure 1. It comprises of two modules: *labeling* and *update*. The labeling module takes as input the AV labels assigned by multiple AV engines to the samples, an optional list of AV engines whose labels to use, a set of tagging rules, an optional set of expansion rules, and a taxonomy that classifies tags into categories and captures parent-child relationships between tags in the same category. For each input sample, it outputs a set of tags ranked by the number of AV engines. AVCLASS2 ships with default tagging rules, default expansion rules, and a default taxonomy. Thus, AVCLASS2 can be used out-of-the-box without the need for any configuration. However, AVCLASS2 is fully configurable, so the analyst can easily plug-in its own tagging rules, expansion rules, and taxonomy.

Malware is an ever-evolving ecosystem. Over time, known families exhibit new behaviors and file properties (e.g., novel obfuscations); new families are introduced with their corresponding aliases; and novel malware classes are occasionally created. The update module tackles the challenge of keeping AVCLASS2 up-to-date with this natural evolution. The update module takes as input co-occurrence statistics output by the labeling module, tagging rules, expansion rules, and taxonomy. It first identifies strong relations between tags, which generalize knowledge beyond individual samples, e.g., that a family is ransomware or sends SMS. Then, it uses inference rules on the relations to automatically propose new tagging rules, new expansion rules, and taxonomy updates, which are then fed back to the labeling.

III. EVALUATION

We have evaluated AVCLASS2 on 42M samples and compared it with AVCLASS and EUPHONY, the two state-of-the-art malware family labeling tools. We have also evaluated the update module for updating the taxonomy, tagging, and expansion input files. For more details about AVCLASS2 and its evaluation, we refer the reader to the full publication [15].

[1] https://github.com/malicialab/avclass.

REFERENCES

[1] H. Halpin, V. Robu, and H. Shepherd, "The Complex Dynamics of Collaborative Tagging," in *International Conference on World Wide Web*, 2007.
[2] C. Körner, D. Benz, A. Hotho, M. Strohmaier, and G. Stumme, "Stop thinking, start tagging: Tag semantics emerge from collaborative verbosity," in *International Conference on World Wide Web*, 2010.
[3] "Malware Attribute Enumeration and Characterization 5.0," 2017, https://maec.mitre.org/.
[4] M. Bailey, J. Oberheide, J. Andersen, Z. M. Mao, F. Jahanian, and J. Nazario, "Automated Classification and Analysis of Internet Malware," in *International Symposium on Recent Advances in Intrusion Detection*, 2007.
[5] J. Canto, M. Dacier, E. Kirda, and C. Leita, "Large Scale Malware Collection: Lessons Learned," in *IEEE SRDS Workshop on Sharing Field Data and Experiment Measurements on Resilience of Distributed Computing Systems*, 2008.
[6] F. Maggi, A. Bellini, G. Salvaneschi, and S. Zanero, "Finding Non-Trivial Malware Naming Inconsistencies," in *International Conference on Information Systems Security*, 2011.
[7] A. Mohaisen and O. Alrawi, "AV-Meter: An Evaluation of Antivirus Scans and Labels," in *Detection of Intrusions and Malware, and Vulnerability Assessment*, 2014.
[8] M. Sebastián, R. Rivera, P. Kotzias, and J. Caballero, "AVClass: A Tool for Massive Malware Labeling," in *International Symposium on Research in Attacks, Intrusions and Defenses*, 2016.
[9] M. Hurier, G. Suarez-Tangil, S. K. Dash, T. F. Bissyandé, Y. Le Traon, J. Klein, and L. Cavallaro, "Euphony: Harmonious Unification of Cacophonous Anti-Virus Vendor Labels for Android Malware," in *International Conference on Mining Software Repositories*, 2017.
[10] U. Bayer, P. M. Comparetti, C. Hlauschek, C. Kruegel, and E. Kirda, "Scalable, Behavior-Based Malware Clustering," in *Network and Distributed System Security*, 2009.
[11] R. Perdisci, W. Lee, and N. Feamster, "Behavioral Clustering of HTTP-Based Malware and Signature Generation Using Malicious Network Traces," in *USENIX Symposium on Networked Systems Design and Implementation*, 2010.
[12] K. Rieck, P. Trinius, C. Willems, and T. Holz, "Automatic Analysis of Malware Behavior using Machine Learning," *Journal of Computer Security*, vol. 19, no. 4, 2011.
[13] D. Arp, M. Spreitzenbarth, M. Huebner, H. Gascon, and K. Rieck, "Drebin: Efficient and Explainable Detection of Android Malware in Your Pocket," in *Network and Distributed System Security*, 2014.
[14] 2020, https://virustotal.com/.
[15] S. Sebastin and J. Caballero, "AVClass2: Massive Malware Tag Extraction from AV Labels," in *Proceedings of the 2020 Annual Computer Security Applications Conference*, Virtual Event, December 2020.

Sesión de Investigación A3: Ciberataques e inteligencia de amenazas
http://doi.org/10.18239/jornadas_2021.34.18

A review of Cybersecurity Threat Intelligence Knowledge Exchange based on blockchain

Raúl Riesco Granadino
Universidad Politécnica de Madrid
raul.riesco.granadino@alumnos.upm.es

Xavier Larriva-Novo
Universidad Politécnica de Madrid
x.larriva@alumnos.upm.es

Victor A. Villagrá
Universidad Politécnica de Madrid
victor.villagra@upm.es

***Resumen*—Although cyber threat intelligence (CTI) exchange is a theoretically useful technique for improving security of a society, the potential participants are often reluctant to share their CTI and prefer to consume only, at least in voluntary based approaches. Such behavior destroys the idea of information exchange. On the other hand, governments are forcing specific entities and operators to report them specific incidents depending on their impact. Obligations and sanctions are usually discouraging participants to share information voluntarily. We propose a paradigm shift of cybersecurity information exchange by introducing a new way to encourage all participants involved, at all levels, to share relevant information dynamically. Participants will have new and specific incentives to share, invest and consume threat intelligence and risk intelligence information depending on their different roles (producers, consumers, investors, donors and owner). Our proposal leverages from standards like Structured Threat Information Exchange (STIX™), W3C semantic web standards and from the Ethereum Blockchain to enable a workspace of knowledge related to behavioral threat intelligence patterning to characterize tactics, techniques and procedures (TTP) introducing new type of incentives.**

***Index Terms*—STIX™, SWRL, OWL, Dynamic Risk Management (DRM), Cyber Threat Intelligence (CTI) , Ethereum Blockchain Smart contracts.**

Tipo de contribución: *Investigación ya publicada [1]*

I. Introduction

When a certain threat shares the same motivation among different organizations, all of them are in danger. Once a piece of knowledge about such threat is available (the threat is characterized somehow), all potential affected organizations could benefit from having access to that knowledge. Until today, Indicators of Compromise (IoC) are used as the de facto type of information to be shared about threats, especially if we want automatic and actionable intelligence.

On the other hand, unsuccessful voluntary sharing has several and different root causes. Several studies have been analyzing why people is often reluctant to share [2] [3] [4] [5] [6] [7].

In the paper published we presented in its table 1 an inventory of the open challenges and limits of existing solutions in information sharing nowadays. The table includes references from the bibliography to support each concept. All the open challenges can be grouped into the following categories:

- The lack of trust (infrastructure, admin and peers)
- The lack of incentives (business cases) provided to all roles simultaneously.
- The asymmetry between consumers and producers.
- The reliability and accuracy of CTI data.
- The lack of semantics (unambiguous data) to exchange knowledge (beyond single pieces of data).
- The effectiveness and efficiency of platforms (automation).

We propose a solution, combining the use of semantic web ontologies, STIX™and Ethereum Blockchain, in order to cover all open challenges at the same time. This summary of a published paper will elaborate more on the lack of incentives, asymmetry and partially on effectiveness and efficiency. The contributions are the following:

- A new incentive model for Cyber Threat Intelligence (CTI) sharing, based on Ethereum Blockchain and a the CTI token (ERC20 compliant).
- A new enhanced version of peers. A semantic approach of CTI sharing systems within Dynamic Risk Management (DRM) processes. For that, we support the exchange of CTI semantic web algorithms (beyond the exchange of IoC data) in the format of SWRL and an OWL enhanced version of STIX™v2, at all levels.
- Simulations for model optimization, and experimentation, to demonstrate its benefits and its limits, especially in terms of costs.

II. Design

II-A. Model

We will define the dynamic or evolutionary benefit B at time t of a entity x as seen in equation 1

$$B_t(x, p_c) = I_t(x, p_c) - C_t(x, p_c) \tag{1}$$

where: p_c is the token prize at time t, I_t is the income (gross benefit without cost) at time t and C_t is the cost at time t.

In our proposal, CTI Data producers and the owner, will receive cash (IC_t), due to the introduction of taxes into each CTI read operation. At the same time, these payments will also represent variable costs VC_t for CTI Data consumers. CTI Data producers will also receive tokens ITK_t in case they

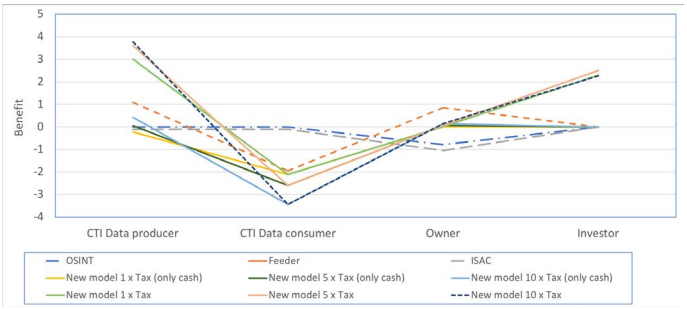


Figura 1. Montecarlo simulation to compare models

decide to invest when uploading CTI Data to the system. We have an evolutionary token value p_c, depending on the balance of the smart contract. The smart contract will sum the cash of all the investments to the 40 percent received from the applied taxes. Having new and balanced incentives we foster a sustainable growing community.

On the other hand, the total cost C_t of an entity x, can be defined as seen in equation 2:

$$C_t(x, p_c) = FC_t(x, p_c) + VC_t(x, p_c) \quad (2)$$

where:

- FC_t is the fixed cost at time t,
- VC_t is the variable cost at time t.

To simplify our calculations, we propose zero fixed costs ($FC_t = 0$), due to the use of a decentralized infrastructure (by definition only has variable tx costs). Variable costs VC_t will be associated to any transaction. We have two types of variable costs: blockchain network fees (gas) and taxes. The gas used to write (store) CTI rules is much higher than the gas used to read (query) those rules. In order to calculate these type of variable costs, we implemented, deployed and evaluated a draft smart contract.

Equation 1 can then be divided into different equations as seen in 3:

$$B_t(x, p_c) = BC_t(x, p_c) + BTK_t(x, p_c) \quad (3)$$
$$BC_t(x, p_c) = IC_t(x, p_c) - CC_t(x, p_c) \quad (4)$$
$$BTK_t(x, p_c) = ITK_t(x, p_c) - CTK_t(x, p_c) \quad (5)$$
$$I_t = IC_t(x, p_c) + ITK_t(x, p_c) \quad (6)$$
$$C_t = CC_t(x, p_c) + CTK_t(x, p_c) \quad (7)$$

where:

- BC_t is the cash benefit at time t,
- BTK_t is the token benefit at time t,
- IC_t is the cash income (cash gross benefit) at time t,
- ITK_t is the cash income (token gross benefit) at time t,
- CC_t is the cost paid in cash at time t,
- CTK_t is the cost paid in tokens at time t.

II-B. Market growth: Montecarlo simulation of new incentives

In order to better design the system, we have made theoretical calculations as seen in figures 1,2. We made a Montecarlo simulation to define the minimum tax value of our system equivalent to a breakeven calculation (a positive **cash benefit** $BC_t(x, p_c) > 0$ of the CTI Data producers). We also used simulations to forecast the potential market growth, in case of introducing our new incentive model. It helped us to better design and select which are the key variables of the model, specially the tax fees. In addition to this, it enabled us to evaluate it versus current legacy information sharing platforms (OSINT, Feeders and ISAC), taking into consideration the specificities of each role.

III. Conclusions

Any entity is exposed to several cybersecurity threats everyday. CTI data is considered then, one of the most valuable assets of any organization, to better detect, prevent and response to cybersecurity threats on time. Its value is related to the quality, understood as the timing (when and how fast is available), the reliability and accuracy of the data. Because of that, there is a very high demand of CTI data, however, there is a limited size of providers, compared to the demand size. Entities are using different taxonomies without enough expressivity to define complex relationships, which are needed to create context-aware or behavioral rules. STIX™is a promising standard but it still lacks of semantics. Furthermore, users are reluctant to share.Trust is one of the main reasons behind, but there are much more reasons.

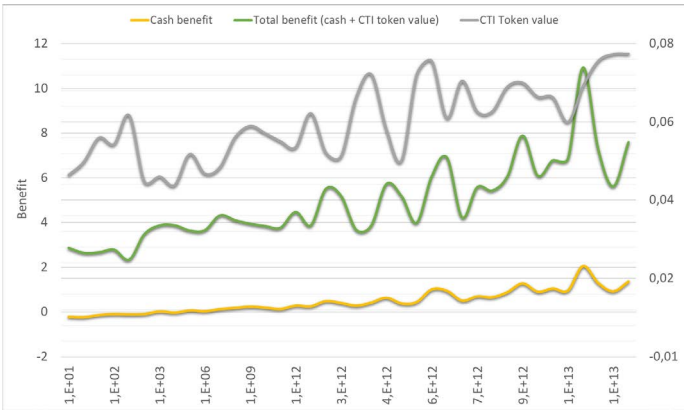


Figura 2. Evolutionary tax and CTI Value

This paper presents a new model, to provide Cybersecurity Intelligence Exchange, based on Blockchain. It provides new economic incentives to all roles involved, as well as an enhanced version of the peers. In order to operate, share and consume semantic advanced intelligence automatically, a semantic reasoner is considered a key and a powerful building block but it needs to understand the data and it should be standard. For that reason we created a OWL version of STIX v2 and our algorithms use SWRL to make relationships between STIX objects standard as well.

Referencias

[1] Riesco et al. en *Çybersecurity Threat Intelligence Knowledge Exchange based on blockchain"*, Telecommunication Systems (2019) https://link.springer.com/article/10.1007/s11235-019-00613-4

[2] NIST Guide to CTI sharing, *"Guide to Cyber Threat Information Sharing"*, Special Publication 800-150 (2016) http://dx.doi.org/10.6028/NIST.SP.800-150

[3] Vishik C., Sheldon F., Ott D., *.Economic Incentives for Cybersecurity: Using Economics to Design Technologies Ready for Deployment"*, in Reimer H., Pohlmann N., Schneider W. (eds) ISSE 2013 Securing Electronic Business Processes. Springer Vieweg, Wiesbaden (2013);133-147. https://doi.org/10.1007/978-3-658-03371-2_12

[4] Tosh, D., Sengupta, S., Kamhoua, CA., Kwiat, KA. *.Establishing evolutionary game models for CYBer security information EXchange (CYBEX)"*, In J Comput Syst Sci. (2018) ;98:27-52. https://doi.org/10.1016/j.jcss.2016.08.005

[5] Skopik, F, Settanni, G, Fiedler, R. *A problem shared is a problem halved: a survey on the dimensions of collective cyber defense through security information sharing*, in Comput Secur. 2016;60:154-176. https://doi.org/10.1016/j.cose.2016.04.003

[6] de Fuentes, JM., González-Manzano, L., Tapiador, J., Peris-Lopez, P. *"PRACIS: privacy-preserving and aggregatable cybersecurity information sharing"*, in Comput Secur. 2017;69:127-141. https://doi.org/10.1016/j.cose.2016.12.011

[7] Ring, T. *"Threat intelligence: why people don't share"*, In Comput Fraud Secur. 2014;2014(3):5-9. https://doi.org/10.1016/S1361-3723(14)70469-5

http://doi.org/10.18239/jornadas_2021.34.19

Hacia un modelo analítico de APT basado en factores técnico-geopolíticos

Lorena Gonzalez-Manzano
U. Carlos III de Madrid
ORCID:0000-0002-3490-621X
lgmanzan@inf.uc3m.es

Jose M. de Fuentes
U. Carlos III de Madrid
ORCID:0000-0002-4023-3197
jfuentes@inf.uc3m.es

Cristina Ramos
U. Carlos III de Madrid
ORCID:0000-0002-4917-8905
crramosi@pa.uc3m.es

Florabel Quispe
U. Carlos III de Madrid
ORCID:0000-0001-8529-4658
fquispe@der-pu.uc3m.es

***Resumen*—Las amenazas sustentadas por Estados (ASE) son aquellas en que un Estado trata de infligir un daño a otro. Dada su creciente sofisticación, es preciso mejorar su entendimiento. No obstante, hasta el momento no existe una caracterización técnica que permita distinguirlas de otros tipos de amenazas. Además, no se han considerado las cuestiones socio-políticas que rodean a la ciberamenaza. El objetivo de este trabajo es contribuir a paliar las dos carencias en un tipo concreto de ASE, denominada amenaza persistente avanzada (APT, en inglés). En particular, se sientan las bases para un futuro modelo que vincule ambas cuestiones. Para ello, se presentan resultados preliminares centrados en el caso de Rusia.**

***Index Terms*—Amenaza persistente avanzada, APT, geopolítica, factores socio-económicos.**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

Actualmente vivimos en un mundo globalizado en el que tanto los eventos políticos, como económicos, culturales y sociales están cada vez más interconectados [1]. El estudio de las relaciones internacionales ha sido una constante a lo largo de la Historia como medio para alcanzar un mayor conocimiento de la situación socio-económica [2]. No obstante, los avances tecnológicos han introducido nuevas variables que afectan a dicha situación. Ya en 1979 se comenzaban a estudiar las repercusiones que la tecnología moderna tendría a nivel comercial, de seguridad nacional y de desarrollo [3]. Estudios más recientes demuestran que el desarrollo tecnológico es un factor tan determinante en las relaciones internacionales como lo es el mero estado de la economía [4].

Esta incidencia tecnológica se ha visto últimamente acentuada en las relaciones internacionales por las turbulencias políticas ocurridas a nivel mundial [5]. Por ejemplo, la reciente guerra comercial entre Estados Unidos y China provocó que un fabricante de teléfonos móviles chino (Huawei) viese limitado su acceso al sistema operativo más utilizado (Android). Así, el ciberespacio introduce nuevos retos en lo concerniente a las relaciones internacionales [6].

A pesar de los aspectos positivos inherentes del ciberespacio, este entorno introduce nuevos riesgos como son los ciberataques. El reciente informe del Foro Económico Mundial los sitúa como una de las principales amenazas de orden mundial no sólo por su probabilidad de ocurrencia, sino también por su impacto socio-económico [7]. Existen muchos potenciales orígenes para este tipo de ciberamenazas, entre los que están las "amenazas sustentadas por Estados" (en lo sucesivo, ASE) y ponen de manifiesto la existencia de motivaciones políticas en el desarrollo de ciberamenazas [8].

A lo largo de la Historia reciente se han sucedido numerosas ASE. En el año 2007, el código maligno "Stuxnet" fue utilizado contra las centrales nucleares de Irán, afectando a su funcionamiento y considerándose la primera ciberarma [9]. En los últimos tiempos han surgido un conjunto de ASE que colectivamente reciben el nombre de "amenazas persistentes avanzadas", o APT por sus siglas en inglés. Las APT se caracterizan generalmente por tratar de alcanzar tres objetivos: sigilo, resiliencia y anonimato [10]. Así, son amenazas informáticas que desarrollan su actividad tratando de pasar desapercibidas, de modo que típicamente la infección se descubre transcurrido mucho tiempo. Por otro lado, deben ser resilientes frente a los posibles cambios en el entorno de la víctima, tales como la instalación de nuevas herramientas. Finalmente, deben permitir que si la infección se descubre no haya riesgo para el atacante.

A pesar de utilizar programas maliciosos (*malware*), tanto la motivación como los fines de las APT difieren de aquellos. Por ello, es necesario caracterizar las APT para permitir su mejor entendimiento. En este trabajo se presentan unos resultados preliminares que sientan las bases para un futuro modelo analítico. Se comienza determinando el alcance técnico del concepto de APT. Seguidamente, se establecen las variables geopolíticas y socioeconómicas vinculadas con las principales APT. Finalmente se estudia la relación EE.UU.-Rusia para ilustrar la viabilidad del futuro modelo.

El trabajo se divide en las siguientes secciones. La Sección II presenta los trabajos relacionados. La Sección III introduce el análisis de APT, mientras que la Sección IV describe variables geopolíticas y socioeconómicas. En la Sección V se plantea la base del modelo. Finalmente, en la Sección VI se concluye el trabajo.

II. Trabajos relacionados

En los últimos tiempos se han desarrollado numerosos esfuerzos en el estudio de APT. Desde una perspectiva técnica,

Lookheed Martin desarrolló la Cyber Kill Chain para describir los pasos seguidos por un ciberataque [11]. Por otra parte, la corporación estadounidense MITRE ha creado ATT&CK, un repositorio de técnicas, tácticas y procedimientos [12]. Además, han estudiado la agrupación de distintos ataques para asociarlos a múltiples APT, creando el catálogo MITRE Groups, donde se asocian APT con sus supuestos orígenes. Por otra parte, en el ámbito académico, [13] y [14] estudian múltiples APT atendiendo a cómo se despliegan y evolucionan, es decir, desde el momento en el que se compromete un sistema hasta que se toma el control. En cambio, en [15] se examinan algunos de los métodos de ataque más comunes utilizados por APT, así como las herramientas que utilizan. [16], por el contrario, estudia los tipos de comportamientos de múltiples APT y medidas de protección. De forma mucho más exhaustiva, en [17] se analizan trabajos existentes sobre APT, determinando aquellos en los que se pueden encontrar los actores, el tipo y el contenido de cada APT. A modo de compendio sobre los esfuerzos llevados a cabo para el conocimiento del concepto de APT, [18] presenta una revisión sistemática de trabajos donde se discuten métodos y técnicas utilizados por las APT y sus métodos de detección. No obstante, el conjunto de APT que se toma como base es radicalmente pequeño en comparación con la cantidad de actores referidos en el mencionado catálogo de MITRE.

Desde la perspectiva sociopolítica, [19] señala que las motivaciones políticas, socioculturales y económicas están correladas con un conjunto de ciberataques, aunque no relacionados con APT. [20] discute la relación de factores políticos, técnicos y científicos respecto a política en materia de ciberseguridad. Por otro lado, [21] analiza ciberataques considerando la dimensión social basándose en 14 noticias. Sin embargo, considerando las variables estudiadas, [22] es la propuesta más relacionada con el trabajo aquí presentado, si bien su objetivo era analizar la relación entre los ciberataques, incluyendo algunos asociados a APT, y el comercio.

A tenor de lo ya expuesto, los esfuerzos investigadores se han concentrado en estudiar las APT utilizando un enfoque eminentemente técnico, pero generalmente para un subconjunto de grupos y técnicas empleadas. Igualmente, los aspectos geopolíticos y socioeconómicos no han sido incorporados en el análisis. Por ello, hasta el momento no se han establecido vínculos entre APT y las relaciones geopolíticas y socioeconómicas entre Estados.

III. Caracterización técnica de APT

La caracterización técnica de una APT viene marcada por dos factores principales: el contexto de uso y el comportamiento de la propia APT. Cada una de estas cuestiones se aborda en los siguientes epígrafes.

III-A. Uso: Atribución y víctimas

Para abordar la atribución se analiza la actividad de 11 grupos asociados a 7 países según MITRE ATT&CK. Estos grupos fueron escogidos por su elevada peligrosidad de acuerdo con el índice de Thales[1]. Así, se estudiaron 286 APT relacionadas con dichos grupos. En el análisis se consideraron diversas fuentes de datos, incluyendo empresas y proveedores de ciberseguridad como FireEye [23], la Agencia de Ciberseguridad y Seguridad de las Infraestructuras de Estados Unidos [24], plataformas colaborativas como Malpedia [25] y blogs de ciberseguridad independientes como Security Affairs [26].

La Tabla I sintetiza el país de origen y las víctimas de cada grupo. A la vista de los resultados, la mayoría de las APT estudiadas tienen origen norcoreano (136), seguido de ruso (48), chino (37), vietnamita (31), iraní (16), indio (11) y estadounidense (7). En lo que respecta a las víctimas, los estudios avalan que hay mucha más variabilidad, y aunque los países europeos y EE.UU. sufren muchos de los ataques, casi todos los países han sido víctimas en algún momento.

III-B. Comportamiento de las APT. El caso de APT29

Cada grupo de APT materializa sus ataques en forma de campañas. Por ejemplo, APT28 desarrolló las campañas Fancy Bear, Pawn Storm o Sofacy entre otras. En ellas se utilizan distintas técnicas en cada fase del ataque, como puede ser el acceso inicial a los sistemas, la persistencia, la escalada de privilegios o la exfiltración de información.

Atendiendo a la clasificación de técnicas propuestas en MITRE ATT&CK, es posible identificar técnicas concretas para cada campaña. Esto permite conocer las similitudes y diferencias operativas a lo largo del tiempo.

A modo ilustrativo, la Tabla II presenta un resumen de las 10 primeras campañas de APT29, una de los grupos de APT rusas más conocidas. Lleva operando desde 2008 y permanece activa, por ejemplo robando información sobre las vacunas de la COVID[2]. De este grupo se han identificado un total de 30 campañas, cada una utilizando multitud de técnicas.

IV. Análisis geopolítico y socioeconómico

Una de las cuestiones clave de las APT es su potencial utilización con fines estratégicos a nivel estatal. Por ello, es necesario caracterizar la situación socioeconómica y geopolítica entre los países identificados como origen y víctima (recuérdese la Sección III-B).

Para definir el estado socioeconómico se consideran variables objetivas, tales como el índice de desarrollo humano, el producto interior bruto, la cantidad de exportaciones e importaciones, el gasto militar o la inversión directa de países extranjeros. De todas ellas se estudia no sólo su estado actual sino también la evolución temporal.

En lo que se refiere a la situación geopolítica, se analiza la base de datos mundial GDELT para la investigación abierta

[1]https://thalesgroup-myfeed.com/THECYBERTHREATHANDBOOK, último acceso 8 de marzo de 2021.

[2]https://www.ncsc.gov.uk/news/advisory-apt29-targets-covid-19-vaccine-development, último acceso marzo de 2021.

Cuadro I
ORIGEN Y VÍCTIMAS DE APT.

	Origen	Víctimas
APT29	Rusia	AUS, AZE, BEL, BGR, BLR, BRA, CAN, CYP, CZE, DEU, ESP, FRA, GBR, GEO, GRC, HUN, IND, IRL, ISR, JPN, KAZ, KGZ, LBN, LTU, LVA, MNE, NLD, NOR, POL, PRT, ROU, RUS, SVN, TUR, UGA, UKR, USA, UZB
APT10	China	ARE, AUS, BEL, BRA, CAN, CHE, CHN, DEU, FIN, FRA, GBR, HKG, IND, JPN, KOR, MEX, NOR, PHL, SGP, SWE, THA, TWN, USA, VNM, ZAF
APT28	Rusia	AFG, ARM, AUS, AUT, AZE, BEL, BGD, BIH, CHE, CHN, CZE, DEU, DNK, ESP, FIN, FRA, GBR, GEO, HUN, IND, IRN, IRQ, ISR, JOR, KAZ, KGZ, KWT, LBN, LTU, MMR, MNE, MNG, MYS, NLD, OMN, PAK, POL, PRT, ROU, RUS, SAU, SRB, SVK, SWE, SYR, TJK, TKM, TUR, TZA, UKR, USA, UZB, ZAF
APT35	Irán	AFG, ARE, CAN, CHE, DEU, DNK, EGY, ESP, FRA, GBR, IND, IRN, IRQ, ISR, JOR, KWT, MAR, PAK, SAU, SYR, TUR, USA, VEN, YEM
Equation	Estados Unidos	AFG, ARE, BEL, BHR, BRA, CHE, CHN, DEU, DZA, ECU, FRA, GBR, IND, IRN, IRQ, KAZ, KEN, LBN, LBY, MAR, MEX, MLI, MMR, MYS, NGA, PAK, PER, PHL, PSE, QAT, RUS, SDN, SGP, SOM, SYR, TUR, UGA, USA, VEN, YEM, ZAF
APT38	Corea del Norte	AUS, BGD, CHL, CHN, ECU, ESP, FRA, GBR, HKG, IND, JPN, KOR, LUX, MEX, MYS, NGA, NOR, PER, PHL, POL, ROU, RUS, TWN, URY, VNM
APT32	Vietnam	AUS, BGD, CCHN, DEU, DZA, FRA, IDN, IND, IRN, JPN, KHM, LAO, MMR, MYS, NPL, PHL, THA, USA, VNM, ZAF
Lazarus	Corea del Norte	ARG, AUS, BEL, BGD, BRA, CAN, CHE, CHN, CZE, DEU, EGY, ESP, EST, FIN, FRA, GBR, GTM, HKG, HUN, IDN, IND, IRL, IRN, ISR, ITA, JPN, KOR, MEX, MUS, NLD, NOR, NZL, PER, PHL, POL, PRT, RUS, SAU, SGP, SVK, SWE, THA, TUR, TWN, UKR, USA, ZAF
APT12	China	DEU, JPN, TWN
Patchwork	India	AUT, DEU, IDN, IRN, JOR, NOR, POL, ROU, THA

Cuadro II
CAMPAÑAS Y TÉCNICAS USADAS POR APT29. (S.F. = SIN FECHA)

Campañas	Técnicas
Noviembre 2008 - Noviembre 2008. Campaña de PinchDuke contra Chechenia [27].	T1003, T1005, T1041, T1071, T1082, T1083, T1192, T1193, T1203, T1204, T1503
Enero 2009 - Diciembre 2012. Campaña GeminiDuke.	T1005, T1007, T1016, T1057, T1071, T1083, T1087, T1192, T1193, T1203, T1204
2009 - Primavera 2010. Ataque a Georgia, Ministerios de asuntos exteriores de Turquía y Uganda y a un ejercicio de la OTAN [27].	T1003, T1005, T1041, T1071, T1082, T1083, T1192, T1193, T1203, T1204, T1503
Abril 2009 - Abril 2009. Ataque a un foro de pensamiento de EE.UU e instituciones gubernamentales de Polonia y República Checa [27].	T1003, T1005, T1041, T1071, T1082, T1083, T1192, T1193, T1203, T1204, T1503
Junio 2009 - Junio 2009. Ataque al Centro de Información de Georgia de la OTAN [27].	T1003, T1005, T1041, T1071, T1082, T1083, T1192, T1193, T1203, T1204, T1503
Julio 2009 - Julio 2009. Ataque al Ministerio de Defensa de Georgia [27].	T1003, T1005, T1041, T1071, T1082, T1083, T1192, T1193, T1203, T1204, T1503
Primavera 2010 - s.f. Campañas de PinchDuke contra miembros de la Commonwealth [27].	T1003, T1005, T1041, T1048, T1056, T1066, T1068, T1071, T1082, T1083, T1113, T1114, T1115, T1192, T1193, T1203, T1204, T1503
Julio 2010 - Julio 2010. Última campaña de PinchDuke conocida [27].	T1003, T1005, T1041, T1071, T1082, T1083, T1192, T1193, T1203, T1204, T1503
Julio 2010 - Febrero 2013. Campañas de MiniDuke [28].	T1001, T1005, T1008, T1023, T1024, T1057, T1060, T1070, T1082, T1083, T1085, T1102, T1105, T1106, T1107, T1116, T1124, T1129, T1132, T1203, T1497
Diciembre 2011 - s.f. Primera campaña de CozyDuke [29].	T1003, T1008, T1016, T1027, T1033, T1036, T1043, T1048, T1050, T1053, T1059, T1060, T1063, T1064, T1071, T1082, T1085, T1102, T1105, T1113, T1122, T1497

[30]. En particular, se extraen las noticias que afectan a cada par de países origen-víctima y se cuantifica su impacto atendiendo a la escala de Goldstein. Dicha escala asigna un factor de impacto en la estabilidad de una relación para cada uno de los eventos considerados en GDELT. Por ello, puede tomar valores tanto positivos como negativos. Así, la suma de todos los eventos permite representar visualmente el estado de las relaciones entre dos países a lo largo del tiempo. Por ejemplo, la Figura 1 representa el estado de las relaciones entre Rusia y Estados Unidos. En ella, la línea verde representa los eventos positivos y la roja los negativos. En general, los eventos positivos son superiores a los negativos excepto en el periodo de 2012 a 2015.

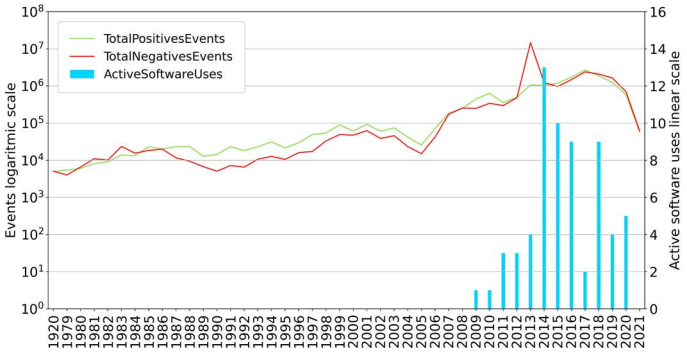


Figura 1. Histórico de eventos entre Rusia y Estados Unidos.

V. Estableciendo vínculos técnicos y socio-políticos

Utilizando los resultados de las secciones anteriores, esta línea de investigación persigue crear un modelo que sea capaz de vincular la perspectiva técnica de las APT con las cuestiones geopolíticas y socioeconómicas. Ello permitirá responder a las siguientes preguntas:

- ¿Existe alguna relación socio-económica entre los orígenes de las APT y las víctimas?
- ¿Existen relaciones entre las técnicas utilizadas en APT y las víctimas?
- ¿Es posible vincular las capacidades de ciberataque más probables a tenor de una circunstancia política concreta?

A modo ilustrativo, en la Figura 1 se presentan, en forma de barras, los ataques presuntamente realizados por Rusia contra Estados Unidos. Se puede observar el máximo en 2014, momento en el que la tensión geopolítica entre ambos países es más elevada y se sitúa en un marco de relaciones difíciles en los años inmediatamente anteriores. Esto es un primer paso para demostrar que los ciberataques de APT tienen relación directa con los eventos geopolíticos.

Para cuantificar la solidez de la relación entre ambas cuestiones, se analiza la correlación entre los tipos de eventos y los ataques de APT. Con el fin de ilustrar este análisis se considera nuevamente el caso de Rusia. Así, se observa una correlación respecto al número de noticias negativas de 0.94 y 0.63 como atacante y atacado respectivamente. Así, al menos en el caso ruso, hay una vinculación entre los eventos geopolíticos negativos realizados por Rusia y los ciberataques por APT realizados y, en menor medida, recibidos.

VI. Conclusiones

Dada la gran cantidad de ciberataques que afectan a los Estados, como son las amenazas persistentes avanzadas (APT), es necesario estudiar la ciberseguridad desde distintos ángulos. Así, este trabajo ha planteado la búsqueda de vínculos entre APT y relaciones geopolíticas y sociopolíticas de los países origen y víctima de los ciberataques. Se han presentado los primeros indicadores que apuntan a la viabilidad del modelo, cuestión que será abordada en el trabajo futuro.

Agradecimientos

Este trabajo está financiado por el proyecto CAVTIONS-CM-UC3M financiado por la UC3M y la CAM; por el MINECO, proyecto ODIO/COW(PID2019-111429RB-C21); por la CAM, proyecto CYNAMON-CM(P2018/TCS-4566), cofinanciado con fondos europeos ESF y FEDER; y el programa de Excelencia para investigadores de Universidad. También agradecemos los comentarios del profesor Ángel Sanchez.

Referencias

[1] J. Baylis, *The globalization of world politics: An introduction to international relations*. Oxford University Press, 2020.

[2] M. Woolcock *et al.*, "The place of social capital in understanding social and economic outcomes," *Canadian journal of policy research*, vol. 2, no. 1, pp. 11–17, 2001.

[3] J. V. Granger, "Technology and international relations," 1979.

[4] T. C. Lawton, J. N. Rosenau, and A. C. Verdun, *Strange Power: Shaping the Parameters of International Relations and International Political Economy: Shaping the Parameters of International Relations and International Political Economy*. Routledge, 2018.

[5] J. N. Rosenau, *Turbulence in world politics: A theory of change and continuity*. Princeton University Press, 2018.

[6] J.-F. Kremer and B. Müller, *Cyberspace and international relations: Theory, prospects and challenges*. Springer, 2013.

[7] W. E. F. (WEF), "The global risks report 2019," 2013.

[8] ——, "Centro criptológico nacional," 2018.

[9] K. Zetter, *Countdown to Zero Day: Stuxnet and the launch of the world's first digital weapon*. Broadway books, 2014.

[10] W. is an Advanced Persistent Threat (APT)?, "Kasperksy inc." 2019.

[11] "L.Martin". "the cyber kill chain". [Online]. Available: https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

[12] T. M. Corporation. Mitre attck®. [Online]. Available: https://attack.mitre.org/

[13] M. Ussath, D. Jaeger, F. Cheng, and C. Meinel, "Advanced persistent threats: Behind the scenes," in *2016 Annual Conference on Information Science and Systems (CISS)*. IEEE, 2016, pp. 181–186.

[14] P. Chen, L. Desmet, and C. Huygens, "A study on advanced persistent threats," in *IFIP International Conference on Communications and Multimedia Security*. Springer, 2014, pp. 63–72.

[15] M. A. Siddiqi and N. Ghani, "Critical analysis on advanced persistent threats," *Int. J. Comput. Appl*, vol. 141, no. 13, pp. 46–50, 2016.

[16] I. Jeun, Y. Lee, and D. Won, "A practical study on advanced persistent threats," in *Computer applications for security, control and system engineering*. Springer, 2012, pp. 144–152.

[17] A. Lemay, J. Calvet, F. Menet, and J. M. Fernandez, "Survey of publicly available reports on advanced persistent threat actors," *Computers & Security*, vol. 72, pp. 26–59, 2018.

[18] A. Alshamrani, S. Myneni, A. Chowdhary, and D. Huang, "A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1851–1877, 2019.

[19] R. Gandhi, A. Sharma, W. Mahoney, W. Sousan, Q. Zhu, and P. Laplante, "Dimensions of cyber-attacks: Cultural, social, economic, and political," *IEEE Technology and Society Magazine*, vol. 30, no. 1, pp. 28–38, 2011.

[20] M. Dunn Cavelty and A. Wenger, "Cyber security meets security politics: Complex technology, fragmented politics, and networked science," *Contemporary Security Policy*, vol. 41, no. 1, pp. 5–32, 2020.

[21] A. C. Sharma, R. A. Gandhi, W. Mahoney, W. Sousan, and Q. Zhu, "Building a social dimensional threat model from current and historic events of cyber attacks," in *2010 IEEE Second International Conference on Social Computing*. IEEE, 2010, pp. 981–986.

[22] W. Akoto, "International trade and cyber conflict: Decomposing the effect of trade on state-sponsored cyber attacks," *Journal of Peace Research*, p. 0022343320964549, 2021.

[23] FireEye. Fireeye. [Online]. Available: https://www.fireeye.com/

[24] US-CERT. Cybersecurity and infrastructure security agency. [Online]. Available: https://us-cert.cisa.gov/

[25] F. FKIE. Malpedia. [Online]. Available: https://malpedia.caad.fkie.fraunhofer.de/

[26] P. Paganini. Security affairs. [Online]. Available: https://securityaffairs.co/wordpress/category/apt

[27] F-Secure, "The dukes 7 years of russian cyberespionage," Tech. Rep., 9 2015. [Online]. Available: https://blog-assets.f-secure.com/wp-content/uploads/2020/03/18122307/F-Secure_Dukes_Whitepaper.pdf

[28] Kaspersky. The miniduke mystery: Pdf 0-day government spy assembler 0x29a micro backdoor. [Online]. Available: https://securelist.com/the-miniduke-mystery-pdf-0-day-government-spy-assembler-0x29a-micro-backdoor/31112/

[29] F-Secure, "Cozyduke," Tech. Rep., 4 2015. [Online]. Available: https://blog-assets.f-secure.com/wp-content/uploads/2019/10/15163418/CozyDuke.pdf

[30] K. Leetaru. The gdelt project. [Online]. Available: https://www.gdeltproject.org/

Sesión de Investigación A3: Ciberataques e inteligencia de amenazas
http://doi.org/10.18239/jornadas_2021.34.20

A Review of Spotting political social bots in Twitter: A use case of the 2019 Spanish general election

Javier Pastor-Galindo[1], Mattia Zago[1], Pantaleone Nespoli[1], Sergio López Bernal[1], Alberto Huertas Celdrán[2], Manuel Gil Pérez[1], José A. Ruipérez-Valiente[1], Gregorio Martínez Pérez[1], Félix Gómez Mármol[1]
[1]*Department of Information and Communications Engineering, University of Murcia*, Spain
{javierpg, mattia.zago, pantaleone.nespoli, slopez, mgilperez, jruiperez, gregorio, felixgm}@um.es
[2]*Department of Informatics, University of Zurich UZH*, Switzerland, huertas@ifi.uzh.ch

***Abstract*—Social media interactions represent one of the primary methods to connect and interact with others, and people rely on them as a primary source of news and information in general. Despite considering these contents as trustworthy knowledge, social media manipulation has been demonstrated as one of the most significant problems of the 21st century. Posts and interactions, especially when related to sensitive subjects like politics, are prone to misinformation and polarised stories. The paper at hand summarises the published article "Spotting Political Social Bots in Twitter" by the same authors, presenting the findings that emerged from the analysis of the Twitter interactions in the 38 days leading to the Spanish elections of Nov. 10th, 2019. The collected data and the subsequent analysis confirms non-negligible, automated and coordinated social bots activities related to the five main political parties.**

***Index Terms*—Social media analysis, misinformation, bots**

Type of contribution: *Already published research*

I. Introduction

Undoubtedly, social media are massively used nowadays by billion users worldwide. While those platforms boost connection among people and spreading opinions, automated accounts heavily populate them to deceive humans toward specific ideologies (or against them), the so-called *political social bots*. Indeed, such bots aim at manipulating public opinion by coordinately amplifying the spread of misinformation via crafted viral trends [1]. Over the years, they progressively become more sophisticated tools capable of mimicking human-like patterns with AI to produce credible social media content. The other bots belonging to the network will interact with it to artificially raise the posts engagement. This unsettling menace has grown at an alarming rate during political events, threatening to jeopardize modern democracies by distorting reality and manoeuvring citizens [2].

Given such a dangerous threat against unaware constituents, the research in-here summarised [3] focused on the November 2019 Spanish general elections to shed light on the social bots' activity. The authors collected Twitter data for six weeks around the five main political parties, political events, and trending topics linked with the election. Both the recollected data [4] and analysis source code[1] are publicly available.

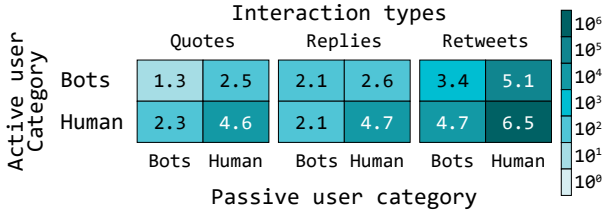


Fig. 1: Interactions involving bots and users by tweet type.

II. Social bots behaviour findings

The collection phase gathered a little more than 5.8 million interactions, of which 88% were retweets, 10% were original tweets, 1% were replies and 1% were quotes, distributed over a total of about 780 thousand different accounts. A bot score has been assigned for each user profile using the heuristic provided by *Botometer*[2]. Users that presented a score below the 75^{th} percentile (about 592 thousand) were considered "human-like" accounts, while users above the 95^{th} percentile (about 40 thousand) were considered "bot-like" accounts. The remaining accounts were considered as "unclear", therefore discarded from the analysis.

Fig. 1 shows, on a logarithmic scale, the volume and type of contents that flowed between the groups. Although human activity was clearly higher, the social bots' activity was also not negligible and devoted to retweet specific human content.

To further examine these interactions' content and precisely to identify clear marks of political polarisation, a machine learning classifier has been designed, developed and thoroughly tested. In the research, 200 verified and politically aligned accounts have been taken into account to create a manually labelled training set. For each tweet, the sentiment analysis (*i.e.*, the machine learning classifier devoted to predict a score for each tweets' text) has been correlated with the main topics and manually labelled with the associated political ideology. Following the same procedure, the model hence inferred the political affiliation for each collected user. Based on the confidence score associated with the model's decision, and taking into account those classified records with at least 80% confidence, there were 825 bots aligned with PP, 1020

[1]https://github.com/CyberDataLab/botbusters-spanish-general-elections

[2]https://botometer.osome.iu.edu/

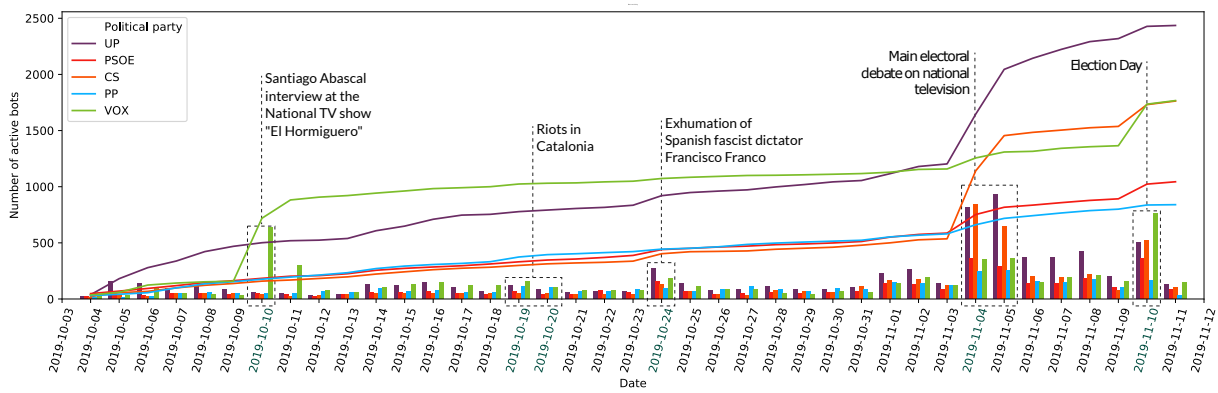


Fig. 2: Active social bots with one-party affinity (cumulative and per day basis).

with PSOE, 1749 with VOX, 1752 with CS and 2417 with UP. For these accounts, Fig. 3 presents their friendship relations.

To further tie the social bots' activities to political events, Fig. 2 counts the number of accounts with their first interaction within the monitoring window during a specific date. The barplot suggests that new or dormant accounts were activated in accord with major political events. Finally, the sentiment analysis also revealed that social bots tend to attack and denigrate political opponents rather than providing encouragement and positive interactions regarding the affiliated political party. Fig. 4 presents the sentiment analysis' statistical metrics for those tweets with a precise and unique political target. In the figure, the zero value indicates an extremely negative attitude, while a value of one refers to the opposite.

III. Discussion and Future Work

Social media's nature favours the creation of echo chambers, *i.e.*, bubbles in which the contents are presented as polarised stories that often include misinformation, fake news, and distorted contents. As such, charts like the one presented in Figure 3 could indicate coordinated accounts with common political goals, or even botnets, aiming to propagate particular ideas in an attempt to affect the beliefs of social media users.

When featuring in the political situation of late 2019, the situation appears even more dramatic: November's general

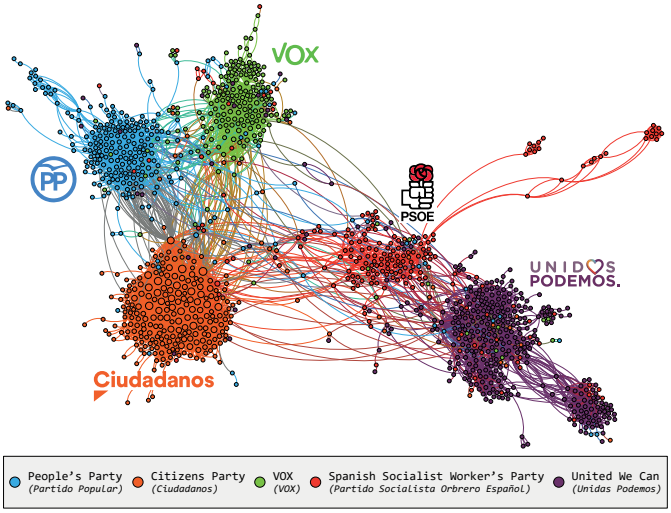


Fig. 3: Friendship relationship among bots.

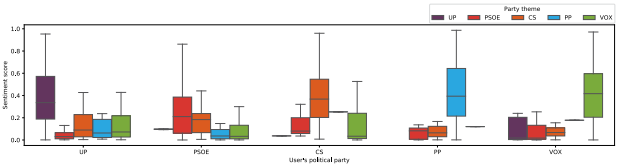


Fig. 4: Sentiment of social bots against party themes.

election was the second attempt to achieve a majority in the Spanish parliament since the PSOE candidate did not obtain enough support to become Prime Minister in the previous round. Indeed, a clear relationship emerged between relevant political events and peaks of bots' activity (the riots in Catalonia or the exhumation of the Spanish fascist dictator Francisco Franco). Besides the timing of the interactions, the content itself provides a key consequence: it appears that social bots focus on attacking opposite ideologies, mainly aiming to introduce hate and disagreement in human conversations rather than directly supporting a candidate or a political party.

Further research is needed to prove these armies' coordinated nature and tie these contents to real individuals or political parties. Indeed, measuring the influence that the social bots have on public opinion represents a yet unsolved challenge, with potentially drastic ramifications.

Acknowledgments

This study was partially funded by the Spanish Government grants FPU18/00304, and RYC-2015-18210 and by a predoctoral grant from the University of Murcia.

References

[1] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "The spread of low-credibility content by social bots," *Nature Communications*, vol. 9, no. 1, p. 4787, 2018.

[2] S. Cresci, "A decade of social bot detection," *Commun. ACM*, vol. 63, no. 10, p. 72–83, Sep. 2020.

[3] J. Pastor-Galindo, M. Zago, P. Nespoli, S. López Bernal, A. Huertas Celdrán, M. Gil Pérez, J. A. Ruipérez-Valiente, G. Martínez Pérez, and F. Gómez Mármol, "Spotting political social bots in Twitter: A use case of the 2019 Spanish general election," *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2156–2170, 2020.

[4] J. Pastor-Galindo, M. Zago, P. Nespoli, S. López Bernal, A. Huertas Celdrán, M. Gil Pérez, J. A. Ruipérez-Valiente, G. Martínez Pérez, and F. Gómez Mármol, "Twitter social bots: The 2019 Spanish general election data," *Data in Brief*, vol. 32, art. no. 106047, 2020.

Sesión de Investigación A3: Ciberataques e inteligencia de amenazas
http://doi.org/10.18239/jornadas_2021.34.21

A review of Leveraging Cyber Threat Intelligence for a Dynamic Risk Framework

Raúl Riesco Granadino
Universidad Politécnica de Madrid
raul.riesco.granadino@alumnos.upm.es

Victor A. Villagrá
Universidad Politécnica de Madrid
victor.villagra@upm.esr

***Resumen*—One of the most important goals in an organization is to have risks under an acceptance level along the time. All organizations are exposed to real time security threats that could have an impact on their risk exposure levels harming the entire organization, their customers and their reputation. New emerging techniques, tactics and procedures (TTP) which remain undetected, the complexity and decentralization of organization assets, the great number of vulnerabilities proportional to the number of new type of devices (IoT) or still the high number of false positives, are only some examples of real risks for any organization. Risk management frameworks (RM) are not integrated and automated with Near Real Time (NRT) risk-related Cybersecurity Threat Intelligence (CTI) information. The contribution of this paper is an integrated architecture based on the Web Ontology Language (OWL), a semantic reasoner and the use of semantic web rule language (SWRL) to approach a Dynamic Risk Assessment and Management (DRA / DRM) framework at all levels (operational, tactic and strategic). We created a new semantic version of STIX™v2.0 for Cyber Threat Intelligence as it is becoming a de facto standard for Structured Threat Information Exchange. Our proposal uses an unprecedented mix of standards to cover all levels of a DRM and ensure easier adoption by users.**

***Index Terms*—STIX™, SWRL, OWL, Cybersecurity, Dynamic Risk Management (DRM), Cyber Threat Intelligence (CTI).**

Tipo de contribución: *Investigación ya publicada [1]*

I. Introduction

Current frameworks and methodologies for Risk Assessment (RA) processes [2] [3] follow an iterative approach in which a partial snapshot of the organization assets and business processes is periodically taken for the estimation of its risk exposure and it is primarily based on expert and subjective theories.

On the other hand, cyber threat landscape as well as the attack surface of any organization (or event unintentional incidents) change constantly. This real and dynamic behavior render these legacy frameworks and methodologies highly ineffective and unreliable for any organization or risk analyst.

Threat intelligence specification drafts like STIX™, or TAXII™are becoming de facto standards but they have still important limitations to describe more complex concepts like TTP (Tactics, Techniques and Procedures) [4], Campaigns [5] or Incidents [6], between others neither in recent version like v2.1.

In this work, we propose a mix of 3 standards to overcome all described limitations: STIX™[7] as an Industry driven standard, as well as OWL [8] and SWRL [9] to overcome all semantic expressiveness and limitations of STIX enabling the understanding by machines and also the inference of new knowledge by reasoners. We consider that if our proposal is based on standards, it will be easier to implement and deploy by several organizations in the future.

II. Contributions

(i) Layered architecture for Dynamic Risk Assessment and Management (DRA/DRM) based on STIX™, OWL ontologies, SWRL and a Pellet semantic reasoner. (ii) Evolution and integration of Cyber Threat Intelligence data within DRA/DRM processes. (iii) Definition of several SWRL rules as algorithms and axioms to support all the business logic made by the Pellet Incremental semantic reasoner used in this work.

III. Approach and Results

In our case, a top manager (CFO) (with access to classified data) reads a third party newspaper daily, the same behavior of a Cybersecurity Expert from the same organization. Web surfing to these external systems is not part of the organization RA/RM scope. Then, despite all efforts at the perimeter and risk management countermeasures, a Watering hole attack could affect not only the victim data but also any data accesible by the victim like the classified data. The DRM will show a very different approach for each user dynamically.

Our approach is based on OWL Ontologies [8] as seen for example in figure 1 and Semantic Web Rule Language (SWRL) [9]. It provides the needed expressiveness of such concepts, rules but also to inference [10] new knowledge. They resolve the lack of interoperability between existing RM frameworks under a common language and understanding in order to expand risk context into a more realistic picture. Today, high speed SIEM to correlate isolated pieces of data by simpler algorithms (even processing several times the same

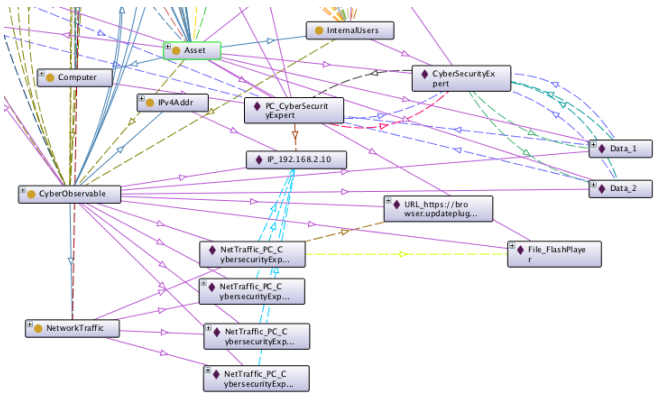


Figura 1. Integrated CTI and DRM OWL architecture

data) is used, here we can just share and add new data to a semantic OWL graph (needing less speed) and it will be automatically classified enriching CTI data and at the same time that new info can help the reasoner to infer new (never received) data by itself automatically. Furthermore our SWRL sharing algorithm perspective avoid sharing useless continuously changing IoC as the TTP does not change like ephemeral IoC.

We can improve the static selected organization's equation (reviewed once per year) to leverage Risk Assessment into something dynamic (depending on time t and security events) and affordable with enough flexibility and granularity like:

$$R_{it} = P_{it} + I_{it} - \Sigma(C1_{it}, C2_{it}, \ldots, CN_{it}) \quad (1)$$
$$R_t = \Sigma R_{it} \quad (2)$$

being:

$$P_{it} = P_{it-1} + \Sigma(EP1_{it}, EP2_{it}, \ldots, EPN_{it}) \quad (3)$$
$$I_{it} = I_{it-1} + \Sigma(EI1_{it}, EI2_{it}, \ldots, EIN_{it}) \quad (4)$$

where:

- P_{it} is the probability of threat i at time t,
- I_{it} is the impact of threat i at time t,
- CN_{it} is the decreasing value of the Impact (Severity) I or Probability P of threat i at time t due to Countermeasure N,
- EPN_{it} is the increasing value of the Probability P of threat i at time t due to the Security Event N,
- EIN_{it} is the increasing value of the Impact (Severity) I of threat i at time t due to the Security Event N,
- R_{it} is the Residual Risk associated to threat i, at time t,
- ΣR_{it} is the sum of all type of Residual Risks associated to threats $i = 1, \ldots, Z$, at time t,
- R_t is the Total Residual Risk at time t of all type of threats $i = 1, \ldots, Z$,

By using our framework (as shown in figures 1 and 2) we have all the needed expressiveness to better know what is really happening along the time, in this case, we know that there is a risk automatically identified of type deliberated malicious SW distribution which has been mitigated by two safeguards and, at the same time, it was increased by different security events. We perfectly know the connection of this risk to the affected assets and services, and all information is consistent. We can query our model to know more about all the relationships and reasoner conclusions but we can also use interactive graphs to see all the relationships as shown in figure 1.

IV. Conclusions

We developed a formal model based on standards to connect real time threats to risk calculation and risk management processes which also provide better automation, enrichment, detection capabilities and simplicity by using standards STIX™[7], OWL [8], a reasoner [11] and SWRL [9].

The selected watering hole attack was motivated due to the interest of the threat actor to access this type of data, which only the CFO (Chief Financial Ofcier) has access to

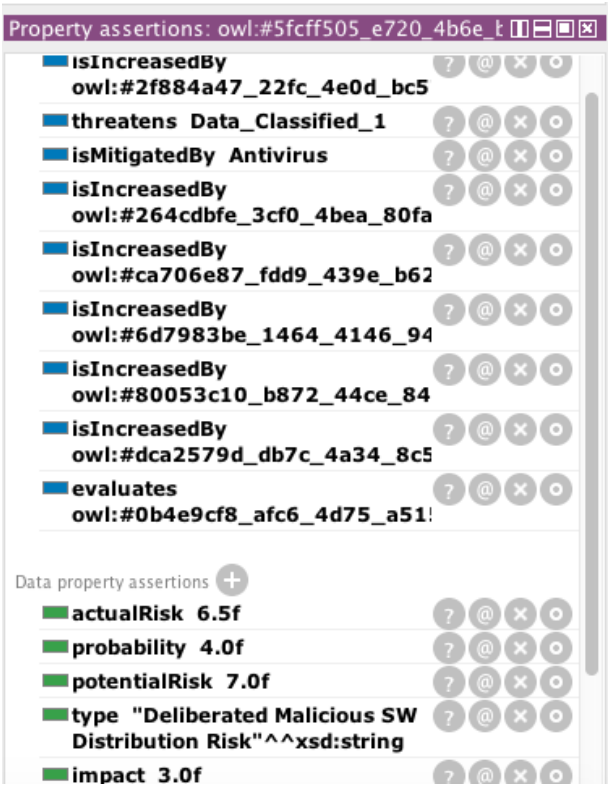


Figura 2. Instance of class Risk Assessment

it. Then, the CFO (Victim_0) is classified automatically by our framework as a potential victim of this type of attack. Different type of risks related to any unauthorized access to classified data will be created by the framework automatically due to the nature of the data (e.g., Bad Reputation Risk when classified data is accessed and leaked, Data Protection Risk, Corporate Bad image, etc.). Once the framework detects that an user with low cybersecurity experience has access to this classified data, it will infer automatic relationships between both type of risks (risk of unauthorized access to classified data and risk of deliberated malicious SW distribution to the user who has access to that data). Risks are also modified along the time based on user behavior and malicious traffic.

Referencias

[1] Riesco at al. *"Leveraging threat intelligence for a dynamic risk frameworkr"*, International Journal of Information Security (2019) https://link.springer.com/article/10.1007/s10207-019-00433-2

[2] *ISO/IEC 27005:2008, Information technology - Security techniques and Information security risk management*, (2008).

[3] *ISO 31000:2018, Risk Management - Guidelines*, (2018).

[4] OASIS, *"TTP (Techniques, Tactics and Procedures"by STIX™*, https://stixproject.github.io/getting-started/whitepaper/#tactics-techniques-and-procedures-ttp

[5] OASIS, *Çampaigns by STIX™"*, https://stixproject.github.io/getting-started/whitepaper/#campaigns

[6] OASIS, *Ïncidents by STIX™"*, https://stixproject.github.io/getting-started/whitepaper/#incidents

[7] OASIS, *"STIX™White paper"*, https://stixproject.github.io/about/STIX_Whitepaper_v1.1.pdf

[8] W3C, *.ºWL"*, https://www.w3.org/OWL/

[9] W3C, *"SWRL Semantic Web Rule Language"*, https://www.w3.org/Submission/SWRL/

[10] W3C, *Ïnference"*, https://www.w3.org/standards/semanticweb/inference

[11] W3C, *Reasoner"*, https://www.w3.org/2001/sw/wiki/Category:Reasoner

Sesión de Investigación A4:
Análisis forense y cibercrimen

Sesión de Investigación A4: Análisis forense y cibercrimen
http://doi.org/10.18239/jornadas_2021.34.22

A Review of "On Challenges in Verifying Trusted Executable Files in Memory Forensics"

Daniel Uroz, Ricardo J. Rodríguez
Dpto. de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Spain
duroz@unizar.es, rjrodriguez@unizar.es

***Abstract*—Memory forensics is a fundamental step in any security incident response process, especially in computer systems where malware may be present. The memory of the system is acquired and then analyzed, looking for facts about the security incident. To remain stealthy and undetected in computer systems, malware are abusing the code signing technology, which helps to establish trust in computer software. Intuitively, a memory forensic analyst can think of code signing as a preliminary step to prioritize the list of processes to analyze. However, a memory dump does not contain an exact copy of an executable file (the file as stored in disk) and thus code signing may be useless in this context. In this paper, we investigate the limitations that memory forensics imposes to the digital signature verification process of Windows PE signed files obtained from a memory dump. These limitations are data incompleteness, data changes caused by relocation, catalog-signed files, and executable file and process inconsistencies. We also discuss solutions to these limitations. Moreover, we have developed a Volatility plugin named `sigcheck` that recovers executable files from a memory dump and computes its digital signature (if feasible). We tested it on Windows 7 x86 and x64 memory dumps. Our experiments showed that the success rate is low, especially when the memory is acquired from a system that has been running for a long time.**

***Index Terms*—memory forensics, Authenticode, digital signature verification, code signing, Volatility**

Tipo de contribución: *Investigación ya publicada en* "On Challenges in Verifying Trusted Executable Files in Memory Forensics," Forensic Science International: Digital Investigation, vol. 32, p. 300917, Apr. 2020. [1]

I. Extended Abstract

A common kind of security incident is caused by the presence of malicious software (*malware*) in a system. Computer and network forensics become fundamental steps during the detection and analysis stage of a security incident response process. Anomalous or unauthorized activity performed by malware in a compromised system can be detected through the analysis of both the device drive and the memory of the system. Disk forensics is related to the analysis of device drives, while memory forensics focuses on the analysis of the data contained in the memory of the system under study [2]. There are situations, however, where the access to the physical device drives is difficult to accomplish (for instance, in Cloud computing). In this paper, we focus on memory forensics.

Memory forensics is usually carried out capturing the current state of the system's memory and dumping it into disk as a snapshot file. This file is also known as *memory dump*. A memory dump can then be taken off-site to analyze it with dedicated software such as Volatility [3], searching for facts about the security incident.

A memory dump contains tons of data that might be of interest for analysis. Among other things, it contains a snapshot of the processes in execution, as well as other system information such as logged users, open files, or open network connections at the time of memory acquisition. Note that the memory state can be inconsistent if it was acquired in a live system, since the system itself evolves over time and system objects might be created or destroyed during the acquisition process. To assess the reliability of analysis results, temporal dimension in memory forensics was recently proposed [4].

Code signing helps to establish trust in computer software, since it allows to authenticate the software publisher and to guarantee code integrity through the validation of the digital signature shipped within the software [5]. Several operating systems rely on code signing to warn the users about the potentially harmful actions that a piece of software may perform. For instance, the execution of a properly signed application in Windows avoids any alert box informing the user about the possible harmful consequences of its execution. Under this premise, malware developers use digital signatures to deceive users to execute their malware and thus compromise their systems, thus subverting the trust in digitally signed software.

Although the use of digital signatures in malware is not a growing trend [6], [7], there are documented cases of signed malware samples in the wild (such as Stuxnet [8], Duqu, or Flame, to name a few). Malware developers use trusted certificates that were either compromised or issued directly to them to sign their software. As the primary defense against these threats, we rely mainly on the revocation process of the abused certificates done by the certification authorities (CAs).

Intuitively, a forensic analyst can think of code signing as a preliminary step to prioritize the list of suspicious processes that need further analysis. The rationale for this thought is correct, but unfortunately is not very fruitful when inspecting a memory dump: a process is an inaccurate representation of the executable files in memory, since parts of the binary code may be paged out of memory or may change at acquisition time [9]. Furthermore, software defenses such as address space layout randomization [10], [11] or position-independent code may change the memory references of certain binary code instructions. But, to what extent can these issues negatively affect the signature computation? Are there any other issues affecting it? Are there any ways to overcome these problems? These questions have motivated our research. Our main research goal in this paper is to explore whether code signing brings any benefit to memory forensics.

Contributions. In this paper, we describe the limitations that memory forensics imposes to the digital signature verification process of Windows PE signed files. In particular, Authenticode is the code signing standard designed to digitally sign files in Windows, introduced in Windows 2000 [12]. We

focused on Windows since it is still the preferred target of malware authors [13]. We have also developed a Volatility plugin to verify digital signatures in a memory dump, named `sigcheck` (as the tool provided by Microsoft for verifying digital signatures on binary files [14]). When feasible, our plugin works on kernel-space file objects that represent executable files, computing the signature and verifying the certificate chain attached to the digital signature. To assess its reliability, we tested it in different scenarios and in signed malware samples. We concluded that the longer a system runs, the fewer file objects can be acquired. Hence, given the current limitations (data incompleteness, data changes caused by relocation, catalog-signed files, and executable file and process inconsistencies), verification of digitally-signed files does not bring any benefit to memory forensics.

Our plugin `sigcheck` relies on a set of plugins shipped with Volatility. In particular, it uses `tasks` (to retrieve the list of processes in execution), `modules` (to retrieve the list of drivers), `devicetree` (to retrieve the driver objects for a given module), `file-scan` (to retrieve the list of file objects), and `dumpfiles` (to obtain the list of memory addresses associated to a `FileObject` structure. The content of these addresses is later read in a programmatic way). Furthermore, the verification of an Authenticode-signed file has been implemented as an independent Python function (named `sigvalidator`), and thus it can be used with image files as well as with executable files.

To foster research in this area and enable the reproducibility of experiments, both the plugin `sigcheck` and the auxiliary tool `sigvalidator` have been released under GNU/GPL version 3 license and are freely available at https://github.com/reversea-me/sigcheck.

Regarding our experiments, the results show that there are more chances of retrieving file objects with complete data at fresh boot. As expected by the way of working of Windows' memory subsystem (page smearing, demand paging, and swap pages), the number of file objects with full content quickly drops as the time evolves. Furthermore, the results in a 32-bit OS are better than in 64-bit OS, for all scenarios. In Windows 7 x86, almost half of driver files are successfully verified as catalog-signed files, while executable and DLL files reach more than 30%. It is remarkable that the content of a huge percentage of file objects is partial, seeming not signed or with an incorrect image base address. In 64-bit DLL files, this percentage increases remarkably. Finally, let us also remark that none of the file objects retrieved in both scenarios contained the Authenticode signature as full content. Only 13 32-bit DLL files contained the certificate header, but it was incomplete due to memory paging issues.

We have also selected a number of signed malware samples [15] from public repositories and analyzed them with `sigcheck` through the auxiliary tool `sigvalidator`. Our results indicate that both Windows UAC and SysInternal's `sigcheck` are more focused on the publisher trustworthiness rather than other aspects of the certificate, such as the certificate validity time. We believe that apart from the publisher trustfulness, a certificate expired should always be reported, as our plugin does. Moreover, the messages shown by the Windows UAC are less intuitive for the users. The case of self-signed certificate is detected by the three tools, although text messages differ slightly. The most interesting case is a malware threat associated to the ransomware MegaCortex [16]. Although SysInternal's `sigcheck` warns that the certificate was revoked by the issuer, surprisingly the Windows UAC tells the user that the file comes from a verified publisher. This difference may be caused by the parameter settings when calling to `WinVerifyTrust`. Similarly, our plugin also returns that the verification process was successful. At the moment `sigcheck` does not perform any certificate revoking checking, since it is not supported by the OpenSSL binary package on which our plugin relies. We are currently implementing the certificate revocation process using the OpenSSL library to fix this issue.

The full version of this paper (with a full description of the experiments and limitations) was published in [1].

Acknowledgments. This work was supported in part by MICINN under grant MEDRESE-RTI2018-098543-B-I00 and by the Aragonese Government (DisCo research group, ref. T21-17R), and by the University of Zaragoza and the *Fundación Ibercaja* under grant JIUZ-2020-TIC-08. The research of D. Uroz was also supported by the Government of Aragón under a DGA predoctoral grant (period 2019-2023).

References

[1] D. Uroz and R. J. Rodríguez, "On Challenges in Verifying Trusted Executable Files in Memory Forensics," *Forensic Science International: Digital Investigation*, vol. 32, p. 300917, Apr. 2020. [Online]. Available: http://webdiis.unizar.es/~ricardo/files/papers/UR-FSIDI-20.pdf

[2] M. H. Ligh, A. Case, J. Levy, and A. Walter, *The Art of Memory Forensics: Detecting Malware and Threats in Windows, Linux, and Mac Memory*. John Wiley & Sons, Inc., Jul. 2014.

[3] A. Walters and N. Petroni, "Volatools: Integrating Volatile Memory Forensics into the Digital Investigation Process," in *BlackHat DC*, 2007.

[4] F. Pagani, O. Fedorov, and D. Balzarotti, "Introducing the Temporal Dimension to Memory Forensics," *ACM Trans. Priv. Secur.*, vol. 22, no. 2, pp. 9:1–9:21, Mar. 2019.

[5] B. Parno, J. M. McCune, and A. Perrig, "Bootstrapping Trust in Commodity Computers," in *2010 IEEE Symposium on Security and Privacy*, May 2010, pp. 414–429.

[6] X. Ugarte-Pedrero, M. Graziano, and D. Balzarotti, "A Close Look at a Daily Dataset of Malware Samples," *ACM Trans. Priv. Secur.*, vol. 22, no. 1, pp. 6:1–6:30, Jan. 2019.

[7] R. Rivera, P. Kotzias, A. Sudhodanan, and J. Caballero, "Costly freeware: a systematic analysis of abuse in download portals," *IET Information Security*, vol. 13, no. 1, pp. 27–35, 2019.

[8] R. Langner, "Stuxnet: Dissecting a Cyberwarfare Weapon," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49–51, May 2011.

[9] A. Case and G. G. Richard, "Memory forensics: The path forward," *Digital Investigation*, vol. 20, pp. 23–33, 2017.

[10] E. Bhatkar, D. C. Duvarney, and R. Sekar, "Address Obfuscation: An Efficient Approach to Combat a Broad Range of Memory Error Exploits," in *Proceedings of the 12th USENIX Security Symposium*, 2003, pp. 105–120.

[11] PaX Team, "PaX address space layout randomization (ASLR)," https://pax.grsecurity.net/docs/aslr.txt.

[12] Microsoft Corporation, "Windows Authenticode Portable Executable Signature Format," [online; http://download.microsoft.com/download/9/c/5/9c5b2167-8017-4bae-9fde-d599bac8184a/authenticode_pe.docx], Mar. 2008, accessed on September 25, 2019.

[13] AV-TEST, "Malware statistics," [Online; https://www.av-test.org/en/statistics/malware/], 2019.

[14] M. Russinovich, "Sigcheck v2.73," [online; https://docs.microsoft.com/en-us/sysinternals/downloads/sigcheck], Sep. 2019, accessed on September 25, 2019.

[15] J. Niemelä, "It's Signed, Therefore It's Clean, Right?" in *CARO 2010 Technical Workshop*, 2010.

[16] A. Brandt, ""MegaCortex" ransomware wants to be The One," [online; https://news.sophos.com/en-us/2019/05/03/megacortex-ransomware-wants-to-be-the-one/], May 2019, accessed on September 30, 2019.

Sesión de Investigación A4: Análisis forense y cibercrimen
http://doi.org/10.18239/jornadas_2021.34.23

A Review of "Camera Attribution Forensic Analyzer in the Encrypted Domain"

A. Pedrouzo-Ulloa
atlanTTic, UVigo, Spain
apedrouzo@gts.uvigo.es

M. Masciopinto
atlanTTic, UVigo, Spain
mmasciopinto@gts.uvigo.es

J. R. Troncoso-Pastoriza
EPFL, Switzerland
juan.troncoso-pastoriza@epfl.ch

F. Pérez-González
atlanTTic, UVigo, Spain
fperez@gts.uvigo.es

***Abstract*—This paper is a review of a work previously published by the authors at IEEE WIFS'18 (Workshop on Information Forensics and Security), which received the Best Paper Award, and contains a summary of its main results. In WIFS'18 we proposed a new framework for the secure outsourcing of the image source attribution problem, in which the Photoresponse Non-Uniformity (PRNU) is used as a fingerprint to decide whether a test image was taken with a specific camera device. This method is fully unattended, that is, the secret key owner does not take part during the process. To this aim, we introduced improvements on the state-of-the-art in secure and unattended solutions for denoising. We also showed how to homomorphically perform filtering, polynomial, denoising and pixel-wise operations in a single round without the need of an interactive protocol.**

***Index Terms*—Photoresponse Non-Uniformity; lattice-based cryptosystems; digital media forensics; camera attribution forensic analyzer**

Type of contribution: *Already published research*

I. Introduction

In this paper we present the results of our research that was previously published at the Workshop on Information Forensics and Security (WIFS) in 2018 [1].

A. Motivation

All digital imaging sensors intrinsically present a noise pattern called PRNU, which is due to tiny and random imperfections on the silicon wafer. PRNU is becoming particularly relevant within digital media forensics, as it can be used as a fingerprint to determine whether a given image was taken by a certain device. Consequently, many works have made use of its uniqueness feature for a wide range of applications; which includes identification and clustering of acquisition devices.

However, an important problem that these applications share is that they are computationally intensive and work with very large databases. Actually, although buying computing power and database storage as needed appears as an interesting solution, the privacy-sensitive nature of forensic data prevents from directly outsourcing it unencrypted.

Recent results from [2], [3] show that the estimated PRNU fingerprints leak a considerable amount of information of the images used for extraction. This constitutes a serious privacy threat and suggests that for some scenarios (e.g., child pornography crimes), camera fingerprints should be protected not only when outsourcing, but at all times during investigations.

B. Main results of [1]

The secure scheme proposed in [1] was exemplified for the case of PRNU extraction/detection, but it covers many other forensic tools.

The main technical results are the following:

- An efficient Wavelet-based denoising primitive is introduced. The main novelty relies on the use of a new homomorphic threshold function by means of the "lowest digit removal" polynomials introduced in [4], [5].
- Further optimizations on the Wavelet denoising primitive are presented, consisting of the use of efficient NTT (Number Theoretic Transforms) packing.
- The previous encrypted denoising primitive is used as a building block in a more complex use case as the PRNU extraction/detection for camera attribution. The proposed method is able to compute the process for extraction/detection in an unattended way, that is, without additional interactions between the client and server.

II. Proposed Scheme

A. Related Works

To the best of our knowledge [6], there are two different approaches for secure camera attribution: (a) Mohanty *et al.,* [7], [8] who combine a trusted environment (ARM TrustZone) for the computation of the PRNU fingerprint, with the Boneh-Goh-Nissim (BGN) cryptosystem for the matching, and (b) ours [1], which proposes a more flexible solution that can be implemented on a general purpose architecture and does not require access to a trusted environment.

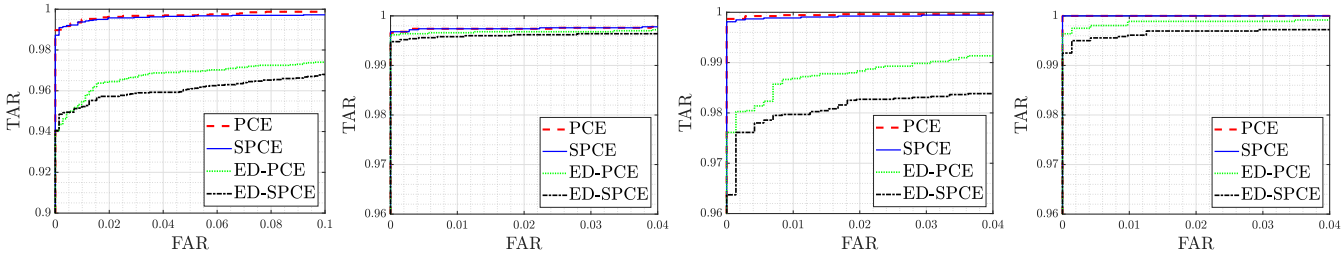
As we discussed in [6], although Mohanty *et al.*'s scheme evaluates most of the computation in the clear, their runtimes do not improve those obtained by our solution. In fact, the PRNU matching in their scheme could be more efficiently calculated by substituting the BGN cryptosystem with more modern lattice-based cryptosystems. In relation to this, it is worth mentioning that, if available, our solution could also use a trusted environment to improve the efficiency.

B. Unattended and Secure Camera Attribution

Our proposed scheme is based on the use of an RLWE (Ring Learning with Errors) cryptosystem equipped with an adequate use of NTT transforms and efficient signal pre-/post-coding operations before/after encryption/decryption.

Due to space restrictions, we refer the reader to [1] for more details. A full diagram of the proposed framework is included in [1, Fig. 1].

The main challenge is the efficient evaluation of the threshold function used in the Wavelet denoising primitive. By approximating this threshold with a quantization operation, we can leverage the "lowest digit removal" polynomials as a mechanism to homomorphically evaluate thresholding. The



(a) Nikon D7000. (b) Nikon D90. (c) Canon 1100D. (d) Nikon D3000.

Fig. 1: True Acceptance Rate (TAR) vs. False Alarm Rate (FAR) for 4 different camera devices. PCE represents the result obtained with the denoising in [9] and the PCE statistic [10], SPCE is the simplified detector in [1, eq. (4)] applying the denoising from [9], ED-PCE is the PCE statistic using the encrypted denoising described in [1, Sec. 3.2], and ED-SPCE stands for the simplified detector discussed in [1, Sec. 3].

use of this functionality results in a considerably reduction of the ciphertext size and the depth of circuit to be computed.

III. Performance Evaluation

We evaluated in [1] our secure framework in terms of efficiency, security and performance. To this aim, we securely performed the PRNU detection test, in which the PRNU estimate is tested against the test image via the statistical distribution of a score on both hypothesis (i.e., the image contains or not the PRNU estimate); whereas the PRNU estimate was obtained in the clear domain.

This scenario corresponds to the case where the police have confiscated the camera of a suspect, and would like to check whether an image has been taken by this camera.

Due to legal restrictions, this test image cannot be outsourced without being previously protected. On the contrary, as we have control of the camera, we can take flatfield images to perform the extraction without any privacy leakage.

A. Implementation and execution times

We implemented our scheme taking advantage of the RNS variant of the FV cryptosystem [1], and execution times were measured on an Intel Xeon E5-2667V3 at 3.2GHz using one core for the non-parallelized choice.

Table I reports the runtimes for encrypted detection assuming that the PRNU estimate and the test image are aligned.

TABLE I: Runtimes for Encrypted PRNU detection (2048×2048 image)

Parallelization (cores)	1	8	16	20
Encrypted Detection (min)	128.33	16.05	8.03	6.53
Encryption + Pre-coding (s)	3.6 (1 core, client-side)			
Decryption + Post-coding (ms)	27 (1 core, client-side)			

The introduced improvements on the unattended denoising primitive result to be fundamental in achieving the above execution runtimes.

B. PRNU Detection Performance

We utilized a database composed of 2639 TIFF images taken from 16 digital camera devices. The fingerprint was extracted for each different camera device from 50 randomly chosen TIFF images. For the detection phase, we considered crops of the JPEG-compressed version of the TIFF images with size 1536×1536 and a quality factor of 95.

Figure 1 compares the performance of the detector in [1, Eq. (2)] (dot product) with the Peak to Correlation Energy (PCE) detector [10], both when the popularly used image denoising in [9] and when our encrypted denoising are used to obtain the residues of the different test images. As the fingerprint estimate is obtained in the clear, we used in all the experiments the denoising method from [9] for extraction.

IV. Conclusions and Future Work

This work reviews the results obtained in a previously published paper [1] by the authors. In [1], we introduced an unattended secure framework for outsourcing computation which could perform the PRNU extraction/detection phases without any additional interaction with the client. We evaluated the performance of our method in a concrete scenario on which the test images have to be protected.

Our results show the feasibility of source camera attribution in the encrypted domain. Even so, there is still room for improvement, and we are currently working on a complete evaluation of the encrypted extraction. This includes further refinements on the encrypted denoising primitive, and a reevaluation of the use of the underlying RLWE cryptosystem profiting from the most recent results in the field.

Acknowledgments

GPSC is funded by the Agencia Estatal de Investigación (Spain) and the European Regional Development Fund (ERDF) under project RODIN (PID2019-105717RB-C21). Also funded by the Xunta de Galicia and the European Union (European Regional Development Fund - ERDF) under projects ED431G2019/08 and Grupo de Referencia ED431C2017/53.

References

[1] A. Pedrouzo-Ulloa, M. Masciopinto, J. R. Troncoso-Pastoriza, and F. Pérez-González, "Camera Attribution Forensic Analyzer in the Encrypted Domain," in *IEEE WIFS*, 2018, pp. 1–7.

[2] S. Fernández-Menduiña and F. Pérez-González, "On the Information Leakage of Camera Fingerprint Estimates," 2020.

[3] F. Pérez-González and S. Fernández-Menduiña, "Prnu-leaks: facts and remedies," in *EUSIPCO 2020*. IEEE, 2020, pp. 720–724.

[4] M. Griffin, "Lowest degree of polynomial that removes the first digit of an integer in base p," https://mathoverflow.net/q/269282, accessed: 10 March 2020.

[5] H. Chen and K. Han, "Homomorphic Lower Digits Removal and Improved FHE Bootstrapping," in *EUROCRYPT*, 2018, pp. 315–337.

[6] A. Pedrouzo-Ulloa, M. Masciopinto, J. R. Troncoso-Pastoriza, and F. Pérez-González, "Efficient PRNU Matching in the Encrypted Domain," in *XoveTIC*. MDPI, 2019.

[7] M. Mohanty, M. Zhang, M. R. Asghar, and G. Russello, "PANDORA: Preserving Privacy in PRNU-Based Source Camera Attribution," in *IEEE TrustCom/BigDataSE*, 2018, pp. 1202–1207.

[8] M. Mohanty, M. Zhang, M. R. Asghar, and G. Russello, "e-PRNU: Encrypted Domain PRNU-Based Camera Attribution for Preserving Privacy," *IEEE Trans. Dependable and Sec. Computing*, pp. 1–1, 2019.

[9] M. K. Mihcak, I. Kozintsev, and K. Ramchandran, "Spatially adaptive statistical modeling of wavelet image coefficients and its application to denoising," in *IEEE ICASSP*, vol. 6, 1999, pp. 3253–3256.

[10] M. Goljan, J. Fridrich, and T. Filler, "Large scale test of sensor fingerprint camera identification," in *Proc. SPIE, Electronic Imaging, Media Forensics and Security XI*, vol. 7254, Feb. 2009, pp. 0I 1-0I 12.

http://doi.org/10.18239/jornadas_2021.34.24

Integrando el Edge Computing en el Desarrollo de una Metodología Forense Dedicada a Entornos IoT

Juan Manuel Castelo Gómez, José Roldán-Gómez, Javier Carrillo-Mondéjar, José Luis Martínez Martínez
Universidad de Castilla-La Mancha.
Instituto de Investigación en Informática de Albacete
C/ Investigación 2 Albacete 02071
juanmanuel.castelo@uclm.es, jose.roldan@uclm.es, javier.carrillo@uclm.es, joseluis.martinez@uclm.es
ORCIDs: J.M. Castelo (0000-0001-6117-482X), J. Roldán (0000-0001-5787-1294), J. Carrillo (0000-0001-8371-4305), J.L. Martínez (0000-0001-5119-2418)

***Resumen*—La aplicación del Internet de las Cosas (IoT) en los múltiples ámbitos de nuestra sociedad no solo ha supuesto buenas noticias para los usuarios, que han conseguido llevar la tecnología a lugares que no hacían uso de ella, sino que, lamentablemente, los cibercriminales también se han visto beneficiados con este cambio de paradigma. La fragilidad de los dispositivos IoT en términos de seguridad, unido a la sensibilidad de los datos que manejan, ha causado que el IoT sea un lugar idóneo en el que llevar a cabo sus ataques. En consecuencia, se necesitan de técnicas forenses que permitan esclarecer cómo se debe proceder a la hora de examinar estos nuevos dispositivos, puesto que tienen características muy distintas a los convencionales. Por ello, en este artículo, además de realizar una evaluación del estado del análisis forense en el entorno IoT y sus requisitos, se propone una metodología forense centrada en este nuevo entorno que hace uso de la tecnología *edge computing*.**

***Index Terms*—Ciberseguridad, Internet de las Cosas, Análisis Forense, Edge Computing, Metodología Forense**

Tipo de contribución: *Investigación original (límite 8 páginas)*

I. Introducción

El auge del Internet de las Cosas (IoT) ha supuesto llevar la tecnología a lugares en los que su uso no era tan extendido o, directamente, no era posible debido a las características de los dispositivos convencionales. El desarrollo de nuevos dispositivos y sistemas ha dado paso a nuevos contextos como la Industria 4.0, las ciudades inteligentes, la eSalud o los hogares inteligentes, y, en consecuencia, al tratamiento de datos extremadamente sensibles y privados. La incorporación de estos nuevos dispositivos en la vida ordinaria de las personas se ha hecho de manera mucho más fluida de lo que era imaginable, hasta el punto de que el IoT se ha convertido en algo que está presente en casi cualquier aspecto de nuestra vida, que ahora es más digital que nunca.

Actualmente, los dispositivos IoT ya son mayores en número que los dispositivos no IoT [1] y, lo que es más importante, no encontramos este tipo de dispositivos en entornos tecnológicos muy específicos, sino que es el entorno doméstico el que mayor uso hace de ellos [2]. Esto significa que no están bajo el control y monitorización de profesionales informáticos, justo al contrario, son personas con rangos de conocimiento tecnológico muy diverso los que se aprovechan de las infinitas funciones que el IoT aporta. A simple vista, esto podría parecer algo insignificante, pero la pobre seguridad de los dispositivos IoT hace que este hecho se convierta en un aspecto clave. Durante el año 2019 se detectaron más de 100 millones de ataques en dispositivos IoT [3], dato que es mucho más grave cuando nos centramos en las estadísticas relativas al último cuarto del año 2020, en las que se puede observar que más del 85 % de los ataques en este tipo de dispositivos tenían como objetivo el servicio Telnet, conocido por ser poco seguro, y que las credenciales por defecto de los dispositivos eran extremadamente simples, con combinaciones como “admin-admin” o “root-1234”, que se mantenían activas y permitían que los ataques por diccionario fueran muy efectivos [4]. Esto, unido a lo comentado anteriormente, hace que el entorno IoT sea un lugar idílico para los cibercriminales en el que llevar a cabo sus ataques, ya que pueden obtener grandes beneficios empleando técnicas de menor complejidad que en los entornos convencionales.

La existencia de este gran número de ataques en dispositivos IoT tiene una consecuencia directa en análisis forense, y es que se necesita de técnicas que permitan realizar investigaciones en este entorno. Sin embargo, el desarrollo de estas técnicas no está resultando tan sencillo. El motivo más importante es que existe una gran cantidad de diferencias fundamentales entre el entorno IoT y los entornos convencionales, lo que se traduce en que las herramientas y metodologías usadas hasta ahora no son del todo efectivas cuando se aplican en este ámbito. Además, al tratarse de una tecnología tan novedosa, los investigadores forenses están un paso por detrás de los cibercriminales, y van planteando soluciones según lo que se va aprendiendo sobre el comportamiento que tienen los *crackers* a la hora de comprometer los sistemas. Por último, no hay que olvidar que el campo forense está estrechamente ligado al mundo legal, por lo que el desarrollo de nuevos procedimientos se debe hacer con cautela. Dos principales razones motivan esta afirmación. La primera es que las leyes limitan la libertad de movimiento de los investigadores a la hora de realizar grandes cambios en la forma de realizar los análisis forenses, ya que una solución que no cumpla con los requisitos legales actuales no podrá ser utilizada en un procedimiento judicial, y el desarrollo de nuevas leyes es una tarea que avanza a una velocidad lenta. A su vez, se debe tener en cuenta que los profesionales que trabajan en el ámbito legal necesitan tiempo para adaptarse a las nuevas tecnologías que van surgiendo, por lo que un cambio brusco en proceder significará un problema de entendimiento entre el mundo legal y el tecnológico, más aún cuando el primero todavía está acomodándose a los procedimientos forenses convencionales.

Por otra parte, los cambios de paradigma abren la puerta a

la inclusión de tecnologías novedosas que no se estaban utilizando hasta el momento. Una de las tecnologías ligadas al IoT es el *edge computing*, que permite solventar las limitaciones computacionales de los dispositivos IoT y aportar un grado adicional de rapidez a la hora de interpretar y analizar datos. Este aspecto es de extrema utilidad en el ámbito forense, en el que es indispensable estudiar un gran número de datos de cara a poder extraer las conclusiones que nos permitan determinar lo que ocurrió en un incidente.

Teniendo todos estos elementos en cuenta, el objetivo de esta investigación es combinar los procedimientos forenses convencionales con los requisitos del entorno IoT y, haciendo uso del *edge computing*, desarrollar una metodología que permita mejorar la forma de llevar a cabo investigaciones forenses en el Internet de las Cosas.

Contribuciones. Las principales contribuciones de esta investigación son las siguientes:

- Se hace un recorrido por las aportaciones de la comunidad científica relacionadas con el desarrollo de metodologías forenses centradas en el entorno IoT y en el uso de la tecnología *edge comptuting* dentro del ámbito de la informática forense.
- Se detallan qué características y requisitos particulares tiene el entorno IoT y los dispositivos que se encuentran dentro de él desde la perspectiva forense, justificando por qué es necesario el desarrollo de nuevas metodologías y modelos que permitan llevar a cabo con garantías las investigaciones.
- Teniendo estos requisitos en cuenta, se propone una metodología para el desarrollo de análisis forenses en entornos IoT que se apoya en el *edge computing* para la realización de tareas críticas dentro del proceso de investigación.

El resto del artículo está organizado de la siguiente forma. La Sección II describe los elementos que motivan la realización de esta investigación. Los trabajos desarrollados por la comunidad científica relativos a las metodologías forenses en IoT y el uso del *edge computing* en el campo del análisis forense se evalúan en la Sección III. La Sección IV detalla la propuesta de metodología forense centrada en entornos IoT que se apoya en la tecnología *edge computing* para realizar las investigaciones. Por último, en la Sección V se presentan las conclusiones extraídas tras la realización de esta investigación.

II. Motivación

Ampliando los dos principales elementos brevemente mencionados en la introducción, esta sección detalla las características de ambos que alimentan el desarrollo de esta investigación.

Necesidad de una metodología forense en IoT. A continuación, se enumeran las principales razones que justifican el diseño de una metodología forense que esté centrada en el Internet de las Cosas y por qué un modelo convencional no es capaz de satisfacer los requisitos de este nuevo entorno.

- Cantidad de dispositivos en la red: la dimensión de las investigaciones forenses cambia por completo debido a la gran cantidad de dispositivos que suelen estar presentes en las redes IoT. Pasamos de un escenario convencional en el que lo habitual es examinar un dispositivo, mientras que en el escenario IoT no es extraño encontrar redes compuestas por decenas de dispositivos, lo que amplia en gran medida el rango de la investigación a la vez que hace más importante la tarea de seleccionar correctamente la prioridad a la hora de adquirir y analizar la información de una fuente de evidencias. Por tanto, es necesario de una metodología que aborde el análisis de un sistema IoT desde un punto de vista global, en el que sea el entorno el que prime por encima de la individualidad de los dispositivos. Este aspecto también afecta a la fase de identificación, puesto que la delimitación del rango de la escena se convierte en una tarea mucho más compleja y que debe apoyarse en técnicas que aseguren la correcta evaluación de todos los dispositivos y las relaciones entre los mismos.
- Especificaciones técnicas de los dispositivos: las unidades IoT cuentan con una menor cantidad de memoria, almacenamiento y capacidad de procesamiento. Esto limita la cantidad de datos que pueden almacenar, por lo que elegir el momento adecuado en el que adquirir las fuentes de evidencia y hacer un filtrado correcto de datos de cara a encontrar la información relevante se convierten en tareas críticas, por lo que la fase de identificación presenta una importancia añadida.
- Intercambio de información: los dispositivos IoT están diseñados para estar intercambiando constantemente información, de ahí que sus capacidades de cómputo sean menores que las de los dispositivos convencionales. Esto se traduce en que los datos se intercambian de forma muy rápida y es más complejo el poder acceder a ellos, puesto que su tiempo de vida es mucho más corto. Esto dota de un aspecto preventivo a los análisis forenses, ya que, de cara a poder acceder a esa información, sería necesario que se almacenase antes de que el incidente se produzca, por lo que la metodología debe indicar cómo debe prepararse un entorno IoT para hacer más fácil la realización de análisis forenses en él.
- Uso de la nube: una de las consecuencias de la baja capacidad de cómputo de los dispositivos IoT es el uso de la nube para llevar a cabo las operaciones que estas unidades no son capaces de ejecutar. Es por eso que no es extraño el ver que una red IoT se apoya en el ella, o incluso que el propio núcleo de la misma está desplegado en la nube. Como es bien conocido, la realización de análisis forenses este entorno es una tarea compleja, puesto que el acceso a los datos está muy restringido por los proveedores de servicio y las leyes del país en el que estén operando. Una metodología forense para el entorno IoT debería abordar este aspecto u ofrecer una alternativa para poder tratar esa información.
- Acceso físico a los dispositivos: una posibilidad que no existía en el entorno convencional es la de no ser capaces de acceder físicamente al dispositivo que debemos investigar. En cambio, esto sí ocurre en el IoT, puesto que un dispositivo puede estar integrado dentro de una máquina o puede darse el caso de que, aunque el acceso físico sea posible, la tarea sea algo inviable, como puede ocurrir en casos en los que los diversos nodos IoT de una red estén separados por grandes distancias. Este hecho se traduce en que el acceso remoto a los dispositivos

puede tener una importancia mayor que la que se le daba en el forense convencional, en el que se trataba de no interactuar con los dispositivos para no comprometer su integridad.
- Heterogeneidad del entorno: por último, y no menos importante, nos encontramos con el hecho de que el Internet de las Cosas es un concepto muy amplio. Esto se traduce en la existencia de dispositivos, sistemas operativos y escenarios que difieren mucho entre sí, pero que todos ellos forman parte del entorno IoT. Por tanto, no es efectivo el abordar el diseño de metodologías desde una perspectiva global, como así hacen los modelos convencionales, sino que es necesario tener en cuenta los requisitos y particularidades de cada contexto para así poder satisfacerlos.

Inclusión del *edge computing* en el análisis forense en IoT. Hay tres motivos principales que hacen que el uso del *edge computing* sea una estrategia interesante a seguir en el desarrollo de soluciones forenses:

- Cantidad de datos intercambiados: como se ha mencionado anteriormente, el entorno IoT está pensado para intercambiar gran cantidad de datos. Desde el punto de vista forense esto se traduce en que la información es mucho más volátil que en un entorno convencional, y el acceder a los datos relevantes intercambiados *on-the-fly* es algo que no es posible una vez que el incidente ha ocurrido. En consecuencia, un dispositivo capaz de capturar y filtrar los datos intercambiados en una red IoT puede marcar la diferencia a la hora de realizar una investigación forense, puesto que permitiría al investigador acceder a datos a los que no podría haber accedido de otra manera.
- Dificultad de acceso a los datos: debido a las características de los sistemas IoT, el hecho del que el proceso de adquisición sea satisfactorio no está garantizado. Nos encontramos con dispositivos que tienen su almacenamiento soldado a la placa que le dota de funcionalidad, por lo que una simple extracción y clonado del sistema de almacenamiento no es posible, hay que optar por técnicas más complejas como el *chip-off*, *Joint Test Action Group (JTAG)/Universal Asynchronous Receiver/Transmitter (UART)* o *in system programming (ISP)*. Del mismo modo, el método de adquisición en vivo no siempre es posible de realizar, primero porque la compatibilidad de la herramienta de adquisición con el sistema que está ejecutando el dispositivo no está del todo asegurada y, además. porque no siempre trabajaremos con dispositivos que ejecuten algún tipo de sistema, como pueden ser sensores que funcionen solamente mediante *firmware*. Por tanto, la existencia de un sistema que sea capaz de acceder a los datos de la red IoT asegurará que sea posible extraer información de los dispositivos que se encuentran en ella.
- Capacidad de cómputo de los sistemas IoT: este aspecto anteriormente mencionado hace complicado el ejecutar tareas con cierta complejidad, como podría ser lanzar una herramienta forense que extraiga algún tipo de información a la hora de hacer un análisis en vivo. El contar con un nodo extra capaz de ejecutar tareas más complejas y de acceder a los datos que se están intercambiando en la red IoT permitirá al investigador forense la adquisición y análisis de datos de una forma más sencilla.

III. Estado del Arte

Hay dos aspectos claves que deben ser estudiados de cara a comprender cómo enfocar el desarrollo de la propuesta. Primeramente, es interesante conocer qué metodologías y modelos han sido propuestos para abordar la problemática de las investigaciones forenses en el Internet de las Cosas. Del mismo modo, de cara a determinar qué puede aportar el *edge computing* en una metodología forense IoT, es útil examinar qué usos previos se han hecho de esta tecnología dentro del campo de la informática forense.

III-A. Desarrollo de metodologías forenses para el entorno IoT

Existe un importante número de investigaciones que abordan el desarrollo de soluciones para modelar el análisis forense en el entorno IoT. Sin embargo, no todo son metodologías, sino que encontramos otras propuestas como *frameworks* o modelos. De hecho, es más habitual encontrar aportaciones de los dos últimos tipos. Por tanto, y puesto que en todas ellas aportan características interesantes, en este apartado evaluamos las más importantes, independientemente de si se tratan específicamente de metodologías o no.

La primera propuesta a destacar es [5], en la que se propone un *framework* genérico para investigaciones IoT que cumple con la norma ISO/IEC 27043:2015, y que está dividido en tres módulos: “proceso proactivo”, “forense en IoT” y “proceso reactivo”. De esta forma, cubre desde el proceso de preparación de un adecuado para desarrollar análisis forenses hasta la respuesta ante incidentes. Respecto al análisis forense como tal, describe qué infraestructuras pueden contener evidencias, clasificándolas en “forense en la nube”, “forense en red” y “forense a nivel de dispositivo”.

En [6] se presenta una metodología centrada en la privacidad de las investigaciones que también cumple con normativa internacional, en este caso la ISO/IEC 29100:2011. La propuesta está dividida en seis fases y depende de la instalación de un software que actúa como testigo forense y recopila y almacena los datos de una red IoT.

La primera propuesta que tiene en cuenta la existencia de diferentes contextos en el Internet de las Cosas es [7], la cuál adapta la investigación en función del escenario en el que se encuentra. En concreto, consta de tres componentes: “Forense según Aplicación”, “Forense Digital”, y “Proceso Forense”. Es en este primer componente en el que se ofrecen varias pautas de cómo debería abordarse el análisis dependiendo del contexto, mencionando los hogares inteligentes, las ciudades inteligentes y los *wearables*. Respecto al proceso forense práctico, éste no se detalla demasiado, aunque se hace una distinción entre “Forense IoT”, “Forense de Red” y “’Forense en la Nube’.

También existen aportaciones como [8], que se centran en una fase concreta del proceso de investigación. En este caso aborda la adquisición de evidencias, dividiéndola en identificación y captura. Para llevar a cabo la primera tarea, se realiza una división de la red en tres zonas: “Red de

Área Personal", "Red de Área Intermedia" y "Red de Área Externa". Respecto a la captura, se mencionan siete acciones a realizar, pero se hace desde un punto de vista muy teórico.

Centrándose en un contexto determinado, [9] aborda los vehículos inteligentes, proporcionando unas pautas breves de cómo se deben examinar los vehículos autónomos, también describiendo cómo preservar los datos adquiridos.

Abordando el mismo contexto, [10] presenta un *framework* muy detallado centrado en el Internet de Vehículos (IoV), aportando directrices para la adquisición de los datos y su almacenamiento seguro en una infraestructura distribuida, proponiendo un algoritmo que permite comprobar la integridad de los datos capturados.

Cambiando de ámbito, pero también centrado en un contexto particular encontramos [11], en el que se propone un *framework* para las investigaciones en hogares inteligentes. Se divide en siete fases y cubre desde la preparación previa al análisis hasta el propio estudio de los datos capturados. Además, aporta un grado interesante de flexibilidad, puesto que indica que no todas las fases son necesarias en la investigación. Sin embargo, peca de falta de enfoque práctico, puesto que se ofrecen directrices sobre dónde podrían encontrarse los datos, pero no se indica cómo proceder a la hora de realizar su captura.

La posibilidad de incorporar nuevas tecnologías a las investigaciones forenses queda reflejada en [12], en el que se aplica el *fog computing* en el desarrollo de un *framework*. Dicha tecnología se usa al contar con un nodo dentro de la red capaz de filtrar y analizar los datos generados por los dispositivos que hay en ella, ofreciendo también la posibilidad de notificarlos cuando se materializa una amenaza y almacenar sus datos en caso de que se hayan visto afectados por la misma.

Abordando la perspectiva holística que caracteriza al entorno IoT, [13] presenta un modelo dividido en tres fases que cubre el proceso proactivo, de inicialización de la investigación y de examinación. Esta última, la más práctica de todas, cubre desde la adquisición hasta el cierre de la investigación, aunque con falta de detalle práctico.

Una metodología centrada en un contexto determinado, pero delimitado por hardware en lugar de por aplicación es [14], que cubre de forma breve desde la preparación de la investigación hasta la presentación de resultados en investigaciones forenses en plataformas hardware de prototipado del Internet de las Cosas. Además, presentan una herramienta para la adquisición de datos volátiles y no volátiles en sistemas Raspbian.

[15] presenta una extensión del *framework* propuesto en [5], que pasa a estar formado por nueve componentes y a cumplir con la normativa ISO/IEC 27043. Aunque no aporta muchos detalles sobre la parte práctica de las investigaciones forenses, sí que menciona que dicho proceso está dividido en tres fases: "Inicialización", 'Adquisición'" e "Investigación".

Por último encontramos [16], que propone una metodología eminentemente práctica centrada en el contexto *wearable*. Aunque solo cubre la inicialización de la investigación y el procesado de los datos, presenta un aporte novedoso, y es que menciona el uso de métodos de adquisición como JTAG, el cuál no había sido contemplado en el resto de propuestas.

III-B. Uso del edge computing en el campo del análisis forense

Antes de comenzar a evaluar las investigaciones centradas en el ámbito forense, es interesante recalcar que existe una diferencia de concepto entre *edge computing* y *fog computing*, aunque en ciertas situaciones dicha diferencia sea mínima. Mientras que *fog computing* se asemeja más al uso de la nube en un entorno intermedio entre la propia nube y el sistema que hace uso de ella, el *edge computing* hace referencia a una situación más específica en la que se utilizan dispositivos en el último nivel de la red en la que se encuentra el sistema y permite añadir nuevas funcionalidades en la misma [17]. En nuestro caso, solo evaluaremos en detalle las que están centradas completamente en *edge computing*, que es la tecnología que trata esta investigación, pese a que no hay demasiadas. En cambio, sí que encontramos varias que abordan el tema de *fog computing* como [18], que se centra en el contexto de las redes vehiculares, o como [19] y [20], que abordan el entorno industrial.

La primera aproximación a un sistema destinado a hacer uso del *edge computing* para asistir en las investigaciones forenses en el Internet de las Cosas se hace en [21], el cual se denomina *Forensics Edge Management System (FEMS)*. El sistema presenta servicios como monitorización del tráfico red, minado de datos, *logging*, parseo y filtrado de datos o creación de líneas temporales, entre otros. Para ello, su estructura se divide en tres capas: la capa de percepción, que se encarga de la recolección de datos y de la detección de eventos, la capa de red, que transmite y procesa la información transmitida por la capa anterior, y la capa de aplicación, que proporciona la interfaz entre el usuario y el sistema. Su funcionamiento es el siguiente: el sistema monitoriza de manera constante la red, almacenando los datos de forma temporal en caso de que ocurra un incidente. Si el sistema detecta un evento relevante, se toma nota de la hora en la que se produjo y todo los datos almacenados a partir de ese momento se marcan como relevantes y el sistema entra en modo forense, en el que realiza el filtrado y parseo de datos, creación de la línea temporal, creación de alertas y presentación de resultados.

Una propuesta más reciente es [22], en la que se presenta un *framework* centrado en la realización de análisis forenses en la Industria 4.0. Dicho *framework* se divide en dos módulos, uno centrado en la detección de ataques y otro encargado en el almacenamiento seguro de los datos recopilados mediante la generación de claves. Su funcionamiento se basa en el uso de patrones de ataque que permiten la detección de los mismos. Una vez que se detecta un ataque, se procede a responder al incidente, extrayendo y preservando las evidencias, y, posteriormente, se genera un informe. Además, el *framework* se evalúa usando *Live Digital Forensic Framework for a Cloud (LDF2C)* y comparándose con otras propuestas similares centradas en la detección de ataques en redes.

Tras analizar las propuestas de la comunidad científica, podemos extraer las siguientes conclusiones:

- El desarrollo de metodologías forenses en IoT está en una etapa inicial en la que el número de propuestas no es muy amplio y muchas de ellas obvian el apartado práctico, por lo que es complicado evaluar su rendimiento.

- Encontramos investigaciones centradas en modelar contextos IoT específicos y, comparándolas con las generales, podemos ver que existen diferencias notables que justifican la delimitación en cuanto a escenarios.
- Respecto al apartado práctico, se puede observar que el análisis en vivo es una opción que gana enteros en este nuevo entorno. Además, también encontramos un denominador común que limita el desarrollo de las metodologías, y es la ausencia de herramientas forenses específicas para IoT.
- El uso del *edge computing* permite la posibilidad de realizar tareas de monitorización y detección de incidentes, ampliando el rango del análisis forense al tener una característica preventiva, y reduciendo el mínimo el tiempo de respuesta ante incidencias.

IV. Metodología Propuesta

Para detallar el funcionamiento de la metodología propuesta, primeramente se explica el modelo que se utiliza como base para la realización de la investigación y, posteriormente, se describe el resultado de incorporar la tecnología *edge computing* a dicho modelo.

IV-A. Metodología de partida

De cara a asegurar la viabilidad y efectividad de la propuesta, esta investigación tiene como base una metodología forense para entornos IoT que ha sido aprobada por la comunidad científica. Dicha propuesta es [23], que, aunque está centrada en un contexto determinado, explica qué elementos de la misma sirven desde el punto de vista general. A su vez, la perspectiva es eminentemente práctica, lo que aporta un nivel adicional de detalle a la hora de abordar las investigaciones. Los elementos más significativos de esta propuesta son:

- Tiene como base una metodología convencional, más concretamente la propuesta en [24].
- Se divide en las siguientes fases: "Preprocesado", "Identificación", "Adquisición", "Análisis", "Evaluación", "Postprocesado" y "Presentación".
- La fase de identificación se realiza mediante la determinación de qué elemento es más relevante en la red y, analizando su información, se delimita el alcance de la investigación, puesto que se identifican qué dispositivos están conectados a él. Una vez evaluado, se pasa a estudiar aquellos que no estén directamente conectados a él.
- La fase de adquisición se centra en los datos no volátiles, y establece el siguiente orden de elección de métodos, dando preferencia a aquellos que respetan la integridad de la fuente de evidencia: extracción y adquisición, JTAG/UART e ISP, *chip-off* y adquisición en vivo.
- La fase de análisis abre la posibilidad a seguir un método en vivo siempre y cuando no sea necesario garantizar la integridad de la fuente de evidencia o cuando no se pueda realizar la adquisición. También se destaca la falta de herramientas forenses para IoT y cómo un investigador debe compensarlo con el conocimiento del sistema a analizar.
- Se añade una fase de evaluación, que tiene por objetivo el estudio de las evidencias extraídas y la extracción de conclusiones desde la perspectiva global y no solamente desde el punto de vista de los dispositivos.

Además, de esta forma es posible determinar de forma más sencilla qué efecto tiene la incorporación del *edge computing* en el desarrollo de metodologías, puesto que se puede realizar una evaluación directa de los resultados comparando ambas propuestas.

IV-B. Descripción de la metodología

El uso de la tecnología *edge computing* en esta metodología se traduce en la inclusión de un nodo dentro de la red IoT capaz de interactuar tanto con los dispositivos que se encuentran dentro de ella como con el exterior. Esto permite el acceso a los datos que se están siendo intercambiados en todo momento en la red IoT y también a los almacenados en los propios dispositivos. Además, al poder comunicarse con el exterior, puede ofrecer información sobre el estado de la red y apoyarse en otro tipo de servicios para ampliar los servicios que puede dotar. En cuanto a las características técnicas de este dispositivo, se necesita de un nodo capaz de ejecutar tareas de cierta complejidad, por lo que no sirve cualquier tipo de dispositivo IoT. Un ejemplo válido sería una Raspberry Pi [25]. Igualmente, se debe tener en cuenta que, de cara a poder acceder ejecutar todos los servicios que debe prestar el nodo *edge*, éste debe disponer de un sistema operativo que permita su despliegue, como podría ser Ubuntu Core [26], que es compatible con placas Raspberry Pi. Por último, debe asegurarse su compatibilidad con los protocolos que se estén usando en la red IoT, de cara a poder capturar y analizar las comunicaciones intercambiadas, aspecto que cumplen ambos de los ejemplos aportados.

Otro elemento clave que puede formar parte de esta metodología es la nube, que puede complementar las limitaciones técnicas del nodo *edge*. De esta manera, se puede encargar del almacenamiento de los datos capturados y de la ejecución de los algoritmos de detección de incidencias, procesos se serán detallados más adelante. En la Figura 1 encontramos una representación gráfica de cómo estaría organizado el entorno.

Respecto a cómo cambian las fases de la metodología usada como referencia, durante esta sección se detalla el impacto que tiene la inclusión de este nodo *edge*. Sin embargo, puesto que las fases de "Evaluación" y "Presentación" no se ven afectadas en gran medida, no se describen de forma específica. Respecto a la primera, la extracción de conclusiones desde la perspectiva global se hace una vez que se han capturado las evidencias del sistema, y esto es una tarea inherente al investigador y en la que el nodo *edge* no puede asistir. En cuanto a la fase de "Presentación", nos encontramos en la misma situación, la realización del informe y la devolución de las fuentes de evidencia es un proceso que no guarda relación con el nodo *edge*. En cambio, el proceso de recuperación del sistema sí que se beneficia de la inclusión de este dispositivo, por lo que pasa a ser una nueva fase en esta propuesta.

Del mismo modo, la posibilidad de llevar a cabo tareas forenses preventivas con el uso de la tecnología *edge computing* hace que sea necesario la creación de una nueva fase destinada a la monitorización, la cuál hemos denominado "Detección".

Por tanto, las fases que conforman esta metodología son:

- Detección: se llevan a cabo las tareas preventivas de monitorización y detección de incidentes en la red IoT.

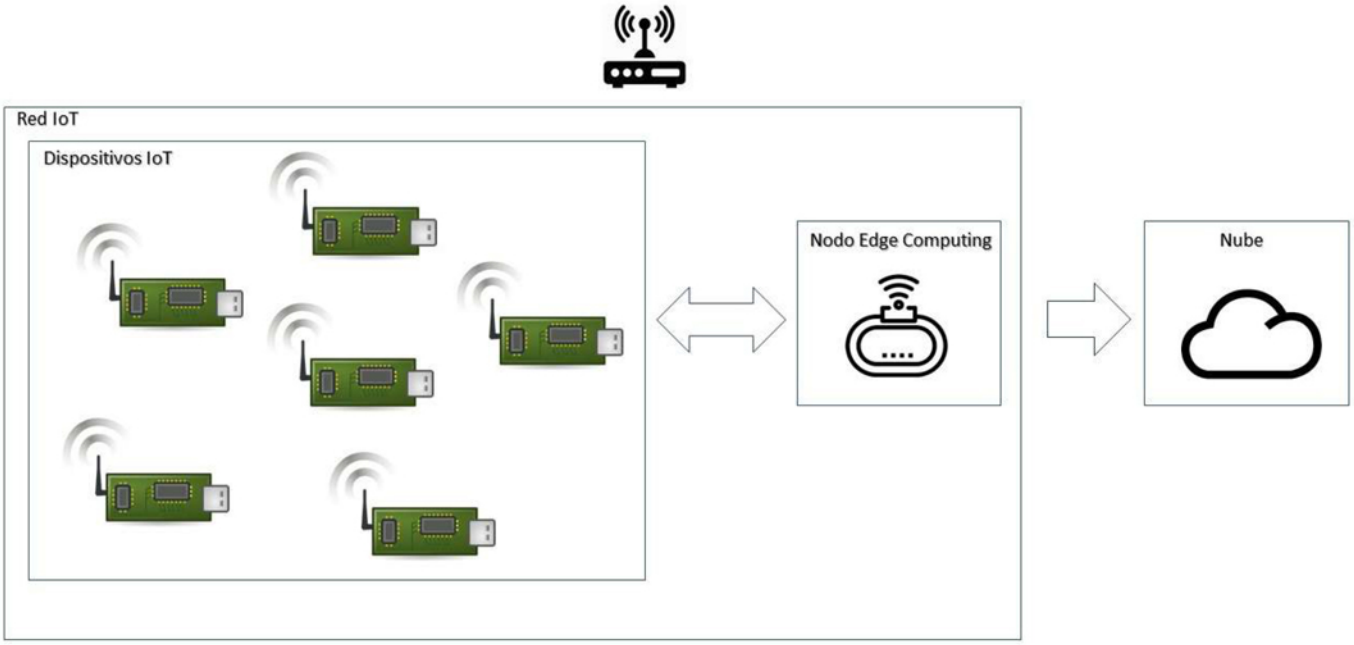


Figura 1. Representación gráfica del entorno.

- Preprocesado: el investigador forense se prepara para llevar a cabo el análisis forense y diseñar el plan de actuación.
- Identificación: se determinan qué elementos en la escena son susceptibles de contener evidencias.
- Adquisición: se realiza la captura de los datos almacenados por las fuentes de evidencia y se preservan para conservar su integridad.
- Análisis: se examinan los datos de cara a detectar evidencias que permitan extraer conclusiones sobre el incidente.
- Evaluación: se evalúan las evidencias detectadas y se extraen conclusiones, esta vez desde la perspectiva de la red.
- Presentación y postprocesado: se presentan los resultados del análisis y se realizan las tareas para dar por cerrada la investigación.
- Recuperación: se devuelve la red IoT a un estado funcional, en caso de que sea necesario.

IV-B1. Detección: El contar con un dispositivo que está en todo momento accediendo a los datos intercambiados en la red hace que podamos reducir el tiempo de respuesta ante la materialización de un incidente al mínimo. Además, también existe la posibilidad de poder acceder a los datos almacenados por un dispositivo IoT en caso de que éste disponga de algún servicio de acceso remoto, como puede ser SSH.

Al tener acceso a estos datos, es posible para el nodo *edge* el llevar a cabo tareas de monitorización con el objetivo de detectar un incidente en la red. Para ello, en esta propuesta se hace uso de la detección de anomalías, que, como muestran [27], [28], [29] o [30], permite localizar un comportamiento anómalo de un dispositivo o de la propia red de forma satisfactoria.

De este modo, se evalúan de forma periódica las comunicaciones de red y el estado de los dispositivos de la misma, analizando tanto sus datos volátiles como no volátiles. En caso de detectar una anomalía, el nodo *edge* se encarga de notificar al encargado de la red y procede a llevar a cabo las tareas correspondientes para asegurar que las posibles evidencias no se pierden, aspecto que detallaremos más adelante.

Esta tarea también puede ser ejecutada por la nube en caso de que el nodo *edge* se vea desbordado por la cantidad de datos y/o dispositivos. En este caso, el nodo envía los datos a la nube y recibe los resultados del análisis de los mismos.

IV-B2. Preprocesado: De cara a determinar el plan de actuación, es importante determinar, entre otros aspectos, el número de dispositivos que se encuentran en la red IoT, para, posteriormente, evaluar si son susceptibles de contener datos relevantes para la investigación. Esta tarea es muy sencilla de llevar a cabo puesto que el nodo *edge* está en contacto con los elementos de la red, lo que permite llevar un seguimiento de los mismos. Esto también permite saber específicamente qué modelo de dispositivos son, lo cuál es de extrema utilidad de cara a que el investigador determine qué opciones tiene de cara a adquirir y analizar los datos que contienen.

También, el hecho de que el nodo *edge* tenga acceso a los dispositivos de la red hace que pueda realizar alguna acción sobre ellos en caso de que el investigador lo desee. Por ejemplo, si se sospecha de la presencia de un malware, se podría enviar la orden de apagado a los dispositivos para que no se extendiese por la red.

IV-B3. Identificación: En esta fase se produce un cambio muy grande a la hora de actuar. Mientras que en la metodología de referencia se habla de analizar aquel dispositivo que tiene mayor importancia en la red de cara a determinar el rango de la escena, en este caso la tarea se simplifica sobremanera gracias al nodo *edge*. Como se ha comentado en el apartado anterior, dicho nodo realiza un seguimiento constante de los dispositivos en la red y las comunicaciones que realizan entre ellos. Por tanto, es muy sencillo saber

el número de dispositivos en la red, cuáles de ellos siguen activos o si va a ser posible o no adquirir de forma directa los datos que almacenan (nos apoyamos en sus especificaciones técnicas). De esta forma el investigador no necesita realizar la adquisición y/o análisis del nodo centrar para llevar a cabo este proceso.

Además, el tener acceso a las comunicaciones realizadas ayuda al analista a determinar la importancia que tienen los dispositivos en la red, pudiendo así priorizar la adquisición y análisis de unos sobre otros.

IV-B4. Adquisición: Mediante la inclusión del nodo *edge* aseguramos el poder llevar a cabo un proceso de adquisición, ya que, como se ha comentado en la Sección II, esto no siempre ocurre cuando se realizan análisis forenses en el Internet de las Cosas. Sin embargo, mediante el uso de esta tecnología no se puede asegurar que la adquisición de los datos de un dispositivo sea haga de forma completa. En el peor de los casos, el analista tiene acceso, al menos, a los datos relativos a las comunicaciones y seguimiento de los dispositivos. De esta forma, se añade un grado importante de flexibilidad en la fase: el investigador tiene la opción, cuando se pueda acceder a él, de realizar la adquisición de forma física mediante métodos como la extracción y adquisición, JTAG/UART o ISP o *chip-off*, pero, a su vez, también tiene acceso a los datos recopilados por el nodo, pudiendo así decidir si es necesario o no llevar a cabo la adquisición física.

El nodo, una vez que ha detectado la anomalía en el sistema, envía la captura de las comunicaciones a la nube, donde queda almacenada. Además, procede a realizar, en los dispositivos a los que tiene acceso, el proceso de adquisición. Si el sistema al que está accediendo permite la realización de una adquisición en vivo, el nodo ejecuta los comandos necesarios para llevar a cabo la tarea y traslada el resultado a la nube. En caso de que no sea posible, procede a realizar una copia lógica de sus datos, que también son enviados a la nube. Junto con los datos, la nube recibe información sobre la fecha y hora a la que se realizó la captura, el código *hash* de la misma y el modelo del dispositivo del que se ha realizado.

Por otra parte, el almacenamiento de los datos en la nube se lleva a cabo con todas las garantías puesto que, como hemos visto en la Sección III, existen técnicas criptográficas para asegurar la integridad y protección de los datos.

IV-B5. Análisis: Aunque el proceso de análisis depende de la habilidad del investigador y de la posibilidad de poder ejecutar o no las herramientas necesarias para poder extraer información, el desempeño del nodo *edge* ayuda al analista a saber qué información del dispositivo se puede extraer de forma online. Es decir, si no existen datos relativos a un dispositivo, esto significa que la interacción con el mismo está descartada y que el analista debe optar por llevar a cabo otro tipo de análisis. Del mismo modo, en el caso de que sí existan datos de un dispositivo, primeramente, esos datos pueden ser analizados por el investigador para determinar si contienen evidencias o no, y, además, le sirve de guía para saber qué datos se pueden extraer de ese dispositivo en concreto. Por tanto, el uso de la tecnología *edge computing*, aunque no facilita el proceso de extracción de conclusiones como tal, sí que ayuda al investigador a tener más información sobre si debe optar por un análisis en vivo, *offline* o debe apoyarse en otros dispositivos para extraer información sobre aquel que está examinando.

Igualmente, al haber estado monitorizando el comportamiento de la red y de los dispositivos, el investigador tiene acceso a un historial de datos que le permite conocer cómo funcionaba la red antes de que se iniciara el caso.

IV-B6. Recuperación: De cara a devolver al entorno IoT en el que ocurrió el incidente a un estado funcional, la incorporación del nodo *edge* proporciona dos ventajas clave. La primera de ellas es que, gracias a su capacidad de monitorización, permite evaluar si la causa del incidente sigue presente en la red o no, ahorrando ese trabajo al analista. Del mismo modo, si dicha causa sigue presente, y es necesario limpiar el entorno y restaurar los sistemas, al realizarse copias periódicas de los mismos, el proceso es mucho más ágil. El investigador solamente tiene que acceder a la nube y realizar la recuperación. Además, de cara a evaluar la efectividad de las acciones llevadas a cabo, el proceso de monitorización se vuelve a erigir como protagonista, ya que se encarga de detectar posibles anomalías remanentes y, gracias al haber capturado las comunicaciones cuando la red funcionaba correctamente, permite al analista comparar de forma sencilla si el comportamiento es el correcto.

V. Conclusiones

En esta propuesta hemos abordado el diseño de metodologías forenses centradas en modelar el entorno IoT. Se han motivado las razones por las que el uso de una metodología convencional para la realización de análisis forenses en el Internet de las Cosas no sería la mejor opción, puesto que las características de los dispositivos demandan un cambio grande a la hora de afrontar las investigaciones. Aspectos como el rango que tienen las mismas, la imposibilidad en ciertos casos de poder adquirir los datos o la volatilidad de los mismos marcan la diferencia con respecto a las investigaciones tradicionales. A su vez, se ha estudiado la posibilidad de hacer uso de una tecnología novedosa como es el *edge computing* para solventar los problemas de procesamiento de información que tienen los dispositivos y sistemas IoT, y así poder dotar a los analistas forenses de un nodo capaz de detectar incidentes y adquirir datos relevantes relativos al resto de dispositivos que conforman la red IoT. La inclusión de ese nodo permitiría a los investigadores forenses contar con un aliado que actúa como intermediario entre el propio investigador y la red, permitiéndole acceder a los datos de forma mucho más sencilla y rápida, e incluso almacenando algunos que no se podrían haber capturado de ninguna otra manera, al ser *on-the-fly*.

Para desarrollar la propuesta se ha adaptado una metodología forense en IoT existente y se ha evaluado cómo se vería modificada al incluir un nodo que hiciera uso de la tecnología *edge computing*. En concreto, se han abordado las fases de "Preprocesado", "Identificación", "Adquisición" y "Análisis"', que son las que más cambian al incluir esta nueva tecnología, y se han incluido dos nuevas denominadas "Detección", que hace referencia al proceso de monitorización y detección de incidentes, y "Recuperación", relativa a la vuelta a la funcionalidad del sistema IoT. De este modo, aseguramos el partir de un modelo aprobado por la comunidad y que se considera efectivo a la hora de abordar las investigaciones,

permitiendo, a su vez, ver de forma clara qué beneficios aporta el *edge computing* al análisis forense en IoT.

V-A. Trabajo Futuro

Esta investigación ha servido para dar un primer paso para evaluar la viabilidad de incorporar la tecnología *edge computing* en el desarrollo de las metodologías forenses que modelan el entorno IoT. Por tanto, existe una gran cantidad de estudios adicionales que se podrían realizar que están relacionados con la materia. Algunos ejemplos son los siguientes:

- La implementación de la metodología propuesta de cara a realizar una evaluación desde el punto de vista práctico.
- La realización de una comparativa práctica en diversos escenarios entre una metodología IoT que no utiliza el *edge computing* y la propuesta desarrollada en esta investigación.
- La evaluación de la metodología en contextos determinados, para así poder ampliar el rango de aplicación de la misma.

AGRADECIMIENTOS

Este trabajado cuenta con el apoyo de la Universidad de Castilla-La Mancha, mediante el contrato con referencia 2018-PREDUCLM-7476 y el proyecto con referencia 2021-GRIN-31042, del Ministerio de Ciencia e Innovación mediante los contratos con referencias FPU 17/03105 y FPU 17/02007, del Ministerio de Asuntos Económicos y Transformación Digital, mediante el proyecto con referencia RTI2018-098156-B-C52, y de la Consejería de Educación, Cultura y Deportes de la Junta de Comunidades de Castilla-La Mancha, mediante el proyecto con referencia SBPLY/17/180501/000353.

REFERENCIAS

[1] I. Analytics, "State of the IoT 2020: 12 billion IoT connections, surpassing non-IoT for the first time," https://iot-analytics.com/state-of-the-iot-2020-12-billion-iot-connections-surpassing-non-iot-for-the-first-time/, 2020.

[2] Gartner Inc., "Gartner Says 8.4 Billion Connected "Things"Will Be in Use in 2017, Up 31 Percent From 2016," https://www.gartner.com/en/newsroom/press-releases/2017-02-07-gartner-says-8-billion-connected-things-will-be-in-use-in-2017-up-31-percent-from-2016.

[3] Dan Demeter and Marco Preuss and Yaroslav Shmelev, "IoT: a malware story - Securelist," https://securelist.com/iot-a-malware-story/94451/.

[4] V. Chebyshev, F. Sinitsyn, D. Parinov, O. Kupreev, E. Lopatin, A. Kulaev, and A. Kolesnikov, "IT threat evolution Q3 2020. Non-mobile statistics," https://securelist.com/it-threat-evolution-q3-2020-non-mobile-statistics/99404/.

[5] V. R. Kebande and I. Ray, "A generic digital forensic investigation framework for internet of things (iot)," in *2016 IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud)*, Aug 2016, pp. 356–362.

[6] A. Nieto, R. Rios, and J. Lopez, "A methodology for privacy-aware iot-forensics," in *2017 IEEE Trustcom/BigDataSE/ICESS*, Aug 2017, pp. 626–633.

[7] T. Zia, P. Liu, and W. Han, "Application-specific digital forensics investigative model in internet of things (iot)," in *Proceedings of the 12th International Conference on Availability, Reliability and Security*, ser. ARES '17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10,1145/3098954,3104052

[8] M. Harbawi and A. Varol, "An improved digital evidence acquisition model for the internet of things forensic i: A theoretical framework," in *2017 5th International Symposium on Digital Forensic and Security (ISDFS)*, 2017, pp. 1–6.

[9] X. Feng, E. S. Dawam, and S. Amin, "A new digital forensics model of smart city automated vehicles," in *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, June 2017, pp. 274–279.

[10] M. Hossain, R. Hasan, and S. Zawoad, "Trust-iov: A trustworthy forensic investigation framework for the internet of vehicles (iov)," in *2017 IEEE International Congress on Internet of Things (ICIOT)*, June 2017, pp. 25–32.

[11] A. Goudbeek, K.-K. R. Choo, and N.-A. Le-Khac, "A forensic investigation framework for smart home environment," 08 2018, pp. 1446–1451.

[12] E. Al-Masri, Y. Bai, and J. Li, "A fog-based digital forensics investigation framework for iot systems," in *2018 IEEE International Conference on Smart Cloud (SmartCloud)*, 2018, pp. 196–201.

[13] L. Sadineni, E. Pilli, and R. B. Battula, "A holistic forensic model for the internet of things," in *Advances in Digital Forensics XV*, G. Peterson and S. Shenoi, Eds. Cham: Springer International Publishing, 2019, pp. 3–18.

[14] N. K. Bharadwaj and U. Singh, "Acquisition and analysis of forensic artifacts from raspberry pi an internet of things prototype platform," in *Recent Findings in Intelligent Computing Techniques*, P. K. Sa, S. Bakshi, I. K. Hatzilygeroudis, and M. N. Sahoo, Eds. Singapore: Springer Singapore, 2019, pp. 311–322.

[15] V. R. Kebande, N. M. Karie, A. Michael, S. Malapane, I. Kigwana, H. S. Venter, and R. D. Wario, "Towards an integrated digital forensic investigation framework for an iot-based ecosystem," in *2018 IEEE International Conference on Smart Internet of Things (SmartIoT)*, 2018, pp. 93–98.

[16] D. H. Kasukurti and S. Patil, "Wearable device forensic: Probable case studies and proposed methodology," in *Security in Computing and Communications*, S. M. Thampi, S. Madria, G. Wang, D. B. Rawat, and J. M. Alcaraz Calero, Eds. Singapore: Springer Singapore, 2019, pp. 290–300.

[17] E. M. Tordera, X. Masip-Bruin, J. Garcia-Alminana, A. Jukan, G.-J. Ren, J. Zhu, and J. Farre, "What is a fog node a tutorial on current concepts towards a common definition," 2016.

[18] J. Ni, A. Zhang, X. Lin, and X. S. Shen, "Security, privacy, and fairness in fog-based vehicular crowdsensing," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 146–152, 2017.

[19] T. Cruz, L. Rosa, J. Proença, L. Maglaras, M. Aubigny, L. Lev, J. Jiang, and P. Simoes, "A cyber security detection framework for supervisory control and data acquisition systems," *IEEE Transactions on Industrial Informatics*, vol. 12, 07 2016.

[20] T. Cruz, J. Barrigas, J. Proença, A. Graziano, S. Panzieri, L. Lev, and P. Simões, "Improving network security monitoring for industrial control systems," in *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, 2015, pp. 878–881.

[21] E. Oriwoh and P. Sant, "The forensics edge management system: A concept and design," in *2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing*, Dec 2013, pp. 544–550.

[22] A. Razaque, M. Aloqaily, M. Almiani, Y. Jararweh, and G. Srivastava, "Efficient and reliable forensics using intelligent edge computing," *Future Generation Computer Systems*, vol. 118, pp. 230–239, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X21000224

[23] J. M. Castelo Gómez, J. Carrillo Mondéjar, J. Roldán Gómez, and J. L. Martínez Martínez, "A context-centered methodology for IoT forensic investigations," *International Journal of Information Security*, Nov. 2020. [Online]. Available: https://doi.org/10,1007/s10207-020-00523-6

[24] Y. Yusoff, R. Ismail, and Z. Hassan, "Common phases of computer forensics investigation models," *International Journal of Computer Science & Information Technology (IJCSIT)*, vol. 3, 07 2011.

[25] Raspberry Pi Foundation, "Buy a Raspberry Pi 3 Model B – Raspberry Pi," https://www.raspberrypi.org/products/raspberry-pi-3-model-b/.

[26] Canonical Group, "Ubuntu Core - Ubuntu," https://ubuntu.com/core.

[27] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: Methods, systems and tools," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 303–336, 2014.

[28] S. Al-Haj Baddar, A. Merlo, and M. Migliardi, "Behavioral-anomaly detection in forensics analysis," *IEEE Security Privacy*, vol. 17, no. 1, pp. 55–62, 2019.

[29] I. Vural and H. Venter, "Mobile botnet detection using network forensics," in *Future Internet - FIS 2010*, A. J. Berre, A. Gómez-Pérez, K. Tutschku, and D. Fensel, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 57–67.

[30] F. Maggi, S. Zanero, and V. Iozzo, "Seeing the invisible: Forensic uses of anomaly detection and machine learning," *SIGOPS Oper. Syst. Rev.*, vol. 42, no. 3, p. 51–58, Apr. 2008. [Online]. Available: https://doi.org/10,1145/1368506,1368514

Sesión de Investigación A4: Análisis forense y cibercrimen
http://doi.org/10.18239/jornadas_2021.34.25

Reinforcement of age estimation in forensic tools to detect Child Sexual Exploitation Material

Rubel Biswas *†, Deisy Chaves *†, Franciso Jáñez-Martino *†, Pablo Blanco-Medina *†, Eduardo Fidalgo *†, Carlos García-Olalla †, George Azzopardi ‡

*Department of Electrical, Systems and Automation, Universidad de León, León, ES
†Researcher at INCIBE (Spanish National Cybersecurity Institute), León, ES
‡Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Groningen, NL
Email:{rubel.biswas, deisy.chaves, francisco.janez, pablo.blanco, eduardo.fidalgo}@unileon.es,
carlos.olalla@incibe.es, g.azzopardi@rug.nl

***Abstract*—Several image-based approaches for estimating the age of a person are available in computer vision literature. However, most of them perform poorly on minors and young adults, especially when the eyes are occluded. This type of occlusion is common in Child Sexual Exploitation Materials (CSEM), in order to hide the identity of victims. We introduce an approach that builds Soft Stagewise Regression Network (SSR-Net) models with natural and eye-occluded facial images, to estimate the age of minors and young adults. Our proposal reduces the Mean Absolute Error from 7.26 to 6.5, and 6.81 to 4.07 for SSR-Net pre-trained models on the IMDB and MORPH datasets, respectively.**

***Index Terms*—Age estimation, Occlusion, SSR-Net model, CSEM, Forensic images**

Type of contribution: *Research already published – Improving Age Estimation in Minors and Young Adults with Occluded Faces to Fight Against Child Sexual Exploitation [1]*

I. Introduction

In forensic applications, accurate and fast age estimation solutions enhance the detection of victims in Child Sexual Exploitation Materials (CSEM) [1]. Forensic tools may also support Law Enforcement Agencies (LEAs) in identifying criminals through enhanced image analysis [2].

Age estimation is a challenging problem due to factors such as pose and illumination variation, which are commonly found in CSEM images [3]. It is also common for offenders to use accessories or black stripes to hide the face or eyes of the victims [4], which presents further challenges to the performance of age estimators.

An increasing number of deep-learning-based age estimators have been proposed during the last years. However, most of these approaches are designed for the age interval between 0 and 60+ years, and are trained with unbalanced data [5], [6]. Thus, many of them do not perform well for minors and young adults, aged between 0 and 25 years old.

To address this problem, we present an improved solution for the age estimation of minors and young adults [1] by training Soft Stagewise Regression Network (SSR-Net) models [5] using natural face images and faces with occluded eyes.

II. Related Work

Due to the advancement of deep learning architectures, the performance of age estimators has improved significantly in recent years [7], [5], [8]. Despite this, to our knowledge, there are very few approaches that estimate the age of minor/young adults [9] or eye-occluded facial images [10].

Zhang et al. [8] introduced an accurate age estimation model by combining Long Short-Term Memory (LSTM) networks which are complex and computationally intensive. In contrast, Yang et al. proposed a lightweight age estimation model, called SSR-Net [5], based on the Deep EXpectation (DEX) model [7]. Likewise, Zhang et al. [6] introduced a compact model using cascaded training and multi-scale context to estimate the age with small-scale facial images. These compact models are preferable for real-time tasks due to a reduced computational cost.

III. Methodology

We introduce a two-fold solution for age estimation of minors and young adults, as presented in Fig. 1.

First, we created a balanced dataset with natural face images of minors and young adults and their corresponding eye-occluded versions. The natural facial images, in the range $[0, 25]$ years, were collected from five different well-known datasets, namely IMDB-WIKI, APPA-REAL, AgeDB, UTKFace, and Diversity in Faces, IBM (DiF).

We gathered a total of 130000 minor and young adult images by inspecting these datasets manually, removing images with an incorrect age label or without any human face. Afterwards, we created the occluded version of these images by locating the eye region using the Multi-Task Cascade Convolutional Neural Network (MTCNN) [11] and then masking it in order to simulate the referred conditions on CSEM.

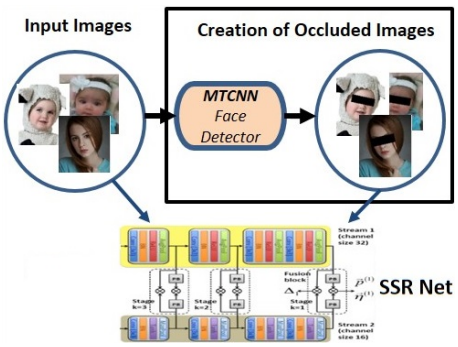


Figure 1. Steps to train age estimation models in minors and young adults.

Lastly, both image sets were merged into one.

Using these images, we implemented a lightweight pre-trained SSR-Net age estimator [5] to build new, fine-tuned age estimation models focused on minor and young adults. Our images were resized to 64×64 pixels to fine-tune the model. Furthermore, we split the dataset into a training (80%) and a test (20%) set using stratified random sampling.

IV. Experimental results

We evaluated the age estimation performance using the Mean Absolute Error (MAE) of the SSR-Net pre-trained models that have been trained with face images considering the age range $[0, 25]$ years from four balanced datasets varying in size, [6500-130000], and with two unbalanced datasets, namely MORPH and IMDB.

Then, we measured the MAE's performance enhancement of fine-tuned age estimators using our non-occluded (Org.), eye-occluded (Ocl.), and a combination of both types (Org. - Ocl.) of minor and young adult facial images. Hence, eight different models were assessed per each image type: Org., Ocl., and Org.-Ocl. Our results are presented in Table I.

We noticed that the age estimation performance was more stable in SSR-Net models —pre-trained on the IMDB dataset— fine-tuned with our merged dataset. These models achieved the best MAE of 3.58 and 4.19 for non-eye-occluded and eye-occluded images, respectively. Regarding MAE distribution, errors are heterogeneous: the MAE for age groups 0, 4-9, and 23-25 years is higher than for age range 0-25 years.

Furthermore, we compared our results with the best SSR-Net model against a state-of-art approach, VGG16-based DEX model, trained with our merged dataset. The proposed models outperformed the DEX model with MAE of 6.5 for non-eye-occluded and eye-occluded facial images. In addition, the size of the SSR-Net-based age estimators was much lower than the DEX age estimator, with sizes of $< 1MB$ and $500MB$, respectively. Moreover, it predicts the age in 0.006 seconds from a facial image.

Lastly, we have successfully integrated our proposal, i.e. fine-tuned SSR-Net age estimation model, into the 4NSEEK[1] tool to support the detection of minors on CSEM.

Table I
MAE values of SSR-Net age estimation models. The best MAE values are highlighted in bold.

Model	# images per age		MORPH dataset MAE Test		IMDB dataset MAE Test	
	Train	Test	Org.	Ocl.	Org.	Ocl.
Pre-trained MORPH	–	50	7.16	6.53	7.51	6.93
	–	100	7.19	6.56	7.52	6.93
	–	200	7.17	6.56	7.54	6.94
	–	1000	7.06	6.55	7.52	6.99
Fine-tuned Org. Img.	200	50	5.37	7.04	4.56	6.25
	400	100	4.40	6.82	4.27	6.53
	800	200	4.24	7.32	4.13	6.64
	4000	1000	**3.63**	7.93	**3.58**	6.46
Fine-tuned Ocl. Img.	200	50	6.57	5.67	5.73	5.22
	400	100	6.03	5.29	5.63	5.08
	800	200	5.80	4.97	5.72	5.07
	4000	1000	5.71	**4.22**	5.35	4.58
Fine-tuned Org. - Ocl. Img.	200	50	6.08	5.81	5.00	5.13
	400	100	6.01	5.19	4.91	5.23
	800	200	4.61	4.75	4.47	4.66
	4000	1000	3.93	4.44	3.95	**4.19**

V. Conclusions

We present an improved age estimator focused on minors and young adults with SSR-Net models, fine-tuned using natural and eye-occluded face images. Results show that our solution performs better in minors and young adults (MAE of 4.07) in comparison to the DEX model (MAE of 6.5), being more robust against eye occlusion.

Moreover, our SSR-Net-based estimators are compact models and suitable for any hardware despite memory capability, as well as forensic applications of child detection on CSEM. As future work, the impact of the gender on the MAE for age estimation will be analyzed.

Acknowledgements

This work was supported by the framework agreement between the Universidad de León and INCIBE (Spanish National Cybersecurity Institute) under Addendum 01. Also, this research has been funded with support from the European Commission under the 4NSEEK project with Grant Agreement 821966. This publication reflects the views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein. Finally, we acknowledge NVIDIA Corporation with the donation of the TITAN Xp and Tesla K40 GPUs used for this research.

References

[1] D. Chaves, E. Fidalgo, E. Alegre, F. Jánez-Martino, and R. Biswas, "Improving age estimation in minors and young adults with occluded faces to fight against child sexual exploitation," in *VISIGRAPP*, 2020, pp. 721–729.

[2] A. Gangwar, E. Fidalgo, E. Alegre, and V. González-Castro, "Pornography and child sexual abuse detection in image and video: A comparative evaluation," in *8th International Conference on Imaging for Crime Detection and Prevention (ICDP)*, 2017, pp. 37–42.

[3] D. Chaves, E. Fidalgo, E. Alegre, and P. Blanco, "Improving speed-accuracy trade-off in face detectors for forensic tools by image resizing," in *V Jornadas Nacionales de Investigación en Ciberseguridad (JNIC)*, 2019, pp. 222–223.

[4] R. Biswas, V. González-Castro, E. Fidalgo, and D. Chaves, "Boosting child abuse victim identification in forensic tools with hashing techniques," in *V Jornadas Nacionales de Investigación en Ciberseguridad (JNIC)*, 2019, pp. 344–345.

[5] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, and Y.-Y. Chuang, "SSR-Net: A compact soft stagewise regression network for age estimation," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018, pp. 1–7.

[6] C. Zhang, S. Liu, X. Xu, and C. Zhu, "C3AE: Exploring the limits of compact model for age estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 579–12 588.

[7] R. Rothe, R. Timofte, and L. V. Gool, "DEX: Deep expectation of apparent age from a single image," in *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015, pp. 10–15.

[8] K. Zhang, N. Liu, X. Yuan, X. Guo, C. Gao, Z. Zhao, and Z. Ma, "Fine-grained age estimation in the wild with attention LSTM networks," *IEEE Trans. Circuits and Systems for Video Technology*, pp. 1–12, 2019.

[9] F. Anda, D. Lillis, A. Kanta, B. A. Becker, E. Bou-Harb, N.-A. Le-Khac, and M. Scanlon, "Improving borderline adulthood facial age estimation through ensemble learning," in *14th International Conference on Availability, Reliability and Security (ARES '19)*, 2019, pp. 1–8.

[10] L. Ye, B. Li, N. Mohammed, Y. Wang, and J. Liang, "Privacy-preserving age estimation for content rating," in *IEEE 20th International Workshop on Multimedia Signal Processing*, Aug 2018, pp. 1–6.

[11] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[1] https://www.incibe.es/en/european-projects/4nseek

http://doi.org/10.18239/jornadas_2021.34.26

A Review of "Pre-processing Memory Dumps to Improve Similarity Score of Windows Modules"

Miguel Martín-Pérez
Universidad de Zaragoza, Spain
miguelmartinperez@unizar.es

Ricardo J. Rodríguez
Universidad de Zaragoza, Spain
rjrodriguez@unizar.es

Davide Balzarotti
EURECOM, France
davide.balzarotti@eurecom.fr

***Abstract*—Memory forensics is useful to provide a fast triage on running processes at the time of memory acquisition in order to avoid unnecessary forensic analysis. However, due to the effects of the execution of the process itself, traditional cryptographic hashes are unsuitable in memory forensics. Similarity digest algorithms allow an analyst to compute a similarity score of inputs that can be slightly different. In this paper, we focus on the issues caused by relocation of Windows processes and system libraries when computing similarities between them. To overcome these issues, we introduce two methods (**GUIDED DE-RELOCATION **and** LINEAR SWEEP DE-RELOCATION**) to pre-process a memory dump. The goal of both methods is to identify and undo the effect of relocation in every module contained in the dump, providing *sanitized* inputs to similarity digest algorithms that improve similarity scores between modules.** GUIDED DE-RELOCATION **relies on specific structures of the Windows PE format, while** LINEAR SWEEP DE-RELOCATION **relies on a disassembling process to identify assembly instructions having memory operands that address the memory range of the module. We have evaluated them in different scenarios. Our results demonstrate that pre-processing memory dumps with these methods significantly improves similarity scores between memory modules. In addition, both methods have been integrated in a Volatility plugin.**

***Index Terms*—similarity digest algorithms, memory forensics, Windows, relocation**

Tipo de contribución: *Investigación ya publicada en* "Pre-processing memory dumps to improve similarity score of Windows modules," Computers & Security, vol. 101, p. 102119, 2021 [1].

I. EXTENDED ABSTRACT

Memory forensics is a branch of the computer forensics process, normally carried out as part of the detection and analysis stage in an incident response process [2]. In particular, memory forensics, unlike disk forensics, deals with the recovery of digital evidence from computer memory instead of computer storage media. Furthermore, the initial triage in memory forensics is faster than in persistent storage forensics since the quantity of data to be analyzed is smaller.

A memory forensic analyst can triage the list of processes running at the acquisition time to discard well-known processes or to focus her attention on particular ones. Thus, she needs some way to identify processes. In disk forensics, cryptographic hash (one-way) functions [3] such as MD5, SHA-1, or SHA-256 functions are commonly used for data integrity and file identification of a seized device [4]. A desirable property of any cryptographic hash function is the avalanche effect property [5], which guarantees that the hash values of two similar, but not identical, inputs produce radically different outputs. Due to this property, these crypto hash functions are unsuitable for identifying common processes that belong to the same binary application, but in different executions.

A common pitfall is to think that the content of a running process and its corresponding executable file are identical. In fact, the OS loader may apply a number of transformations when the executable file is mapped into memory. For instance, software defense techniques such as Address Space Layout Randomization ensure that executable files are mapped in memory regions that are different among consecutive executions or among consecutive system reboots (in Windows systems). Furthermore, the size of the executable file in the memory may be larger than on disk due to memory alignment issues, as the granularity of the memory subsystem OS manager determines the minimum quantity of allocated memory space (for instance, 4 KiB in Windows, macOS, and GNU/Linux).

To overcome these limitations, approximate matching or *similarity digest algorithms* (SDA) have emerged in recent years as a prominent approach that is more robust against active adversaries than traditional hashing [4]. SDA identify similarities between two digital artifacts providing a measure of similarity, normally in the range of $[0, 100]$. This similarity score enables an analyst to find out whether artifacts resemble each other or whether an artifact is contained in another artifact [6].

As mentioned above, the differences between processes are mainly motivated by the work of the relocation process. These differences, in turn, negatively affect the similarity scores provided by the similarity digest algorithms (in some cases even resulting in a similarity close to zero).

To minimize the effect of these differences, in this paper we propose two methods to process the input given to a similarity digest algorithm prior to computing its similarity hash. Both pre-processing methods undo the work performed by the relocation process, but in different ways: the method called GUIDED DE-RELOCATION relies on particular kernel-space structures that might be contained in the memory dump. These structures point to the affected bytes and allow a precise de-relocation.

While the LINEAR SWEEP DE-RELOCATION method performs a linear sweep disassembly of the binary code of a process. Specifically, it identifies all possible sub-structures of the PE format and then it performs a linear sweep disassembly of the unidentified bytes, selecting the longest

sequence of instructions as the most probable correct sequence. As the last step, instructions with operands that point to module space are normalized.

Both methods work with small memory page granularity (4 KiB). We have evaluated them by comparing the similarity scores generated by the `dcfldd`, `ssdeep`, `sdhash`, and `TLSH` SDAs, and shown that the similarity score is improved when any method is used. Figure 1 shows the similarity scores between related pages when no de-relocation is performed for 32-bit architecture, which has the worst similarity result. Similarly, Figure 2 and 3 show the improvement of the similarity when the GUIDED DE-RELOCATION method and the LINEAR SWEEP DE-RELOCATION method are applied, respectively. The results in the latter case are worse than the former one because LINEAR SWEEP DE-RELOCATION identifies many, but not all, of the bytes affected by relocation. Nevertheless, the improvement of the results shown is enough to identify successfully similar modules. Last, in Figure 4 we display the result of comparing modules processed with different methods, showing that both methods are compatible and the results are equal to the worst case. On the contrary, the results of applying de-relocation methods in 64-bit architectures are very similar to the results without any pre-processing. This is motivated because in 64-bit mode the RIP-relative addressing form was introduced, which facilitates the construction of position-independent code and therefore the bytes affected by relocation will be only the ones related to function addresses of shared libraries. We refer the reader to [1] for more details in this issue.

In addition, we have also evaluated to what extent the similarity score of each algorithm considered in the paper is affected by the loading process. We found that these algorithms are particularly sensitive to byte modifications, and that *intelligent* byte modifications can dramatically affect the similarity score for some of these algorithms.

We have developed a Volatility plugin that implements the de-relocation methods and the SDA considered in this paper. For the sake of open science, we have released it under the GNU/GPLv3 license [7]. The tool is designed in an extensible way, allowing for quick additions of SDAs.

The full version of this paper (with a full description of the experiments and limitations) was published in [1].

Figure 1. Not pre-processed 32-bit architecture

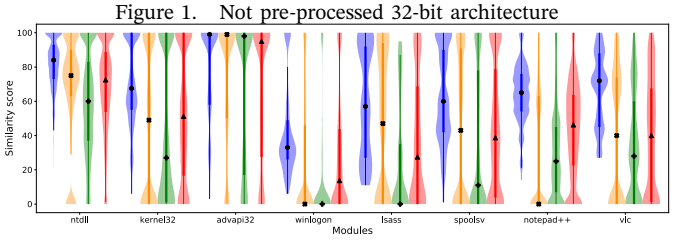


Figure 2. GUIDED DE-RELOCATION 32-bit architecture

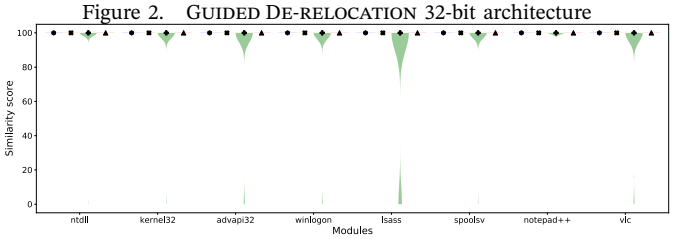


Figure 3. LINEAR SWEEP DE-RELOCATION 32-bit architecture

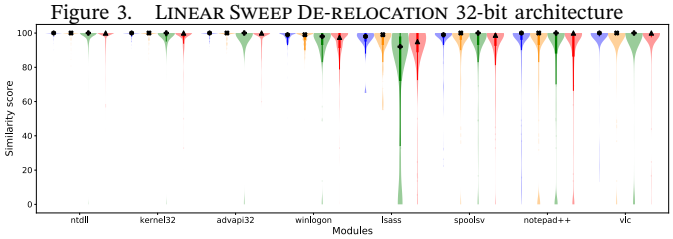
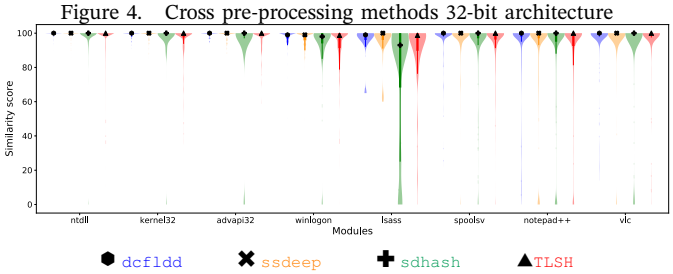


Figure 4. Cross pre-processing methods 32-bit architecture



ACKNOWLEDGEMENTS

The research was supported by the Spanish Ministry of Science, Innovation and Universities under grant MEDRESE-RTI2018-098543-B-I00 and by the University, Industry and Innovation Department of the Aragonese Government under *Programa de Proyectos Estratégicos de Grupos de Investigación* (DisCo research group, ref. T21-20R). It was also supported by the Spanish National Cybersecurity Institute (INCIBE) "Ayudas para la excelencia de los equipos de investigación avanzada en ciberseguridad", grant numbers INCIBEC-2015-02486 and INCIBEI-2015-27300. This work was also supported in part by the European Research Council (ERC) under the European Union Horizon 2020 research and innovation programme (grant agreement No 771844 BitCrumbs). This research has been developed during a short-research term in EURECOM supported by *Campus de Excelencia Internacional del Valle del Ebro* (Campus Iberus), "Convenio de subvención Erasmus+ Educación Superior para prácticas Consorcio Iberus+", and *Universidad de Zaragoza, Fundación Bancaria Ibercaja y Fundación CAI* "Programa Ibercaja-CAI de Estancias de Investigación", grant number IT 7/19.

REFERENCES

[1] M. Martín-Pérez, R. J. Rodríguez, and D. Balzarotti, "Pre-processing memory dumps to improve similarity score of Windows modules," *Computers & Security*, vol. 101, p. 102119, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404820303928

[2] P. Cichonski, T. Millar, T. Grance, and K. Scarfone, "Computer Security Incident Handling Guide," National Institute of Standards and Technology (NIST), techreport SP 800-61 Rev. 2, Sep. 2012, special Publication (NIST SP).

[3] O. Goldreich, *Foundations of Cryptography: Volume 1*. New York, NY, USA: Cambridge University Press, 2006.

[4] V. S. Harichandran, F. Breitinger, and I. Baggili, "Bytewise Approximate Matching: the Good, the Bad, and the Unknown," *Journal of Digital Forensics, Security and Law*, vol. 11, no. 2, 2016.

[5] A. F. Webster and S. E. Tavares, "On the Design of S-Boxes," in *Advances in Cryptology — CRYPTO '85 Proceedings*, H. C. Williams, Ed. Springer Berlin Heidelberg, 1986, pp. 523–534.

[6] F. Breitinger, B. Guttman, M. McCarrin, V. Roussev, and D. White, "Approximate Matching: Definition and Terminology," National Institute of Standards and Technology, techreport NIST Special Publication 800-168, May 2014.

[7] M. Martín-Pérez, "Similarity Unrelocated Module Volatility plugin," [Online; https://github.com/reverseame/similarity-unrelocated-module], Jul. 2020.

http://doi.org/10.18239/jornadas_2021.34.27

The H2020 project RAYUELA: A fun way to fight cybercrime

Gregorio López[1], Nereida Bueno[2], Mario Castro[1], María Reneses[2], Jaime Pérez[1], María Riberas[2], Manuel Álvarez-Campana[3], Mario Vega-Barbas[3], Sonia Solera-Cotanilla[3], Leire Bastida[4], Ana Moya[4], Rubén Fernández[5], Violeta Vázquez[6], Germán Zango[6], Pedro Vicente[7]
[1]Instituto de Investigación Tecnológica, ICAI, Universidad Pontificia Comillas, Madrid, Spain
[2]Facultad de Ciencias Humanas y Sociales, Universidad Pontificia Comillas, Cantoblanco, Spain
[3]ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain
[4]Fundación Tecnalia Research and Innovation, Derio, Spain
[5]Policía Local de Valencia, Valencia, Spain
[6]Zabala Innovation Consulting, Madrid, Spain
[7]Pedro Vicente, Polícia Judiciária, Lisboa, Portugal
0000-0001-9954-3504, 0000-0003-1442-7905, 0000-0001-6328-8994, 0000-0002-9708-6896, 0000-0001-6044-0022, 0000-0003-2030-0310, 0000-0003-2747-9798, 0000-0003-4506-6284, 0000-0003-3516-4489, 0000-0002-2399-2757, 0000-0001-6180-2662, 0000-0001-8507-070X

Abstract- **As in the case of maieutics, this paper aims to unveil the most important goals and features of the recently funded European project RAYUELA by answering some important questions, such as: why, who, what, how, what the main challenges and novelty of the project are, and what will be next (although the project is still in its earlier stages).**

Index Terms- **Connected devices, Cyberbullying, Cybercriminality, Cybersecurity, Data analysis, Human Trafficking, Misinformation, Online grooming, Serious games**

Tipo de contribución: *Investigación en desarrollo*

I. WHY?

Based on the UNICEF report 'The State of the World's Children 2017: Children in a Digital World', children and adolescents under 18 already account for an estimated one in three Internet users around the World [1]. Although these children and teenagers may be considered digital natives, sometimes they are not fully aware of the risks and threats, or of the benefits and opportunities that technology and the Internet offer. This very important issue can be tackled mainly from two different perspectives:

- Prevention: by teaching and training minors to make proper use of the Internet and associated technologies.
- Mitigation: by identifying potential risk profiles and implementing policies to protect them.

And is there a better way to do so than by playing? This is indeed the approach of the EU H2020 project RAYUELA (empoweRing and educAting YoUng pEople for the internet by pLAying) [2]. Fig. 1 shows the logo and motto of the project.

The name of the project is in turn inspired in the kid game hopscotch (*rayuela*, in Spanish) and in the famous Cortazar's novel with the same name, which was very provocative and innovative when it was published because its story depends on the decision the reader makes. In such a novel, Cortazar himself explained the kid game as follows [3]:

"*Hopscotch is played with a pebble that you move with the tip of your toe. The things you need: a sidewalk, a pebble, a toe, and a pretty chalk drawing, preferably in colors. On top is Heaven, on the bottom is Earth, it's very hard to get the pebble up to Heaven, you almost always miscalculate and the stone goes off the drawing. But little by little you start to get the knack of how to jump over the different squares (spiral hopscotch, rectangular hopscotch, fantasy hopscotch, not played very often) and then one day you learn how to leave Earth and make the pebble climb up into Heaven".*

So what is the Heaven of our particular RAYUELA? The answer to this question is that none other than contributing to make the Internet a better and safer place for minors.

Fig. 1. Logo and motto of the project

II. WHO?

The RAYUELA project was funded with around € 5M under the subtopic 2 of the call H2020-SU-FCT01-2019, entitled "Understanding the drivers of cybercriminality, and new methods to prevent, investigate and mitigate cybercriminal behaviour". It is a 36-months project that started in October 2020.

The project is led by Universidad Pontificia Comillas and the consortium is composed of 17 partners from 9 different countries covering the main European geographical areas, as Fig. 2 illustrates.

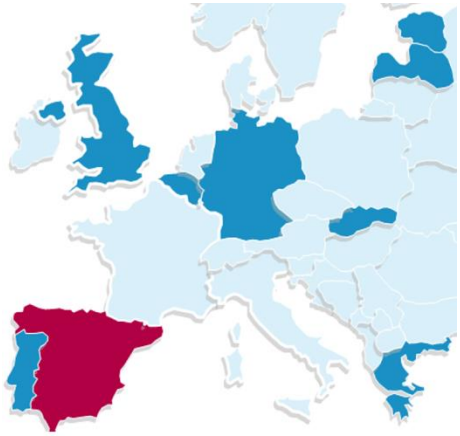


Fig. 2 RAYUELA's consortium map. In bold the countries where the project has footprint.

The consortium includes LEA (Law Enforcement Agencies), large industry companies and SME (Small to Medium Enterprise), research and academia, and educational institutions and associations. It stands out for its interdisciplinarity, bringing together LEAs, sociologists, psychologists, anthropologists, legal experts, ethicists, philosophers, educators, and computer scientists and engineers, as Fig. 3 shows.

NO.	Participant organization name	Country	Type of entity role
1 (CO)	COMILLAS – Universidad Pontificia Comillas	ES	**University**: expertise in psychology and anthropology. Complex system modelling, gaming development, data analysis.
2	UPM – Universidad Politécnica de Madrid	ES	**University**: IoT and cybersecurity: threat modelling, wearables, connected devices.
3	TECNALIA – Tecnalia Research and Innovation	ES	**Research**: serious game design and development experts.
4	TIMELEX – Timelex SCRL	BE	**SME**: Legal and GDPR expert. Privacy and data security.
5	BPI – Bratislava Policy Institute	SK	**Research**: policy experts. Specialists in qualitative research for societal threats.
6	TARTU – University of Tartu	EE	**University**: experts in ethics, philosophy, criminology, and privacy.
7	PJ – Polícia Judiciária	PT	**LEA**: Law Enforcement Agency
8	PLV – Policía Local de Valencia	ES	**LEA**: Law Enforcement Agency
9	PSNI – Police Service of Northern Ireland	UK	**LEA**: Law Enforcement Agency
10	UGENT – University of Ghent	BE	**University:** expertise in criminology and psychology.
11	TILDE – Tilde SIA	LV	**SME:** Machine Translation and Open Data experts.
12	EA – Ellinogermaniki Agogi	GR	**Educational institution**: cluster of Greek schools.
13	UCLL – UC Leuven-Limburg	BE	**Educational institution:** University College Teaching Education Department and Art of Teaching research group.
14	ALLDIGITAL – All Digital	BE	**Association:** pan-European association working with 25,000 digital competence centres.
15	ZABALA – Zabala Innovation Consulting	ES	**SME:** innovation management, communication and **dissemination** expert.
16	EPBG – Politsei- ja Piirivalveamet	EE	**LEA**: Law Enforcement Agency
17	NEC – NEC Laboratories Europe GmbH	DE	**Large industry:** Machine learning and Deep learning algorithms. Serious game data analytics.

Fig. 3 RAYUELA's list of partners including country, type of entity and expertise.

The project also counts with an IAB (International Advisory Board) which brings together international experts and relevant institutions, including LEA, public administrations and organizations, civil associations, and educational institutions. The main duties of such an IAB include providing guidance on the project direction and goals and feedback on the project progress and results, helping with generating awareness about the project and with reaching target users, and supporting the development of policies. Current members of the IAB are:

- INCIBE (Spanish National Cybersecurity Institute)
- OCC (Spanish Cybernetic Coordination Office)
- Belgian Federal Judicial Police
- Save The Children (Spain)
- Lifelong Learning Platform (Belgium)
- The Regional Department of Education, Youth and Sport of the Community of Madrid (Spain)
- CIPFP Misericordia, one of the Spanish national reference centres in vocational education

In addition, due to the importance of ethics and legal aspects, the project also counts with two external Ethics Advisors: Ofelia Tejerina-Rodríguez and Caroline Gans-Combe.

III. WHAT?

The overall goal of the project is to bring together experts from different areas of knowledge from all over Europe to develop novel methodologies that allow better understanding the drivers and human factors affecting certain relevant ways of cybercriminality, as well as empowering and educating young people (children and teenagers primarily) in the benefits, risks and threats intrinsically linked to the use of the Internet, thus preventing and mitigating cybercriminal behavior.

As it has already been said, the project aims to achieve such an overall goal "by playing", which represents a novel method to do so. In particular, the project aims to develop an interactive story-like game that, on the one side, will allow minors to learn good practices on the use of the Internet and associated technology by playing, and, on the other side, will allow modelling, in a friendly and non-invasive manner, online habits and potential risk profiles related to cybersecurity and cybercriminality, providing LEA with scientifically sound foundations to define appropriate policies.

The cybercriminality and cybersecurity topics covered in the project include cyberbullying, online grooming, human trafficking for sexual exploitation, misinformation and deception, and the technological threats and risks associated to the connected devices used by minors.

IV. HOW?

Fig. 4 illustrates the research methodology that will be followed to achieve the aforementioned goal.

The first stage of the project (shown in the left-hand side of Fig. 4) will be twofold. On the one side, as it is shown in the upper left-hand side of Fig. 4, thorough research will be carried out on the sociological, anthropological, and psychological factors affecting the considered cybercrimes (i.e., cyberbullying, online grooming, human trafficking for sexual exploitation, and misinformation and deception). Traditional research methods in Social Sciences, such as semi-structured and in-depth interviews to victims, offenders, and experts, or focus group, will be applied in this stage.

On the other side, as it is shown in the lower left-hand side of Fig. 4, thorough research will be also carried out on the technological threats associated to the use of IoT (Internet of Things) devices (e.g., connected toys, wearables, or smart personal assistants), as well as on how human factors affect to the impact of such threats. In this case, traditional research methods in engineering, such as SLR (Systematic Literature Review), will be applied together with hand-in research, such as penetration testing or honeypot deployment and analysis.

As it is show in the centre of Fig. 4, the main findings of this first research stage will be translated into the interactive story-like game, which will address these topics through different cyber-adventures in which players may end up in a risky or safe situation depending on the decisions they make. Thus, the players may "live" different stories depending on

the decisions they make while playing (and learn from them), the same way as the well-known Cortázar novel involves different stories depending on the decisions the reader makes while reading it. As a result, the game will be a safe environment where minors will face certain situations, in which they may fail and make wrong decisions, but they will have new chances to make the right ones, so they will learn good practices for behaving online in the real virtual world without taking any risk, the same way as pilots learn how to fly an actual plane in flight simulators.

Once the first prototype of the game is launched, it will be tested in several pilots across Europe, as shown in the right-hand side of Fig. 4. Such pilots will involve at least 150 secondary education students aged from 13 to 15 from the

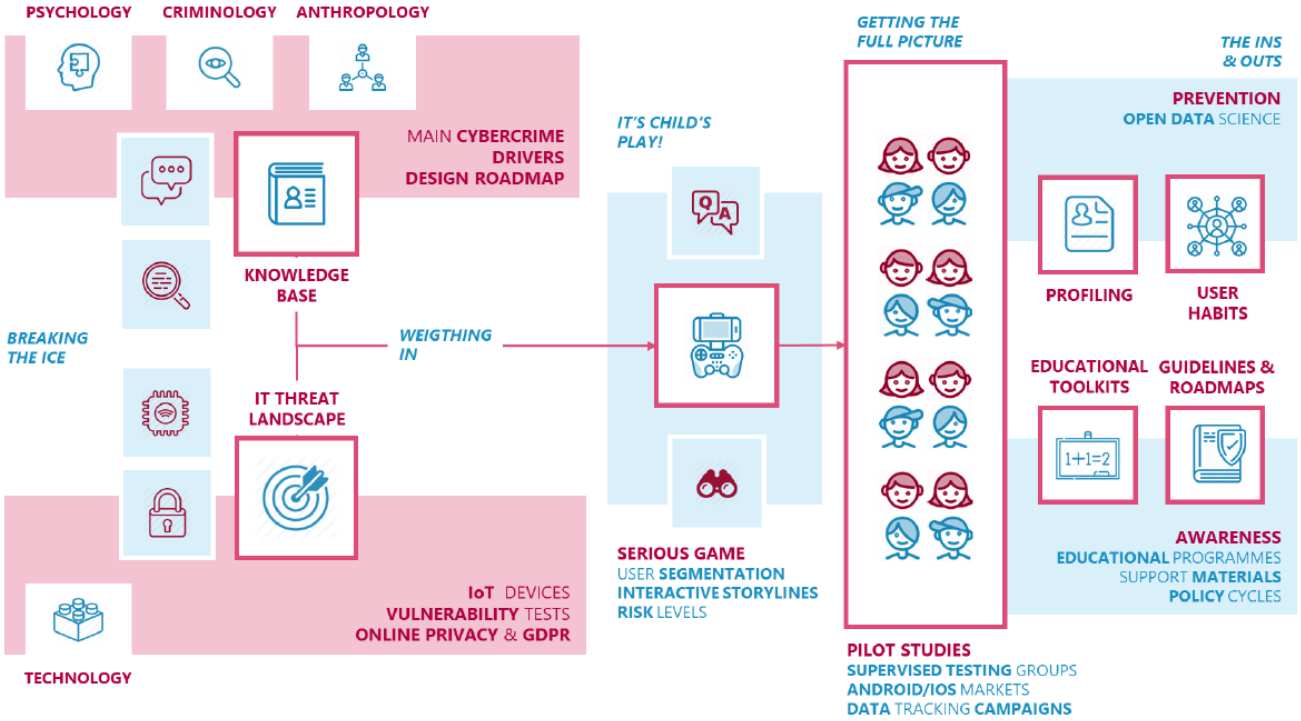


Fig. 4 Overview of the project.

Greek cluster of German schools participating in the project, but also many more youngsters in both controlled (e.g., workshops, organized events) and uncontrolled environments (e.g., downloading the game from a market).

The data gathered through the game will represent a large and diverse sample covering the most representative geographical areas in Europe. Such data will be pseudonymized and processed, combining Bayesian methods and Machine Learning/Deep Learning algorithms, and will be eventually interpreted jointly with the psychologists, sociologists, anthropologist, educators, and LEAs of the project to identify potential risky online habits and user profiles related to the considered topics.

The main conclusions of such analysis and interpretation will serve as input to the LEAs to develop evidence-based policies. Furthermore, the project will generate material to increase awareness and for capacity building among the interested stakeholders (e.g., LEAs, educators, minors, parents).

All this work will be carried out paying special attention to ethical and legal issues to avoid discrimination, stigmatization, or to prepare specific procedures for accidental findings well in advance. The workflow that has just been explained is organized in the work package structure shown in Fig. 5.

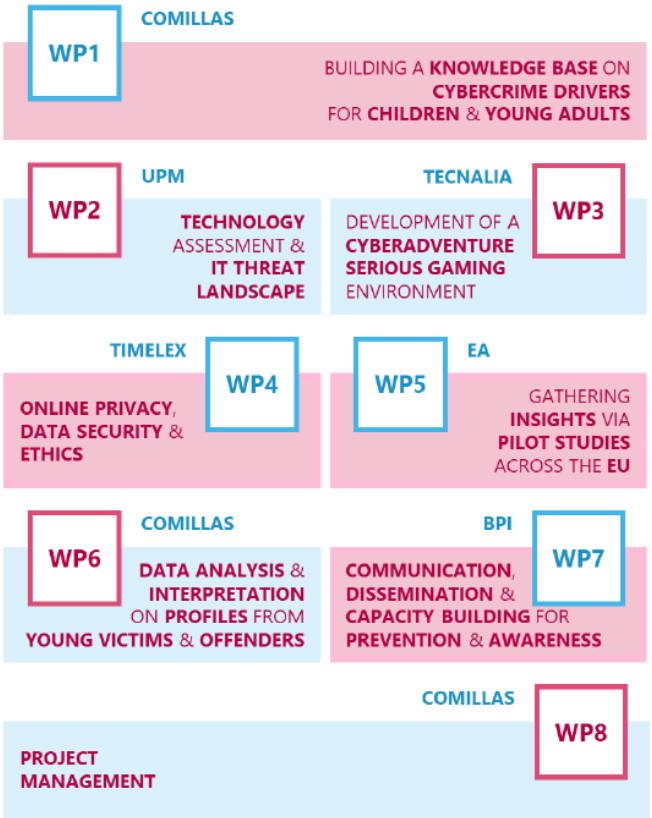


Fig. 5 RAYUELA's work package structure including work package leaders.

V. What are the main Challenges and Novelty?

Although serious games have been out there for quite a while, they have not been extensively applied to the knowledge area the project is focused on, which represents one of the innovations of the project.

Furthermore, serious games have been applied so far mainly for learning purposes and for assessing such a learning, but in this project the data gathered through the game is intended to be processed for profiling purposes as well, which represents one of the main challenges of the project. In this sense, the serious game will work as an amplifier of the traditional research methods used in Social Sciences.

Another great challenge of the project related to data analysis has to do with the lack of available data in this domain, which will make to explore, as part of the project, different approaches to generate synthetic data to test, select, and train the algorithms in advance.

In addition, unlike traditional research approaches where the impact on target users is unclear, in this case the target population will benefit from the main takeaways of the project directly by playing the game.

Last, but not least, ethical and legal issues represent definitely a challenge to carry out the research activities planned in the project being compliant with the highest standards in this regard, required by the target users of the game.

VI. What Next?

Although the project still has 30 months ahead, as a kind of outlook the project will try to promote further research by developing new serious games or by analysing the data gathered through ours for investigating, preventing and mitigating the effects of other online cybercrimes.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 882828. The authors would like to thank all the partners within the consortium for the fruitful collaboration and discussion. The sole responsibility for the content of this document lies with the authors and in no way reflects the views of the European Union.

References

[1] UNICEF, "The State of the World's Children 2017: Children in a Digital World," 2017

[2] RAYUELA's webpage: https://www.rayuela-h2020.eu

[3] J. Cortázar, "Hopscotch," 1963.

Sesión de Investigación A4: Análisis forense y cibercrimen
http://doi.org/10.18239/jornadas_2021.34.28

Development of a Hardware Benchmark for Forensic Face Detection Applications

Javier Velasco-Mata*†[a], Deisy Chaves*†[b], Verónica De Mata†[c], Mhd Wesam Al-Nabki*†[d], Eduardo Fidalgo*†[e], Enrique Alegre*†[f], and George Azzopardi‡[g]

*Department of Electrical, Systems and Automation, Universidad de León, León, ES
†Researcher at INCIBE (Spanish National Cybersecurity Institute), León, ES
‡Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Groningen, NL
Email:{**javier.velasco**, **deisy.chaves**, wesam.alnabki, eduardo.fidalgo, enrique.alegre}@unileon.es,
veronica.demata@incibe.es, g.azzopardi@rug.nl
ORCID iDs: [a]0000-0002-7923-1658, [b]0000-0002-7745-8111, [c]0000-0003-2337-1050, [d]0000-0002-3975-3478,
[e]0000-0003-1202-5232, [f]0000-0003-2081-774X, [g]0000-0001-6552-2596

***Abstract*—Face detection techniques are valuable in the forensic investigation since they help criminal investigators to identify victims/offenders in child sexual exploitation material. Deep learning approaches proved successful in these tasks, but their high computational requirements make them unsuitable if there are time constraints. To cope with this problem, we use a resizing strategy over three face detection techniques —MTCNN, PyramidBox and DSFD— to improve their speed over samples selected from the WIDER Face and UFDD datasets across several CPUs and GPUs. The best speed-detection trade-off was achieved reducing the images to 50% of their original size and then applying DSFD. The fastest hardware for this purpose was a Nvidia GPU based on the Turing architecture.**

***Index Terms*—Face Detection, GPU, CPU, Benchmark**

Type of contribution: *Already Published Research — Assessment and Estimation of Face Detection Performance Based on Deep Learning for Forensic Applications* [1]

I. Introduction

With the popularity of cameras and sharing the produced content online, offenders found a easy way to distribute Child Sexual Exploitation Materials (CSEM) [2]. As counterpart, the criminal investigation of this content has shown a growing interest worldwide [3]. However, manual analysis to identify CSEM content is infeasible due to the large amount of data to review in most investigations. Thus, there is an urgent need to develop automatic and fast systems for its detection [4].

An approach to detect CSEM is to combine a face detector with an age estimator to detect minors as possible victims [5]. Chaves et al. [3] selected three popular methods according to their processing time and accuracy and attempted to improve their speed-performance trade-off using a strategy consisting of downsizing the images. These three face detectors are the Multi-Task Cascade CNN (MTCNN) [6], the Context-Assisted Single Shot Face Detector (often referred as PyramidBox) [7], and the Dual face Shot Detector (DFSD) [8]. However, that study is limited to a GTX 1060 GPU, so it can significantly be a reference to end-users, such as law enforcement analysts, who often need to determine the most suitable hardware for these demanding computational tasks.

This study evaluates five CPUs and seven different GPU models to present a more comprehensive comparison between speed and accuracy. We evaluated the three mentioned methods from [3] using a set of images chosen from the WIDER Face Dataset [9] and the Unconstrained Face Detection Dataset (UFDD) [10] as explained in Section III.

II. Related Work

Nowadays, deep learning methods became state-of-the-art on many computer vision applications, including face detection. MTCNN [6] consists of three CNNs that simultaneously solve face detection and alignment problems. In contrast, the PyramidBox method [7] integrates multi-scale feature maps with multi-level semantic information to improve the detection of small faces. Similarly, DSFD [8] aggregates multi-scale and semantic information with enhanced features corresponding to information context to increase face detection accuracy.

Typically, face detectors are evaluated in terms of mean Average Precision (mAP). In the literature, MTCNN reports an mAP of 75.2%, PyramidBox 92.6% and DSFD 93.3% on the WIDER Face [9] dataset. However, face detection speed, which is a critical factor for end-users, is seldom reported. In [11] Nardelli et al. compared the required training time for several object detectors based on deep learning with various CPUs and GPUs, but not at the testing phase. To the best of our knowledge, no study could serve as a benchmark for face detection performance over several hardware options.

III. Methodology

We evaluated the performance of face detection by analyzing images from two datasets, WIDER Face [9] and UFDD [10]. Both datasets were chosen because they consider diverse acquisition conditions in terms of illumination, scale, pose and occlusion. Besides, in order to replicate the usual number of subjects involved in CSEM, only images with less than five people were analyzed. We manually chose 1994 images from the WIDER Face dataset and 2222 from the UFDD dataset. Following the study by Chaves et al. [3] we evaluated the performance of MTCNN, PyramidBox and DSFD face detectors (see Section II) using a resizing strategy in order to decrease the time for face detection processing. This strategy first reduces the image size and then uses the face detection method on the reduced image. Finally, the bounding boxes containing face locations are scaled back to the original image

TABLE I
AVERAGE MAP AND F1 SCORES BY FACE DETECTOR OVER IMAGES IN THE ORIGINAL SIZE OR RESIZED TO 75%, 50% AND 25%. THE IMAGES CAME FROM WIDER FACE AND UFDD DATASETS. BESIDES, IT SHOWS THE MEAN OF SECONDS TO PROCESS AN IMAGE ON DIFFERENT CPUS AND GPUS.

	100%			75%			50%			25%		
Method	MTCNN	PyramidBox	DSFD	MTCNN	PyramidBox	DSFD	MTCNN	PyramidBox	DSFD	MTCNN	PyramidBox	DSFD
Avg. mAP	37.02	71.88	**79.02**	36.69	70.13	**78.59**	34.59	64.60	**74.55**	27.03	46.68	**61.49**
Avg. F1	0.363	0.720	**0.815**	0.357	0.696	**0.804**	0.331	0.612	**0.748**	0.250	0.398	**0.572**
i5-3450	0.321	6.631	17.881	0.197	3.707	9.864	0.108	1.657	4.354	0.048	0.443	1.033
i7-8650	0.424	8.081	17.487	0.257	4.553	10.600	0.135	2.070	4.770	0.057	0.560	1.276
i7-4790K	0.215	5.679	11.546	0.129	3.161	6.300	**0.067**	1.403	2.624	**0.028**	0.370	0.565
i9-8950HK	**0.192**	**4.435**	9.714	**0.126**	**2.365**	**2.750**	0.072	**1.040**	2.292	**0.028**	**0.271**	**0.544**
Xeon E5	0.540	4.579	**8.642**	0.356	2.758	4.208	0.208	1.381	**2.203**	0.096	0.478	0.858
Tesla K40c	0.189	0.696	0.824	0.121	0.433	0.488	0.065	0.240	0.263	0.034	0.118	0.129
TITAN Xp	0.135	0.201	0.277	0.088	**0.138**	0.180	0.052	**0.092**	0.110	0.030	0.056	0.067
GTX 1050 ti	0.110	0.627	0.642	0.069	0.351	0.365	0.039	0.182	0.170	**0.018**	0.060	0.072
GTX 1060	0.121	0.347	0.356	0.071	0.220	0.214	0.039	0.115	0.107	0.020	0.049	0.050
GTX 1070	0.117	**0.258**	0.315	0.076	0.165	0.198	0.044	0.093	0.110	0.022	**0.045**	0.063
RTX 2060	**0.107**	0.458	**0.256**	**0.067**	0.268	0.165	**0.038**	0.125	**0.076**	0.019	**0.045**	**0.041**
RTX 2070	0.115	0.458	0.282	0.073	0.262	**0.162**	0.042	0.126	0.078	0.021	0.047	0.048

dimensions. We assessed the processing time of the face detectors on five Intel CPUs – i5-3450, i7-4790K, i7-8650U, i9-8950HK, and Xeon E5-2630, and seven Nvidia GPUs — Tesla K40c, TITAN Xp, GTX1050, GTX 1060, GTX 1070, RTX 2060 and RTX 2070.

IV. EXPERIMENTAL RESULTS

Table I presents the weighted averages of mAP and F1 scores obtained in each of the two datasets and the time required to process a picture in the indicated hardware. The results are classified by the resizing strategy employed: using the full-size image or resizing it to 75%, 50% or 25% of the original image. The detection scores of the three models over full sized images were lower than the ones provided by Yang et al. [9] on the full WIDER Face dataset because we used a harder-to-detect subset of samples, i.e., the ones with five people or less.

Results show that the resizing strategy speeds up the process but decreases the mAP and F1 scores, which is more noticeable at 25% of the original size. Besides, the best compromise between speed and detection was achieved with the DSFD model with an image reduction of 50%. Comparing the different CPUs, the overall best time was achieved by i9-8650HK, which excels at its bus speed[1]. GPU results, however, overcome the CPU ones, and we noticed that the GPU architecture determines the detection speed. RTX GPUs with a Turing architecture performed better than GPUs with Pascal —GTX and TITAN— and Keppler –Tesla– architectures, despite that these GPUs disposed of a large number of cores, memory video or memory bandwidth.

V. CONCLUSIONS

This work evaluated the speed-performance trade-off among three face detector models –MTCNN, PyramidBox and DSFD– on five Intel CPUs and seven Nvidia GPU cards. The images used in the experimentation came from two datasets, WIDER Face and UFDD, and to improve the inference speed we used a resizing strategy at different scales. We found that the speed-up from resizing was more noticeable on more sophisticated detectors. Thus, the best speed-accuracy trade-off was yielded by DSFD on images resized to 50% of the original size on GPUs of Turing architecture. This benchmark could be used to generate accurate models to predict the performance of future pieces of hardware.

[1]Specifications consulted from https://cpu.userbenchmark.com/

ACKNOWLEDGEMENTS

This research was funded with support from the European Commission under the 4NSEEK project with Grant Agreement 821966. This publication shows the authors' point of view, and the European Commission cannot be held responsible for any use which may be made of the information contained therein. This work was also supported by the framework agreement between the University of León and INCIBE (Spanish National Cybersecurity Institute) under Addendum 01. We acknowledge NVIDIA Corporation with the donation of the TITAN Xp and Tesla K40 GPUs used for this research.

REFERENCES

[1] D. Chaves, E. Fidalgo, E. Alegre, R. Alaiz-Rodríguez, F. Jáñez-Martino, and G. Azzopardi, "Assessment and estimation of face detection performance based on deep learning for forensic applications," *Sensors*, vol. 20, no. 16, p. 4491, 2020.

[2] M. W. Al-Nabki, E. Fidalgo, E. Alegre, and I. de Paz, "Classifying illegal activities on Tor network based on web textual contents," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 35–43.

[3] D. Chaves, E. Fidalgo, E. Alegre, and P. Blanco, "Improving speed-accuracy trade-off in face detectors for forensic tools by image resizing," *V Jornadas Nacionales de Investigación en Ciberseguridad*, 2019.

[4] D. Chaves, S. Saikia, L. Fernandez-Robles, E. Alegre, and M. Trujillo, "A systematic review on object localisation methods in images," *Revista Iberoamericana de Automática e Informática Industrial*, vol. 15, no. 3, pp. 231–242, 2018.

[5] D. Chaves, E. Fidalgo, E. Alegre, F. Jáñez-Martino, and J. Velasco-Mata, "CPU vs GPU performance of deep learning based face detectors using resized images in forensic applications," in *Proceedings of the ICDP-2019*, 2019, pp. 93–98.

[6] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[7] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: A context-assisted single shot face detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 797–813.

[8] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "DSFD: dual shot face detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5060–5069.

[9] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "WIDER Face: A face detection benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5525–5533.

[10] H. Nada, V. A. Sindagi, H. Zhang, and V. M. Patel, "Pushing the limits of unconstrained face detection: a challenge dataset and baseline results," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–10.

[11] R. Nardelli, Z. Dall, and S. Skevoulis, "Comparing Tensorflow deep learning performance and experiences using CPUs via local PCs and cloud solutions," in *Future of Information and Communication Conference*. Springer, 2019, pp. 118–130.

Sesión de Investigación A4: Análisis forense y cibercrimen
http://doi.org/10.18239/jornadas_2021.34.29

A Review of "Bringing Order to Approximate Matching: Classification and Attacks on Similarity Digest Algorithms"

Miguel Martín-Pérez
Universidad de Zaragoza, Spain
miguelmartinperez@unizar.es

Ricardo J. Rodríguez
Universidad de Zaragoza, Spain
rjrodriguez@unizar.es

Frank Breitinger
University of Liechtenstein, Liechtenstein
frank.breitinger@uni.li

***Abstract*—Fuzzy hashing or similarity hashing (a.k.a. bytewise approximate matching) converts digital artifacts into an intermediate representation to allow for efficient (fast) identification of similar objects, e.g., for deny-listing. Over the past decade, new algorithms have been developed and released to the digital forensics community. When releasing algorithms (e.g., as part of a scientific article), they are frequently compared with other algorithms to outline the benefits and sometimes also the weaknesses of the proposed approach. However, given the wide variety of algorithms and approaches, it is impossible to provide direct comparisons with all existing algorithms. In this paper, we present the first classification of approximate matching algorithms which allows for an easier description and comparisons. Therefore, we first reviewed existing literature to understand the techniques various algorithms use and to familiarize ourselves with the common terminology. Our findings allowed us to develop a categorization relying heavily on the terminology proposed by NIST SP 800-168. In addition to the categorization, this paper presents an abstract set of attacks against algorithms and why they are feasible. Lastly, we detail the characteristics needed to build robust algorithms to prevent attacks. We believe that this paper helps newcomers, practitioners, and experts alike to better compare algorithms, understand their potential, as well as characteristics and implications they may have on forensic investigations.**

***Index Terms*—Similarity digest algorithm, Approximate matching, Fuzzy hashing, Similarity hashing, Bytewise, Classification scheme**

Tipo de contribución: *Investigación ya publicada en* "Bringing Order to Approximate Matching: Classification and Attacks on Similarity Digest Algorithms," Forensic Science International: Digital Investigation, 2021 [6].

I. Extended abstract

According to NIST SP 800-168, "approximate matching is a promising technology designed to identify similarities between two digital artifacts" [1]. This identification of similarities between two or more artifacts can happen on three different levels of abstraction: *bytewise*, when the comparison relies on the raw sequence of bytes that form the digital artifacts; *syntactic*, when the internal structures of the digital artifacts under analysis are used instead of merely byte sequences; or *semantic*, when the comparison relies on contextual attributes to interpret the digital artifacts and estimate their similarity. Furthermore, algorithms may either compare artifacts directly (e.g., Levenshtein distance or Hamming distance), or they may first convert them into an intermediate representation (e.g., a fingerprint, hash, digest) that can then be compared. This latter case is often referred to as fuzzy hashing or similarity hashing and aims at complementing cryptographic hash functions by allowing for the identification of *similar* objects instead of *completely identical* objects.

In this paper we focus on algorithms/literature that operate on the byte-level[1] and utilize an intermediate representation, i.e., a digest/fingerprint. We define these kinds of algorithms as *similarity digest algorithms* (SDA)[2]. These algorithms gained popularity around 2006 when `ssdeep` was published [2]. Over the years, many more algorithms have been proposed such as `sdhash` [3], `mrsh-v2` [4] or `TLSH` [5], to name a few.

In order to compare algorithms, the community mostly focuses on obvious metrics such as runtime efficiency or precision and recall rates. However, due to the various design decisions researchers and practitioners have made during the development, we argue that a finer granular comparison is necessary as there may be instances where precision and recall are insufficient. For instance, some implementations have difficulties handling extremely small files, while others are susceptible if the difference in file size between two objects is too large (e.g., 5 MiB vs. 5 GiB). Consequently, this paper has the following contributions:

- The first categorization for SDA, allowing the community to better discuss and compare the various existing algorithms. Categorizations are useful for scientific fields, as they allow structuring a domain.
- A comprehensive discussion of the algorithms with respect to the categorization and its implication for practitioners.
- A discussion of the categorization with respect to why these characteristics are important and how practitioners may contribute from it, describing an abstract set of attacks.
- In addition, we also provide insights on the desirable properties to build a robust SDA against attacks.

In order to develop the classification, we have identified the six phases of a SDA. Five of these phases can be grouped in an "artifact processing and digest generation phase", while the other is devoted to digest comparison. Each phase consists of various dimensions and procedures

[1] Inputs/artifacts are treated as a byte stream and are processed without any interpretation of the data.

[2] In this paper, we use the SDA interchangeably as a singular and plural acronym.

Algorithm	Feature generation				Feature Processing		Feature Selection		
	Length	*Support Function*	*Intersection*	*Cardinality*	*Mapping Function*	*Bit Reduction*	*Selection Function*	*Domain*	*Coverage*
dcfldd	Static (512)	None	No	Variable ($L/512$)	Hashing	None (128)	None	(n/a)	Full
Nilsimsa	Static (3)	None	Yes	Variable ($6L$)	Hashing	None (8)	None	(n/a)	Full
ssdeep	Dynamic ($L/64$)	Trigger function	No	Fixed (64)	Hashing	Ratio (6/32)	None	(n/a)	Full
md5bloom	Static (512)	None	No	Variable ($L/512$)	Hashing	Ratio (40/128)	None	(n/a)	Full
MRS hash	Dynamic (234)	Trigger function	No	Variable ($L/234$)	Hashing	Ratio (44/128)	None	(n/a)	Full
SimHash	Static (1)	None	Yes	Variable ($8L$)	Identifier	None (8)	Block matching	Feature	Partial
sdhash	Static (64)	None	Yes	Variable (L)	Hashing	Ratio (55/160)	Minimum probability	Feature	Partial
MRSH-V2	Dynamic (320)	Trigger function	No	Variable ($L/320$)	Hashing	Ratio (55/64)	None	(n/a)	Full
mvHash-B	Static ($20, 50$)	None	Yes	Variable ($8L$)	Encoding	Ratio (1/32)	Block similarity	Feature	Full
TLSH	Static (3)	None	Yes	Variable ($6L$)	Hashing	None (8)	None	(n/a)	Full
saHash	Static (1)	None	Yes	Variable ($4L$)	None	None (8)	None	(n/a)	Full
LZJD	Dynamic ($1 + \log_{256} L$)	Unique	No	Variable ($L/(1 + \log_{256} L)$)	Hashing	None (128)	Minimum value	Processed feature	Partial
FbHash	Static (7)	None	Yes	Variable (L)	Hashing	None (64)	None	(n/a)	Full

Table I
CLASSIFICATION OF SIMILARITY DIGEST ALGORITHMS ACCORDING TO OUR PROPOSED CLASSIFICATION SCHEME (*feature generation*, *feature processing*, AND *feature selection* PHASES).

Algorithm	Digest generation				Feature Deduplication		Digest comparison			
	Digest Size	*Storing Structure*	*Order*	*Requirements*	*Type*	*Occurrence*	*Requirements*	*Output Score*	*Score Trend*	*Space Sensitivity*
dcfldd	Input dependent	Processed feature concatenation	Absolute	None	None	(n/a)	None	Interval	Ascending	Total
Nilsimsa	Fixed	Counter	Processed feature-aware	None	Consecutive	Comparison	Minimum commonality	Interval	Ascending	None
ssdeep	Input dependent with max	Processed feature concatenation	Absolute	Minimum features	Consecutive	Comparison	Minimum commonality, Similar input size	Interval	Ascending	Total
md5bloom	Input dependent	Set concatenation	Set-absolute	None	In-Scope	Generation	None	Interval	Ascending	Partial
MRS hash	Input dependent	Set concatenation	Set-absolute	None	In-Scope	Generation	None	Interval	Ascending	Partial
SimHash	Fixed	Counter	None	None	None	(n/a)	Similar input size	Half-bounded	Descending	None
sdhash	Input dependent	Set concatenation	Set-absolute	Diversity	In-Scope	Generation	Minimum amount	Interval	Ascending	Partial
MRSH-V2	Input dependent	Set concatenation	Set-absolute	None	In-Scope	Generation	Minimum amount	Interval	Ascending	Partial
mvHash-B	Input dependent	Set concatenation	Set-absolute	Diversity	In-Scope	Generation	Similar input size	Interval	Ascending	Partial
TLSH	Fixed	Counter	Processed feature-aware	None	None	(n/a)	None	Half-bounded	Descending	None
saHash	Fixed	Counter	Processed feature-aware	None	None	(n/a)	None	Binary value	(n/a)	Total
LZJD	Fixed	Set	None	None	None	(n/a)	None	Interval	Ascending	None
FbHash	Fixed	Counter	Processed feature-aware	Document frequency	None	(n/a)	None	Interval	Ascending	None

Table II
CLASSIFICATION OF SIMILARITY DIGEST ALGORITHMS ACCORDING TO OUR PROPOSED CLASSIFICATION SCHEME (*digest generation*, *feature deduplication*, AND *digest comparison* PHASES).

which themselves are based on characteristics (i.e., its values). For instance, the *feature generation phase* has, among other dimensions/procedures, *length*, *support function* and *intersection*. Every dimension can have different characteristics, such as *static* or *dynamic* (length) and *trigger function* or *unique* (support function). The classification of SDA according to our proposed classification scheme is given in Tables I and II.

Regarding the set of attacks against SDA, we have distinguished between two types of attacks:

- **Attacks against the similarity score**, which divides into *Reduction of Similarity* attacks (the input is crafted to minimize the similarity score when it is compared against other input) and *Emulation of Similarity* attacks (the artifact is manipulated to yield a high similarity score to a another [non-similar] artifact).
- **Attacks against impeding the last phases of an SDA**, which splits into *Impeding the Digest Generation Phase* attack (a deliberate modification of the input in such a way that the SDA is unable to generate a similarity digest due to insufficient conditions) and *Impeding the Digest Comparison Phase* attack (an adversary crafts an input so that the similarity digest generated cannot be compared or the similarity score computed is always very low – i.e., no similar).

As the last contribution of the paper, we have highlighted the properties needed to build a robust SDA against these attacks. For the sake of space, we have deliberately omitted a more detailed description of these properties in this paper. Further details are given in [6].

The full version of this paper (with a full description of the classification dimensions, as well as a detailed explanation on attacks and on building a robust SDA) was published in [6].

ACKNOWLEDGEMENTS

The research was supported in part by the Spanish Ministry of Science, Innovation and Universities under grant MEDRESE-RTI2018-098543-B-I00 and by the University, Industry and Innovation Department of the Aragonese Government under *Programa de Proyectos Estratégicos de Grupos de Investigación* (DisCo research group, ref. T21-20R). It was also supported by the Spanish National Cybersecurity Institute (INCIBE) "Ayudas para la excelencia de los equipos de investigación avanzada en ciberseguridad", grant numbers INCIBEC-2015-02486 and INCIBEI-2015-27300.

REFERENCES

[1] F. Breitinger, B. Guttman, M. McCarrin, V. Roussev, and D. White, "Approximate Matching: Definition and Terminology," National Institute of Standards and Technology, techreport NIST Special Publication 800-168, May 2014.

[2] J. Kornblum, "Identifying almost identical files using context triggered piecewise hashing," *Digital Investigation*, vol. 3, pp. 91–97, 2006, the Proceedings of the 6th Annual Digital Forensic Research Workshop (DFRWS '06).

[3] V. Roussev, "Data Fingerprinting with Similarity Digests," in *Advances in Digital Forensics VI*, K.-P. Chow and S. Shenoi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 207–226.

[4] F. Breitinger and H. Baier, "Similarity Preserving Hashing: Eligible Properties and a New Algorithm MRSH-v2," in *Digital Forensics and Cyber Crime*, M. Rogers and K. C. Seigfried-Spellar, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 167–182.

[5] J. Oliver, C. Cheng, and Y. Chen, "TLSH – A Locality Sensitive Hash," in *2013 Fourth Cybercrime and Trustworthy Computing Workshop*. IEEE, 2013, pp. 7–13.

[6] M. Martín-Pérez, R. J. Rodríguez, and F. Breitinger, "Bringing Order to Approximate Matching: Classification and Attacks on Similarity Digest Algorithms," *Forensic Science International: Digital Investigation*, 2021.

Sesión de Investigación A5:
Ciberseguridad industrial y aplicaciones

http://doi.org/10.18239/jornadas_2021.34.30

Diseño de un IDS basado en anomalías para IoT: caso de estudio en SmartCities

Rafael Estepa
Universidad de Sevilla
Email: rafa@us.es
https://orcid.org/0000-0001-8505-1920

Antonio Estepa
Universidad de Sevilla
Email: aestepa@us.es
https://orcid.org/0000-0003-1841-3973

Jesús Díaz-Verdejo
Universidad de Granada
Email: jedv@ugr.es
https://orcid.org/0000-0002-8424-9932

Agustín W. Lara
Universidad de Sevilla
Email: alarar@us.es
https://orcid.org/0000-0003-4809-5654

Germán Madinabeitia
Universidad de Sevilla
Email: german@us.es
https://orcid.org/0000-0001-6376-4620

José A. Morales Sánchez
Wellness Techgroup
Email: jamorales@wellnesstg.com
https://orcid.org/0000-0002-6668-8667

Resumen- **Los sistemas de Smart-City constituyen un campo específico en el IoT. Las soluciones de ciberseguridad IT tradicionales son excesivamente genéricas y poco eficientes para este tipo de instalaciones con escasos recursos computacionales y de coste limitado. Por ello, en conjunción con una empresa del sector, se está desarrollando un proyecto para la detección de incidentes de seguridad de un sistema de Iluminación Inteligente. En este artículo se describen los resultados iniciales del proyecto.**

***Index Terms*- Ciberseguridad en IoT, Ciberseguridad Smart Cities, Sistema Detección Anomalías (A-IDS)**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

La ciberseguridad en IoT es un subcampo de aplicación dentro de los sistemas de control industrial (ICS) cuyas principales particularidades según [1] son: (I) pilas de protocolos diferentes a las empleadas en IT, (II) dispositivos con recursos limitados y (III) tráfico de aplicación con alta periodicidad. Los ataques en entornos ICS también presentan peculiaridades, siendo en ocasiones dirigidos y difíciles de detectar. Por ello, el sistema de bastionado suele utilizar técnicas de detección de anomalías en IoT [2]. Dado que las comunicaciones en cada escenario tienen un patrón de comportamiento particular, las soluciones de seguridad ICS presentan diferencias con las soluciones tradicionales del mundo IT, estando adaptadas a los distintos escenarios (plantas de gas [3], *Smart Grid* [4], tratamiento de aguas[5], IoT[6], etc.).

En este trabajo nos centraremos en el caso particular de un sistema de iluminación inteligente de Smart-City. Se diseñará una solución de seguridad para este entorno que proporcione un nivel de protección adecuado. Los resultados mostrados en el presente trabajo se corresponden con los primeros avances de un proyecto en curso con la empresa *Wellness Techgroup*, propietaria del sistema de *Smart Lighting* implantado en más de 10 países y usado en este artículo.

II. Sistema IoT de *Smart Lighting*

En esta sección presentamos las líneas generales del funcionamiento del sistema de iluminación inteligente, así como sus principales amenazas.

A. Sistema de Smart Lighting

Los elementos que conforman este sistema son: (I) luminarias inteligentes, (II) controladores de luminarias, (III) servidor de aplicación IoT (COAP/MQTT) y (IV) sistema de gestión. En la Fig. 1 podemos apreciar la disposición de estos elementos. El controlador de luminarias alimenta a las luminarias en los momentos apropiados, que pueden depender del orto/ocaso, del nivel de luminosidad, etc. Las luminarias pueden ser también configuradas para atender a eventos como la detección de personas, luminosidad, etc. El sistema sigue las pilas de protocolos IoT y la aplicación M2M utiliza como protocolos COAP o MQTT, según el caso. La comunicación entre controlador y servidor usa la red NB-IoT de un operador o una red LPWA (p.ej., Lora o SigFox), según las circunstancias del despliegue. El servidor COAP o MQTT se encarga de volcar los datos recibidos en una base de datos *Time Series (TSdB)* y de hacer llegar las consignas o comandos a los controladores y luminarias. El sistema se completa con una aplicación de gestión que permite monitorizar y configurar todos los elementos. La comunicación con los servidores COAP / MQTT se realiza sobre una VPN, y entre el gestor del sistema y la aplicación de gestión utiliza HTTPS sobre TLS1.2 con la apropiada autenticación. Los dispositivos envían al servidor COAP / MQTT cada 300 s un resumen de los eventos y consumos detectados.

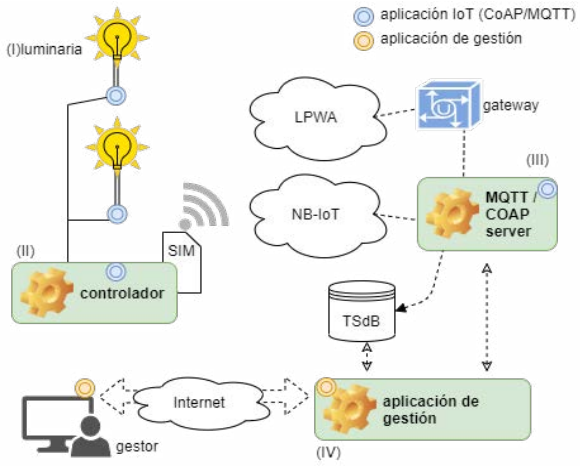


Fig. 1. Esquema simplificado del sistema de *Smart Lighting*.

B. *Principales amenazas en el sistema de Smart Lighting*

En este trabajo nos centramos en tres tipos de ciberamenazas:

A1. Ataques contra el servidor COAP/MQTT. Estos pueden llevarse a cabo mediante la incautación o compromiso de una luminaria o controlador, que pasaría a estar dentro de la red privada y a disposición del atacante, lo que permitiría realizar acciones tales como: envenenar la TSdB, DoS del servidor, etc.

A2. Robo de credenciales del gestor del sistema, lo que permitiría configuraciones maliciosas, como apagar todas las luminarias durante la noche.

A3. Ataques contra el servidor de la aplicación de gestión, que pudieran derivar en la toma de control del servidor con las mismas consecuencias que la amenaza A2.

Sería deseable también incorporar a las amenazas anteriores circunstancias anómalas como el robo de energía eléctrica, sabotaje de un equipo, errores de software, etc., que degradan el rendimiento del sistema.

C. *Condicionantes y requisitos del sistema de detección de anomalías*

El sistema propuesto será compatible con el uso de sistemas de detección de intrusiones (IDS) convencionales del mundo IT basados en firmas (p.ej., Suricata, WAF, etc.). Dado que la detección de amenazas A3 está suficientemente cubierta con este tipo de herramientas y que una de sus principales consecuencias sería la amenaza A2. Por ello, el sistema a desarrollar en este trabajo se centrará en la detección de A1, A2 y el resto de las circunstancias que afectan al rendimiento del sistema.

Los requisitos de la solución propuesta son: (R1) detectar patrones de ataques basados en anomalías, (R2) no interferir el normal funcionamiento del sistema (pasivo), y (R3) bajo consumo de recursos computacionales para no afectar al coste del despliegue de la solución.

III. Sistema propuesto

El sistema propuesto se refleja en la Fig. 2, y consta de tres módulos: (I) Detección de anomalías a nivel de flujos de tráfico, (II) Detección de anomalías a nivel de aplicación M2M y (III) Sistema de correlación de eventos y generación de alertas priorizadas. Dado el elevado número de sistemas luminaria/controlador, la escasez de recursos computacionales, y el tráfico generado por estos, se considera más eficiente la co-ubicación de los sistemas de protección junto a los sistemas de monitorización y gestión existentes. A continuación, se esboza cada uno de los módulos que componen el sistema.

A. *Detector de anomalías en flujos de datos IPFIX*

Este módulo proporciona protección contra la amenaza A1 y se ubica en los servidores (CoAP/MQTT o web), o bien sobre un dispositivo independiente que reciba una copia del tráfico de los servidores (empleando puertos espejos en los conmutadores). Su operación es periódica, realizando secuencialmente los siguientes pasos:

1) Se captura el tráfico (formato *pcap*) durante *T* segundos (por defecto 5 minutos).
2) Generación de una traza IPFIX del tráfico capturado mediante la herramienta *Tranalyzer*, compilada para generar una traza que incluya información sobre: protocolo ARP, periodicidad de los paquetes y DPI (*p0f*).
3) Creación de un resumen de los flujos IPFIX en forma de *matriz de tráfico* con los estadísticos promedio, varianza, máximo y mínimo de ciertas características de los flujos entre dos equipos (*IPsrc – IPdst*).
4) Actualización de un conjunto de indicadores de interés a partir de la matriz de tráfico anterior en función de los ataques a detectar (p.ej., número medio de conexiones). Se definen indicadores a nivel global (actúan sobre toda la matriz de tráfico) y a nivel de equipo (con estadísticos sobre el servidor, básicamente). En total se definen unos 25 indicadores de interés.
5) Ejecución de detectores de comportamientos anómalos sobre los indicadores de interés anteriores en función de la naturaleza del evento a detectar y el comportamiento esperado. Los detectores pueden usar diversas técnicas (valor umbral, EWMA, *Isolation Forest* [8], etc.) y tienen asociadas acciones a ejecutar cuando advierten un comportamiento anómalo. Cada detector funciona de manera independiente y, en general, genera un mensaje con el evento detectado.

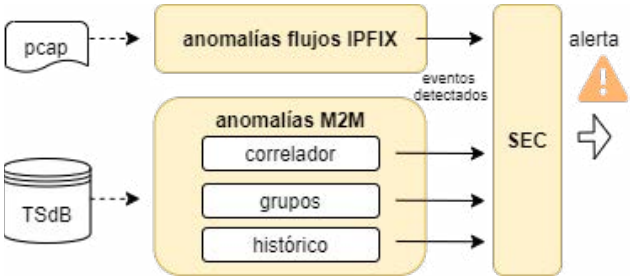


Fig. 2. Diagrama de bloques del sistema.

La matriz de tráfico también es empleada para el mantenimiento automático del inventario de activos y sus propiedades (dirección MAC, sistema operativo, etc.) lo que permite detectar suplantaciones de identidad.

B. *Detección de anomalías a nivel de aplicación M2M*

El sistema utilizará los datos de la TSdB para la detección de anomalías, aprovechando la fuerte periodicidad del modelo de aplicación M2M, muy influenciado por la hora del orto y ocaso (que varían a lo largo del año). Se desea detectar anomalías en consumos, errores de funcionamiento en controladores y luminarias, comportamientos anómalos respecto al histórico temporal u otros equipos similares como reflejo de la amenaza A2, o de otros comportamientos ilícitos (robo de energía, fallos, etc.).

El sistema propuesto está formado por varios detectores:

a) *Correlador*: Cada dispositivo tiene un analizador de red eléctrica con 26 variables. Dado que estas variables responden a un modelo físico, cambios en la correlación entre ellas implican errores en los sensores correspondientes (medidas erróneas). Este módulo identifica grupos de variables correladas dentro de un dispositivo y detecta cambios en la correlación entre las variables que conforman dichos grupos.

b) *Detector de anomalías de grupo*: Tiene como objetivo detectar comportamientos anómalos de un dispositivo frente a otros similares. Toma como entrada una variable por cada grupo de correlación de distintos dispositivos.

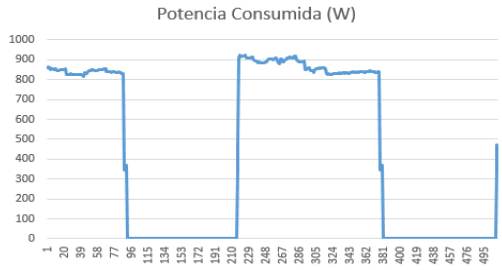


Fig. 3. Valores de las primeras 500 muestras de una variable.

Este sistema realizará una reducción de dimensionalidad empleando *PCA* para luego, con el modelo creado, realizar un *control de residuos Q* [7] con las medidas recibidas y las esperadas según el modelo.

c) *Histórico*: El último sistema detectará anomalías en el histórico de valores de una variable (p.ej., potencia) de un dispositivo. Este módulo está aún en desarrollo.

C. *Correlación de eventos y generación de alarmas*

Este sistema recibe las notificaciones de eventos generadas por los detectores anteriores y genera las alarmas finales. Se basa en el motor de correlación ligero SEC (https://simple-evcorr.github.io/). Algunas situaciones que se deben contemplar en las reglas de correlación incluyen: la validación de que un comportamiento anómalo de aplicación se corresponde un cambio en la configuración y no con un ataque, la búsqueda de la causalidad de la alarma, la correlación entre alarmas, la agrupación de alarmas (para evitar saturar al centro de control) o la priorización de alarmas.

IV. Resultados Preliminares

En el momento de redactar este trabajo tan sólo se dispone del tráfico generado por un controlador durante una semana. Este tráfico consta de un total de 2139 muestras (una cada 300s en media) con 26 variables o características tales como: potencia registrada (real y aparente), consumo de energía, intensidad, voltaje, factor de potencia, etc. En general, los datos exhiben una alta periodicidad, como se puede apreciar en la Fig. 3. Se ha generado un fichero *pcap* con el tráfico con la comunicación COAP correspondiente que se utilizará para alimentar al detector de anomalías en flujos de datos.

A continuación, se presentan las pruebas y avances realizados en los distintos detectores.

A. *Detección de anomalías sobre trazas IPFIX*

Sobre el fichero *pcap* generado se han incorporado una serie de ataques con el fin de validar el funcionamiento del sistema propuesto. Los ataques realizados son:

[T0846] scanningIP (*#nmap –sP*),

[T0841] scanningService *(#nmap –sV serverIP*),

[T0814] Ataque DoS (*#hping3 –p <port> -S –c 2000 –faster <serverIP>*),

[T0830] ARP *spoofing* (*#arpsppof –i eth0 –t <IP> <serverIP>*).

Los códigos asignados a los ataques anteriores se corresponden con las técnicas descritas en la matriz ATT&CK de Mitre. A dichos ataques hay que sumar dos situaciones anómalas a detectar: la inclusión de un nuevo dispositivo en la red (nueva dirección IP) y la caída de un dispositivo. Para las pruebas se han usado los indicadores y detectores recogidos en la Tabla I. Algunos de estos indicadores se calculan sobre el equipo servidor —p.ej. SERVER— y otros sobre la matriz de tráfico completa —p.ej. TOTAL—. Los detectores de anomalías utilizados son: MM=media móvil, VAL=valor, EWMA=*Exponentially Weighted moving average*, INCR = incremento.

Tabla I
Detectores y ataques esperados

ID	Indicador	Detector	Ataques
1	MAX (#FLOWs_ FROM/TO_i)	MM = 0	Robo o malfuncionamiento (*Device Down*)
2	MAX (#IP_ FROM_i_TO_!SERVER)	VAL>0	Posible equipo infectado (comun_not_allowed)
3	TOTAL_ARP_REQ	EWMA	Fase de reconocimiento de ataque (scanning IP)
4	Mean (#dst_port _TO_ SERVER)	EWMA	Fase de reconocimiento (scanning service)
5	TOTAL (MAX(#TCPSYN))	EWMA	Denegación de Servicio (DoS)
6	#ACTIVOS	INCR	Nuevo dispositivo no reconocido en red
7	#CAMBIOS_ ACTIVOS	VAL>0	Intento suplantación MAC o S.O. (MiM)

Tabla II
Resultados detector anomalías en flujos

Ataque	Descripción	Alarmas (detectorID)
T0846	scanningIP	2,3
T0841	scanningServ	4
T0814	DoS	4,5
T0830	MiM	7
New IP	Nuevo equipo	6
Down	Equipo caído	1_i

La Tabla II refleja las alarmas generadas por los detectores correspondientes tras las pruebas. Se puede observar que todos los ataques han sido detectados con la implementación de 7 indicadores. El sistema usado para la realización de las pruebas es un Xeon CPU E5-2620v3@2.40GHz, con 2GB de RAM, y un tráfico de 865 KB (tráfico de 1000 dispositivos en 300 s). Utiliza 530 MB de RAM y tarda 16,2 segundos en realizar el procesamiento, de los que dedica el 55% del tiempo consumido a generar la traza IPFIX y el 42% al cálculo de los indicadores para los detectores. Estos datos avalan que el sistema implementado puede emplearse tanto en máquinas con escasos recursos, tipo *Raspberry Pi*, como en procesos de un servidor.

B. *Detección de anomalías a nivel de aplicación M2M*

Este sistema consultará directamente la base de datos de series temporales (TSdB), por lo que podrá ubicarse en cualquier equipo con acceso a la misma. Dada la limitación que supone para las pruebas tener muestras correspondientes a un sólo controlador, las pruebas se han centrado en:

a) detectar grupos de variables fuertemente correladas a fin de determinar posibles fallos de un sensor en un dispositivo, e identificar las variables a monitorizar,

b) verificar la detección de valores anómalos de potencia consumida frente al comportamiento habitual (anomalías de contexto).

Con respecto a la correlación automatizada de variables, el sistema se ha entrenado con las primeras 288 muestras (1 día de observación), encontrándose una correlación fuerte (más de 0,95) en los siguientes grupos de variables: (G1) intensidades y potencias en las tres líneas, (G2) factor de potencia en las tres líneas, (G3) voltaje en las tres líneas. La Fig. 4 muestra el cambio en la correlación entre potencia e intensidad detectados

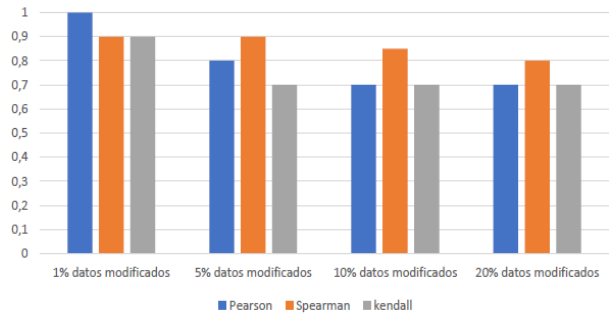


Fig. 4. Correlación para distintos coeficientes y tasas de contaminación.

Tabla III
RESULTADOS DETECTOR ANOMALÍAS EN FLUJOS

#Comp	Max (Q)	E[Q]	$Q_{límite}$ (95%)	False +	True +
1	37.2	5.0	79	0	0
2	11.2	1.7	32	0	0
3	7.0	0.03	6.3	3	0
4	4.3	0.01	4.4	0	0
5	3.9	0.08	3.4	0	2
> 6	<3.8	<0.06	<2.6	0	2

con una ventana deslizante de 24 horas cuando introducimos una contaminación controlada. Observamos que en función del tipo de coeficiente de correlación empleado (*Pearson*, *Spearman* o *Kendall*), somos capaces de detectar cambios cuando entre el 1% y el 5% de las muestras se cambian por un valor aleatorio entre el máximo y mínimo registrado en la fase de entrenamiento. Si aplicamos el algoritmo a una ventana deslizante de 24 h., permitiría detectar fallos de sensores asociados a variables en 72 minutos.

El algoritmo de detección de anomalías de contexto utiliza una aproximación PCA con un control del residuo Q. En dicha aproximación, con los datos en entrenamiento (288 del primer día) se ajusta el modelo mediante combinaciones lineales (componentes principales) de las variables de entrada. Una vez seleccionado el número de componentes, *l*, el residuo, Q, de cada nuevo vector de entrada *Xs* debe ser inferior a un límite Q_{limite} a partir del cual se considera *Xs* un vector anómalo conforme al modelo entrenado. En la Tabla III se muestra el resultado obtenido para distinto número de componentes, el residuo máximo, medio y Q_{limite} para una significación estadística del 95% cuando se han contaminado 2 de las 1851 muestras restantes. Podemos ver que con 3 componentes tenemos 3 falsos positivos (el 0,1% de las muestras) y a partir de 5 componentes no se detecta ningún falso positivo y sí se detectan las 2 muestras contaminadas (el 100% de las muestras). La Fig. 5 muestra la evolución de Q en la ventana temporal de evaluación (segundo día en adelante) cuando hay una muestra contaminada (en rojo Q_{limite}). Estos resultados preliminares permiten presagiar un buen comportamiento del detector ante valores anómalos provocados por las amenazas A2 y A3.

V. CONCLUSIONES

En este trabajo se presentan los resultados preliminares de un proyecto dedicado al diseño de detectores de anomalías en IoT, dentro del campo de las *Smart Cities*, concretamente en *Smart Lighting*. Se ha diseñado una solución de fácil implementación e implantación y bajo coste que cumple con los requisitos de diseño. Los datos de campo utilizados han

Fig. 5. Evolución temporal del residuo Q con 1 muestra contaminada.

permitido realizar una exploración inicial del sistema propuesto. En breve se recibirán los datos de 85 equipos de una instalación completa lo que permitirá ampliar los estudios y diseños realizados y la parte de correlación e inteligencia, que han quedado fuera del alcance del presente trabajo exploratorio.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el proyecto 2020/00000172 dentro del programa de Proyectos singulares de actuaciones singulares de transferencia en los CEI en las áreas RIS3 de la Junta de Andalucía.

REFERENCIAS

[1] Bhamare, D., Zolanvari, M., Erbad, A., Jain, R., Khan, K., & Meskin, N. "Cybersecurity for industrial control systems: A survey". *Computers &Security*, 89, 101677, 2020.

[2] Andrew A. Cook , Göksel Mısırlı, and Zhong Fan: " Anomaly Detection for IoT Time-Series Data: A Survey", en *IEEE Internet of Things Journal, 7(7)*, 6481-6494, 2020.

[3] Beaver, Justin M., Raymond C. Borges-Hink, and Mark A. Buckner. "An evaluation of machine learning methods to detect malicious SCADA communications." 2013 12th international conference on machine learning and applications. Vol. 2. IEEE, 2013

[4] Sridhar, et.al. (2014). Model-based attack detection and mitigation for automatic generation control. IEEE Transactions on Smart Grid, 5(2), 580–591.

[5] Hadžiosmanović, Dina, et al. "Through the eye of the PLC: semantic security monitoring for industrial processes." Proceedings of the 30th Annual Computer Security Applications Conference. 2014

[6] Sandor, H., Genge, B., & Szanto, Z. (2017). Sensor data validation and abnormal behavior detection in the internet of things. 16th Networking in Education and Research RoEduNet International Conference, RoEduNet 2017 – Proceedings

[7] Fouzi Harrou, Farid Kadri, Sondes Chaabane, Christian Tahon, Ying Sun: "Improved principal component analysis for anomaly detection: Application to an emergency department", en *Computers & Industrial Engineering, 88*, 63 - 77, 2015.

[8] S. Hariri, M. C. Kind and R. J. Brunner, "Extended Isolation Forest," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4, pp. 1479-1489, 1 April 2021, doi: 10.1109/TKDE.2019.2947676.

Sesión de Investigación A5: Ciberseguridad industrial y aplicaciones
http://doi.org/10.18239/jornadas_2021.34.31

A Review of MADICS: A Methodology for Anomaly Detection in Industrial Control Systems

Ángel Luis Perales Gómez*[1], Lorenzo Fernández Maimó[1], Alberto Huertas Celdrán[2] and Félix J. García Clemente[1]
[1] Faculty of Computer Science, University of Murcia, 30100 Murcia, Spain
angelluis.perales@um.es; lfmaimo@um.es; fgarcia@um.es
[2]Communication Systems Group CSG, Department of Informatics IfI, University of Zurich UZH, CH-8050, Switzerland
huertas@ifi.uzh.ch

***Abstract*—Diverse cyberattack detection systems have been proposed over the years in the context of Industrial Control Systems (ICS). However, the lack of standard methodologies to detect cyberattacks in industrial scenarios prevents researchers from accurately comparing proposals and results. In this work, we present MADICS, a methodology to detect cyberattacks in industrial scenarios that intends to be a guideline for future works in the field. In order to validate MADICS, we used the popular SWaT dataset, which was collected from a fully operational water treatment plant. The experiments showed that following MADICS, we achieved state-of-the-art precision of 0.984, as well as a recall of 0.750 and F1-score of 0.851, above the average of other works, proofing that the proposed methodology is suitable to be used in real industrial scenarios.**

***Index Terms*—anomaly detection, critical infrastructures, deep learning, industrial control systems, machine learning.**

Tipo de contribución: *Investigación ya publicada en Journal of Symmetry, artículo: "MADICS: A Methodology for Anomaly Detection in Industrial Control Systems".*

I. Introduction

Industrial Control Systems (ICS) are the core of critical infrastructures that provide essential services such as water, power, or communications, among others. ICS support basic services for monitoring and controlling industrial processes. In addition, ICS are complex systems whose networks transport sensors and actuators information together with multimedia data such as images and videos. Thus, it is crucial to adopt security mechanisms to protect these information assets from malicious activities and cyberthreats.

Due to the increment of cyberattacks in industrial scenarios and its high specialization, the research community is moving toward the Anomaly Detection (AD) paradigm using Machine Learning (ML) and Deep Learning (DL) techniques. Although many AD methodologies based on ML/DL have been proposed in different fields, industrial scenarios have several considerations that need to be taken into account in order to detect anomalies successfully.

The current major challenge is the lack of a common methodology in the AD field that establishes guidelines to detect cyberattacks specialized in ICS scenarios. The existing solutions apply different steps without standard criteria resulting in hardly comparable results. The lack of this common methodology may even lead to the results obtained by the researchers being invalid because of wrong steps carried out.

This paper is a review of [1], where we proposed MADICS to overcome the aforementioned challenges. MADICS is a complete methodology based on semi-supervised ML and DL anomaly detection techniques for the ICS field. This methodology is intended to be a step-by-step guide that any researcher can apply to detect cyberattacks in ICS scenarios, aiming to provide a common and unified way to compare their results.

II. Proposed Methodology

The MADICS methodology is divided into five steps: 1) dataset preprocessing, 2) feature filtering, 3) feature extraction, 4) anomaly detection method selection, and 5) validation.

Dataset preprocessing. The first task is to divide the dataset into training, validation, and test. The second task is to explore the dataset to check if the data contain spurious or corrupted values. This type of values is frequent during transitory states or warm-up processes, that is, when the ICS is initializing or after cyberattack is performed. We suggest independently plotting each feature against time and visually identifying corrupted values to remove the corresponding samples. The third task is dedicated to preparing categorical features ready to be used with ML and DL models. At this point, our suggestion is to encode categorical features present in the dataset by applying One-Hot Encoding (OHE). The fourth and final task consists of scaling the continuous features.

Feature Filtering. MADICS proposes several techniques to study the features of the dataset. The goals of these techniques are two, and they are divided into three tasks: to determine if data leakage is present in the dataset (task 1), and to remove features that do not change in the entire dataset (task 2) as well as those whose statistical distributions in the training, validation and test dataset differ significantly (task 3).

Feature Extraction. This step is in charge of extracting higher-order features from the original features, considering that industrial and critical processes frequently perform repetitive actions. It is divided into two tasks: the calculation of the Autocorrelation and Discrete Fourier Transform (DFT).

Anomaly Detection Method. In this step, all the tasks needed to select, fine-tune, and train the most appropriate model to detect anomalies in industrial scenarios are detailed. The first task consists in selecting a proper model. In particular, MADICS suggests a DL regressor model applied to a feature window because 1) it can model more complex behavior than ML techniques, and 2) it is suitable to model time-series data. In the second task, MADICS establishes guidelines to select the proper hyper-parameters range. The

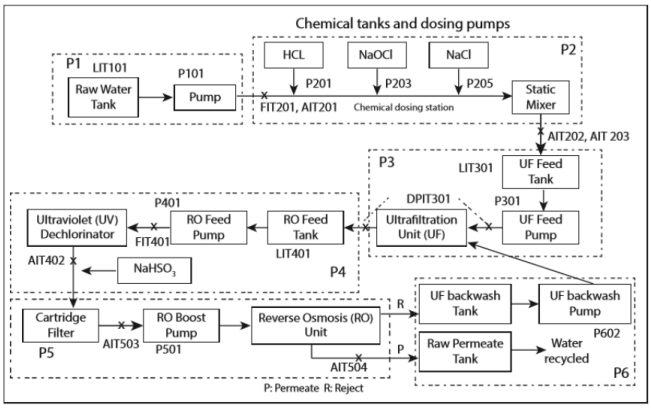


Figure 1. SWaT testbed and its processes

recommendations to select a strategy to search those hyper-parameters are detailed in task 4. Finally, in task 4, MADICS proposes z-score to compute the threshold.

Validation. In the first task, MADICS recommends defining the metrics to validate our model considering the nature of the dataset. When we work with a highly imbalanced dataset, True Positive, True Negative, False Positive, and False Negative are the preferred metrics. However, to facilitate comparison with other works, MADICS uses the precision, recall, and F1-score metrics. Finally, the second task is to check if the results obtained are appropriate for our problem.

III. Methodology Implementation

We evaluated the effectiveness of the proposed methodology using a well-known dataset. The dataset, named SWaT, was captured from a fully operational scaled-down water treatment plant producing 5 gallons/minute of doubly filtered water. As shown in Figure 1, the SWaT testbed consists of 6 industrial processes (labeled from P1 to P6) that work together to treat water for its distribution.

Dataset preprocessing. We started from sensors and actuators logs of the SWaT dataset that are stored in two files: one containing only normal samples and another with both normal and anomalous samples. Following the first task, our training dataset was selected from the first 80% samples of the first file and the validation dataset included the 20% remaining samples. The second file was our test dataset. In the second task, we analyzed the values of each feature, concluding that the first records of the dataset corresponded to the ICS warm-up process. In addition, we removed the next three minutes of normal behavior after each cyberattack because we considered that they were prematurely labeled as normal behavior. As the third and fourth tasks, we used OHE to encode categorical features and scaled the continuous features using standardization.

Feature Filtering. In the first task, we carried out a covariance study to ensure there was no high correlation between the features and the label. In the second task, we carried out a variance study of each feature to determine those features that did not suffer a change in the dataset. Furthermore, in the third task, we used the Kolmogorov-Smirnov test to check if the features from both datasets came from the same distribution.

Feature Extraction. To extract higher-order features from the original features, as suggested by the first and second tasks, we applied autocorrelation and DFT using a 120-second window and then, we computed five statistical measures (mean, standard deviation, minimum, maximum, and range) for each operation. These new features were added to the dataset, resulting in 704 features: 54 features from the original dataset and 650 new autocorrelation/DFT features.

Anomaly Detection Method. As a first task, an LSTM neural network was selected to capture the temporal pattern and predict the future state based on the previous and current state. In the second task, we selected a set of hyper-parameters, and their respective ranges of values, to be fine-tuned during the training process. The hyper-parameter fine-tuning was carried out using a grid search over the possible values in the third task. Finally, in the fourth step, we computed a z-score over the validation dataset, and the maximum value was selected as our threshold.

Validation. In the first step, we chose precision, recall, and F1-score to validate our implementation. Our experimental results achieved the best result in terms of precision. Additionally, our recall and F1-score were above most of the other works as shown by Table I.

Table I
Result comparison of different works using SWaT Dataset

Work	Precision	Recall	F1-score
1D CNN (kravchik et al., 2018)	0.968	0.791	0.871
MLP (Shalyga et al,m 2018)	0.967	0.696	0.812
CNN (Shalyga et al., 2018)	0.952	0.702	0.808
RNN (Shalyga et al., 2018)	0.936	0.692	0.796
LSTM (Zizzo et al., 2019)	-	-	0.817
DNN (Inoue et al., 2017)	0.982	0.678	0.802
OCSVM (Inoue et al,m 2017)	0.925	0.699	0.796
AE Frequency (Kravchik et al., 2019)	0.924	0.827	0.873
DIF (Elnour et al., 2020)	0.935	0.835	0.882
GAN (li et al., 2018)	0.700	0.954	0.810
Ours	0.984	0.750	0.851

IV. Conclusion

In this work, we have presented MADICS, a methodology to detect cyberattacks in ICS scenarios based on semi-supervised ML and DL algorithms and the AD paradigm. This methodology includes five steps: dataset preprocessing, feature filtering, feature extraction, selection, fine-tuning, training of the anomaly detection method, and model validation. The validation of the methodology was performed using the SWaT dataset. In the experiment carried out to test our methodology, the resulting model achieved a state-of-the-art precision score (0.984), whereas both recall (0.750) and F1-score (0.851) were above the average of the rest of the relevant works in the literature.

Acknowledgements

This work has been funded by Spanish Ministry of Science, Innovation and Universities, State Research Agency, FEDER funds, under Grant RTI2018-095855-B-I00, and by the Swiss Federal Office for Defence Procurement (armasuisse) (project code CYD-C-2020003).

References

[1] Á. L. Perales Gómez, L. Fernández Maimó, A. Huertas Celdrán, and F. J. García Clemente, "Madics: A methodology for anomaly detection in industrial control systems," *Symmetry*, vol. 12, no. 10, p. 1583, 2020.

http://doi.org/10.18239/jornadas_2021.34.32

CloudWall: A Cloud-enabled Resiliency Framework for HealthCare IT Infrastructures

Ivan Marsa-Maestre[1], Jose Manuel Gimenez-Guzman[1], Luis Cruz-Piris[1], Susel Fernandez-Melian[1], Jose-Javier Martínez-Herraiz[2], Bernardo Alarcos[1] and Iván Blanco-Chacón[3]
[1]Departamento de Automática, Universidad de Alcalá
[2]Departamento de Ciencias de la Computación, Universidad de Alcalá
[3]Departamento de Física y Matemáticas, Universidad de Alcalá
Edificio Politécnico, Campus Universitario, 28805 Alcalá de Henares (Madrid), Spain
{ivan.marsa,josem.gimenez,luis.cruz,susel.fernandez,josej.martinez,bernardo.alarcos,ivan.blancoc}@uah.es
ORCID IDs: 0000-0002-{5529-2851,1645-8476,9570-2851,1576-4340,2351-7163,4455-5716,4666-019X}

Abstract- **In this paper we present CloudWall, a Cloud-enabled Resilience Framework specially tailored for the needs of the healthcare IT infrastructures. This framework will increase the capability of these infrastructures to prevent and react to cyber attacks. Starting from a risk analysis model based on multi-layer networks, CloudWall will allow to design and implement infrastructure deployments which are more resilient to cyber threats, and specially to malware spread, enforcing good practices in critical infrastructure IT design, such as isolated attack paths and hardware/software diversity in component deployment. The same risk analysis model will help to enhance threat detection using state-of-the-art technologies, like Security Information and Event Management Systems (SIEM). Finally, by deploying components and network elements in a virtualized private-cloud infrastructure when possible, it will leverage recent advances in container-based computing and network function virtualization (NFV) to allow the infrastructure (including both computation and networking assets) to reconfigure itself in the event of a security incident so that the risk of further damage is mitigated.**

Index Terms- reactive resilience, network reconfiguration, security in healthcare IT infrastructures

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

Cyber-attacks over Health Care IT infrastructures have proliferated greatly in the last years [1], since these infrastructures are more connected than ever. Moreover, the heterogeneity of these networks is also increasing radically. Within a healthcare network, we can find nowadays true Internet of Things (IoT) ecosystems [2], comprised of laptops, smartphones and specific medical equipment (drug administration devices, glucometers, blood pressure measurement devices, cardiac pacemakers and defibrillators) accessing or generating medical data. This combination of heterogenous devices together with the desired ubiquity and high connectivity [3] entails a complex challenge from the cybersecurity point of view, and it is becoming even more challenging as the actual trend towards *Bring Your Own Device* (BYOD) initiatives increases.

The impact from cyber attacks over healthcare infrastructures can be severe [4]. Ransomware attacks over the last four years to different hospitals all over the world have repeatedly forced the cancellation of new inpatient admissions, respiratory therapy, radiology exams and procedures and even some surgeries. The frequency of such attacks is increasing alarmingly, and their apparition is globally widespread. Just to cite a few in the last year we have seen successful attacks in Spain (January 2020), in the US (October 2020) and more recently in France (March 2020). And the threat is not only reduced to directed attacks to healthcare. The Wannacry attack was not targeted to health care systems (it was a generic *cryptoworm* which affected Windows machines), but its effects on healthcare were dramatic nonetheless: at least 80 out of 236 hospital trusts across England were affected, 595 GP were disrupted, ate least 1220 diagnostic/treatment devices were disabled and around 19,000 appointments were cancelled. A report published by the government estimates that Wannacry caused approximately £92M of lost output and IT costs, including restoring systems and data in the weeks after the attack. The potential for dramatic consequences increases if we put into the equation medical implants as connected devices [5].

Taking this into account, if we want to keep relying on the increasingly complex and heterogeneous healthcare IT infrastructures for our society wellbeing, new mechanisms should be devised to increase their resilience [6]. Cloudwall aims to increase the capability of health care IT infrastructures for preventing and reacting to cyber attacks. We do so with an approach based on multilayer, graph-based risk analysis and distributed, dynamic network reconfiguration. This approach requires to bring together results and techniques from a variety of disciplines, and to advance the state of the art in some of them. First, we are building in our expertise on graph-based characterization of complex systems, and on distributed optimization of the behaviour of such systems via centralized and distributed techniques. We are particularizing the models for the problem of cyber risk assessment in healthcare scenarios, and are researching new centralized and distributed techniques aimed to reduce this risk in both a preventive and reactive manner. These techniques focus on reconfiguring (reshaping) the IT infrastructure either to make it more resilient (prevention) or to respond to the realization of a particular threat (reaction). Finally, we intend to tackle the technological challenges to apply these techniques to real healthcare IT

scenarios, which involves having a dynamic computing and network deployment framework and to secure the processes regarding the interactions within such framework.

The rest of the paper is structured as follows. Next section reviews some of the most relevant works and efforts around the aforementioned areas. Section III describes the CloudWall infrastructure. Finally, Section IV discusses its potential impact.

II. Related Work

A. Resilient network design and reactive redeployment

The design and deployment of resilient networks and IT infrastructures have been widely studied [7, 8]. However, resilience in healthcare IT infrastructures have not attracted much specific interest from the research community. This is a significant research gap, since these infrastructures have some specific features that make generic models not directly applicable. In addition, most proposals related to cybersecurity in healthcare systems assume that the network is already deployed, so resilient network design is not addressed. In other critical infrastructure domains, there are some promising works from which we are building on, such as [9] and [10]. This latter, along with [11], suggest the use of careful segmentation and component diversity to difficult attack propagation, but they focus on the static design of the network, which allows for preventive risk mitigation, but not reaction. To the best of our knowledge, we are pioneers in proposing reactive network reconfiguration (reshaping) for resilience in critical network infrastructures [12, 13]. In CloudWall we are building upon our previous results to provide an infrastructure for preventive and reactive resilience in the healthcare sector.

B. Multilayer network reshaping

Our reshaping techniques take as a starting point a multilayer network model of the healthcare IT infrastructure, its assets and its risks. This allows us to build on all the advances on complex network theory, including the pioneering contribution [14], where authors analyze the robustness of complex networks to localized attacks, to determine how much damage a network can sustain before it collapses. The differences between these efforts and our own are clear, as these papers are more theoretical, and we consider more realistic computer network models, based on multi-layer graphs [15, 13]. These models have a twofold objective, as they let us to capture the specific features of healthcare networks and also to generate large-scale testing scenarios prior to test our proposals with a smaller-scale prototype testbed to demonstrate the readiness of the technologies and the solution. Regarding the kind of techniques we are researching for network reshaping, we are relying on the families of centralized and distributed techniques that we have successfully exploited in our previous work. This includes centralized optimization hiper-heuristics such as simulated annealing [16] or coral reef optimization [17] and distributed techniques such as belief propagation [Mar17] and nonlinear negotiation [18, 19].

C. Implementation of the reshaping framework.

As we mentioned above, CloudWall central paradigm is network reshaping in reaction (or prevention) to security incidents. This network reshaping, in the end, involves the reconfiguration of network and computation assets as the result of the system optimization. Practical implementation of such reconfigurations requires to address two main technological challenges: first, to ensure that we have an underlying physical network and computational infrastructure which enables the kind of flexibility that we need to dynamically deploy the configurations resulting from the reshaping techniques; and second, to address the potential points of attack inherent to this underlying physical network.

To provide the required flexibility to the infrastructure, we will rely on technologies which are already contributing to the Infrastructure as Code (IaC) paradigm [20], such as Software Design Networking (SDN), Network Function Virtualization (NFV) and container technologies. These are technologies that already enable to have dynamic and flexible deployments. SDN, for instance, through the OpenFlow southbbound API [21], allows to repurpose and reconfigure network elements, thus changing network topologies in real time. If we pair this with container technologies such as Docker [22] and orchestration systems like Kubernetes [23], we have the potential to dynamically re-instantiate components with different configurations and in different places in the network as needed.

Of course, such a flexible network and computational infrastructure is not exempt of drawbacks, and especially regarding security considerations [22]. Typical SDN or NFV infrastructures rely on a centralized controller, and this could be a critical point of failure should this centralized controller be compromised during an attack [24]. Although this central point of failure issue can be mitigated by means of distributed approaches, this raises the issue of how to detect and counter rogue nodes within the network. In this line, we are building upon some of the latest developments in the security literature [25], along with our own work on handling behaviour deviations in complex networks for other domains [18,16,19].

Finally, many of the reshaping techniques we are considering (and specially the distributed ones) rely to some extent on message passing between nodes. We need to ensure that these messages are protected in terms of, among others, authentication, integrity and confidentiality. We are researching on communication schemes based on homomorphic encryption to address these issues [26,27,28].

III. The CloudWall Resilience Framework

A. Healthcare IT risk model based on multilayer networks

We are developing a formal representation that accurately describes healthcare IT infrastructures and the risks of their different assets. This model, which takes advantage of recently developed formulations on multilayer networks [15], will be useful both to capture the features of real healthcare IT infrastructures and to systematically generate environments for testing. This model builds upon the preliminary multilayer risk models that we developed for conceptual critical IT infrastructures [13]. Figure 1 shows an example of the model, which defines four layers. The asset layer, on top, is extracted from risk analysis, and captures the relative importance of the assets and their interdependencies. The bottom layer is the infrastructure layer, which represents the actual infrastructure elements (hosts, links and network appliances) upon which the assets are deployed within the network. Between these two end layers, the component layer represents the components (e.g. databases, backend or frontend elements) upon which assets

depend on to provide their functionality, while the instance layer represents the actual instantiation of these components with particular configurations and degrees of redundancy.

In the final model, this representation is being augmented and refined to capture the particularities of the different healthcare IT infrastructures (e.g. a hospital with its different operational units and their associated assets). A knowledge-based representation will complete the model to facilitate its understanding and use by experts and stakeholders. Also, the latest trends in human-operated ransomware targeting the healthcare sector have motivated us to add an Authentication Layer (greatly inspired by the graph model in [29]) to represent the trust/control relationship between different elements in Active Directory, which are the most frequent path taken by attackers to move laterally to compromise healthcare infrastructures at the moment.

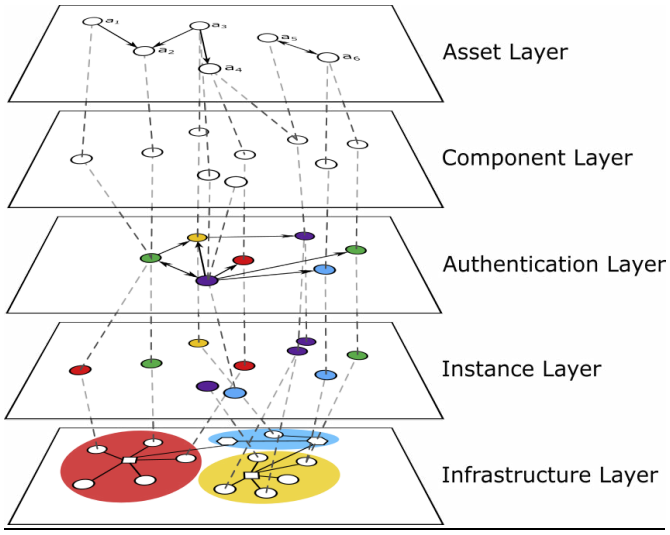


Fig. 1. Multilayer network model

B. *Resilient healthcare IT configurations*

The model developed, along with carefully-designed, extensive experimentation, is allowing us to gain insights into the particularities of risk assessment in healthcare IT infrastructures. From these insights, we are researching on techniques which allow, given a specific scenario and its associated multilayer risk model, to derive a network configuration which minimizes cyber risks. These resilient network configurations take into account the known attack patterns, and specially malware propagation patterns. For instance, one of the most usual propagation techniques in malware involve moving from machine to machine using a remote code execution vulnerability (RCE), such as the ones addressed in MS17-010 Windows patch, one of which was used by WannaCry to spread. Network topologies which are segmented make this propagation harder. In addition, network diversity can be used to further difficult malware spread [11]. For instance, in the WannaCry case, malware could not spread through Unix machines. For assets and components which allow for different configurations (e.g. a web server or a database), a deployment which takes into account configuration diversity could enhance resilience. We are researching mechanisms to derive these optimal network and component configurations from the risk model. This involves a *reshaping* of the multilayer network into a more resilient one.

In this process we are taking a twofold path. On one hand, we try to achieve IT configurations which minimize overall risk at the design stage, which will help decision makers to build more resilient networks in a preventive manner. On the other hand, we are researching on reactive network reshaping in the event of a security incident, that is, when a risk has already materialized and the goal is to minimize damage. This reactive reshaping will difficult attack propagation and will help to avoid the usual cascading failures that occur after cyber attacks by isolating the compromised components or assets.

C. *Technological challenges for the implementation of the solutions*

The CloudWall infrastructure will provide a number of reshaping techniques tailored to protect healthcare IT infrastructures against cyber attacks. In this section we briefly discuss the implementation issues of these techniques in current and near-future network environments. This is of paramount importance to guarantee that the framework can be useful in real environments.

First, to be able to have real dynamic topologies and configurations, we must deploy our infrastructures on virtualized environments, as legacy physical networks do not let us to configure them dynamically with the proper response times required in cybersecurity. For that reason, these infrastructures are the base of the underlying physical network.

Second, as we are preparing IT infrastructures against cyber attacks, the operation of our reshaping infrastructure must be also robust and resilient. For example, if the operation of the reshaping technique is based on some kind of information exchange between nodes, we must guarantee the security of this exchange in terms of, among others, authentication, integrity and confidentiality. We must also take into account the effect of compromised nodes in the operation of the reshaping infrastructure, since the CloudWall scenario involves the reaction to security incidents. Therefore, we are researching and developing mechanisms for the resilience of the reshaping framework itself. We are following an iterative approach, considering initially the scenarios where a small fraction of virtual components has been compromised, and building up from them to finish studying the effect of compromises of physical assets.

Finally, we will explore the particularities and limitations of different implementation strategies. Our exploration will cover three main stages. A first stage will assume physical, on premises devices in the healthcare center, which is closer to the concept shown in our previous works. A second stage will assume a significant number of the assets are in a private data center, which is the actual setting in most public healthcare IT infrastructures in Spain. At a third explore, we will explore the idea of applying this approach to deployments in a public cloud (e.g. Azure or AWS), with the advantages and limitations of such an approach regarding reliability and control (e.g. different “control jurisdictions” for cloud provider and tenant).

IV. Expected Results

The impact of the CloudWall framework is expected to be very significant, both scientifically and for society. Regarding scientific impact, we will advance the state of the art in a number of disciplines. The large scale and complexity of

healthcare infrastructures will challenge current metrics in complex network theory and in cybersecurity risk assessment [11]. Also, the distributed reshaping techniques in response to attacks will require collective automated decision making in a time scale currently not explored, which will impact the fields of collective intelligence and crowd agent deliberation.

CloudWall will also have a significant societal contribution. By providing a novel, multi-layer risk model specifically tailored for healthcare infrastructures, Cloudwall will contribute to enhance the cybersecurity situational awareness of the entire healthcare sector, allowing for better decision making at all levels. The centralized and distributed mechanisms to derive new infrastructure configurations which reduce risks will enable the design of more resilient critical infrastructures, and also the timely reaction to security incidents while minimizing the impact of breaches. Although our work is specifically focused on healthcare scenarios and malware infections, we expect the results to be generalizable to other threat actors such as APTs or targeted attacks, and also to other critical infrastructures outside the healthcare sector. This has the potential for a twofold economic impact: first, by ensuring the availability of critical services such as healthcare, and second, by reducing the recovery costs resulting from cyber attacks, like the aforementioned £92M from Wannacry in the UK health system.

ACKNOWLEDGEMENTS

All authors are funded by Project PID2019-104855RB-I00/AEI/10.13039/501100011033 of the Spanish Ministry of Science and Innovation. Ivan Marsa-Maestre, Jose Manuel Gimenez-Guzman, Luis Cruz Piris, Susel Fernandez-Melian and Iván Blanco-Chacón are also partially supported by Project UCeNet (CM/JIN/2019-031) of the Comunidad de Madrid and University of Alcalá. Ivan Marsa-Maestre, Jose Manuel Gimenez-Guzman, Luis Cruz Piris and Susel Fernandez-Melian are also partially supported by Project SBPLY/19/180501/000171 of the Junta de Comunidades de Castilla-La Mancha. Susel Fernandez-Melian is also supported by Project ESCUDO (CCG20/IA-041) of the University of Alcala. Jose-Javier Martinez-Herraiz is also supported by the European Union's Horizon 2020 Research and innovation program, under grant agreement No. 826284 (ProTego). Bernardo Alarcos is also supported by Project RTI2018-101962-B-I00 of the Spanish Ministry of Science and Innovation. Iván Blanco-Chacón is also supported by Project MTM2016-79400-P of the Spanish Ministry of Science and Innovation and by CCG20/IA-057 of the University of Alcala.

REFERENCES

[1] Safavi, S., Meer, A. M., Melanie, E. K. J., & Shukur, Z. (2018, November). Cyber Vulnerabilities on Smart Healthcare, Review and Solutions. In 2018 Cyber Resilience Conference (CRC) (pp. 1-5). IEEE.

[2] D.V. Dimitrov, Medical internet of things and big data in healthcare, Healthcare Inf. Res. 22 (2016) 156–163.

[3] T. Walker, Interoperability a must for hospitals, but it comes with risks, Manag. Healthc. Exec. (2017) http://managedhealthcareexecutive.modernmedicine.com/

[4] Oz, H., Aris, A., Levi, A., & Selcuk Uluagac, A. (2021). A Survey on Ransomware: Evolution, Taxonomy, and Defense Solutions. *arXiv e-prints*, arXiv-2102.

[5] Zhang, G., Zhang, G., Yang, W., Valli, C., Shankaran, R., & Orgun, M. (2018). From WannaCry to WannaDie: Security trade-offs and design for implantable medical devices.

[6] Smart, W. (2018). Lessons learned review of the WannaCry ransomware cyber attack. *London: Skipton House*.

[7] Sterbenz, J. P., Hutchison, D., Çetinkaya, E. K., Jabbar, A., Rohrer, J. P., Schöller, M., & Smith, P. (2010). Resilience and survivability in communication networks: Strategies, principles, and survey of disciplines. Computer Networks, 54(8), 1245-1265.

[8] Smith, P., Hutchison, D., Sterbenz, J. P., Schöller, M., Fessi, A., Karaliopoulos, M., ... & Plattner, B. (2011). Network resilience: a systematic approach. IEEE Communications Magazine, 49(7), 88-97.

[9] Kreutz, D., Malichevskyy, O., Feitosa, E., Cunha, H., da Rosa Righi, R., & de Macedo, D. D. J. (2016). A cyber-resilient architecture for critical security services. Journal of Network and Computer Applications, 63, 173-189.

[10] Zhang, M., Wang, L., Jajodia, S., Singhal, A., & Albanese, M. (2016). Network diversity: A security metric for evaluating the resilience of networks against zero-day attacks. IEEE Transactions on Information Forensics and Security, 11(5), 1071-1086.

[11] Li, T., Feng, C., & Hankin, C. (2018). Improving ICS Cyber Resilience through Optimal Diversification of Network Resources. *arXiv preprint arXiv:1811.00142*.

[12] Enrique de la Hoz, José Manuel Giménez-Guzmán, Iván Marsá-Maestre, Luis Cruz-Piris and David Orden, "Distributed, Multi-Agent Approach to Reactive Network Resilience," 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS '17).

[13] Iván Marsá-Maestre, José Manuel Giménez-Guzmán, David Orden, Enrique de la Hoz and Mark Klein, "REACT: REActive resilience for critical infrastructures using graph-Coloring Techniques. Journal of Network and Computer Applications," 2019.

[14] Shao, S., Huang, X., Stanley, H. E., & Havlin, S. (2015). Percolation of localized attack on complex networks. New Journal of Physics, 17, 023049.

[15] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno y Mason A Porter. "Multilayer networks". Journal of complex networks 2.3 (2014), pp. 203-271.

[16] de La Hoz, E., Marsa-Maestre, I., Gimenez-Guzman, J. M., Orden, D., & Klein, M. (2017, May). Multi-agent nonlinear negotiation for Wi-Fi channel assignment. In Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (pp. 1035-1043).

[17] Camacho-Gómez, C., Marsa-Maestre, I., Gimenez-Guzman, J.M., Salcedo-Sanz, S. "A Coral Reefs Optimization algorithm with substrate layer for robust Wi-Fi channel assignment", Soft Computing, 2019.

[18] Gimenez-Guzman, J. M., Marsa-Maestre, I., Orden, D., de la Hoz, E., & Ito, T. (2018). On the goodness of using orthogonal channels in WLAN IEEE 802.11 in realistic scenarios. Wireless Communications and Mobile Computing, 2018.

[19] Marsa-Maestre, I., de la Hoz, E., Gimenez-Guzman, J. M., Orden, D., & Klein, M. (2019). Nonlinear negotiation approaches for complex-network optimization: a study inspired by Wi-Fi channel assignment. Group Decision and Negotiation, 28(1), 175-196.

[20] Rahman, A., Mahdavi-Hezaveh, R., & Williams, L. (2019). A systematic mapping study of infrastructure as code research. *Information and Software Technology*, *108*, 65-77.

[21] Mirghiasaldin Seyedebrahimi, Faycal Bouhafs, Alessandro Raschellà, Michael Mackay y Qi Shi. "SDN-based channel assignment algorithm for interference management in dense Wi-Fi networks". IEEE European Conference on Networks and Communications, 2016 (EuCNC). 2016, pp. 128-132.

[22] Martin, Antony, Simone Raponi, Théo Combe, and Roberto Di Pietro. "Docker ecosystem–vulnerability analysis." *Computer Communications* 122 (2018): 30-43.

[23] Gawel, M., & Zielinski, K. (2019, July). Analysis and Evaluation of Kubernetes Based NFV Management and Orchestration. In *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*. IEEE.

[24] Ahmad, I., Namal, S., Ylianttila, M., & Gurtov, A. (2015). Security in software defined networks: A survey. IEEE Communications Surveys & Tutorials, 17(4), 2317-2346.

[25] Alshehri, M., & Panda, B. (2019, July). An Encryption-Based Approach to Protect Fog Federations from Rogue Nodes. In *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage* (pp. 225-243). Springer, Cham.

[26] Gentry, C., & Boneh, D. (2009). A fully homomorphic encryption scheme (Vol. 20, No. 09). Stanford: Stanford University.

[27] Luo, F., Wang, F., Wang, K., & Chen, K. (2019). Fully homomorphic encryption based on the ring learning with rounding problem. *IET Information Security*

[28] Blanco, I. (2021) RLWE/PLWE equivalence for totally real cyclotomic subextensions via quasi-vandermonde matrices. *Journal of Algebra and its Applications*. In press.

[29] Powell, B. A. (2020). The epidemiology of lateral movement: exposures and countermeasures with network contagion models. *Journal of Cyber Security Technology*, *4*(2), 67-105.

Sesión de Investigación A5: Ciberseguridad industrial y aplicaciones
http://doi.org/10.18239/jornadas_2021.34.33

Towards decentralized and scalable architectures for Access Control Systems in IIoT environments

Santiago Figueroa-Lorenzo[1,2] Saioa Arrizabalaga[1,2] Javier Añorga[1,2]
[1]CEIT-Basque Research and Technology Alliance (BRTA), Manuel Lardizabal 15, 20018 Donostia / San Sebastián, Spain.
[2]Universidad de Navarra, Tecnun, Manuel Lardizabal 13, 20018 Donostia / San Sebastián, Spain.
Email: {sfigueroa, sarrizabalaga, jabenito}@ceit.es

Abstract- **Security is one of the main challenges of Industrial Internet of Things environments, for which, access control has emerged as a solution for mitigating some of their vulnerabilities and threats. This manuscript provides an overview of the research carried out within the last three years in the topic of access control system architectures for IIoT environments, that have derived in seven scientific papers. The research has taken profit of the evolution of blockchain technologies, as well as the level of access control implementation on smart contract. This paper describes the progress of access control systems towards decentralized and scalable architectures in IIoT environments, analyzing, based on scientific contributions of our research, benefits, and limitations of each of proposed architectures.**

***Index Terms*- Access Control Systems, IIoT, Blockchain, SSI**

Contribution type: *Extended abstract*

I. Introduction

The Industrial Internet of Things (IIoT) architecture is complex due to, among other things, the convergence of protocols, standards, and buses. This convergence not only makes interoperability difficult, but also makes security one of the main challenges of IIoT. For this reason, a comprehensive survey of the thirty-three most useful protocols, standards, and buses in IIoT environments was performed based on features such as architecture, topology, protocol, and security [1], which correspond to the contribution (1A) of our roadmap (Fig. 1). From this analysis, we determined the existence of security problems not only in the design or configuration, but also in the assets that support these protocols, standards, and buses. Therefore, it was necessary to perform a comprehensive assessment to measure the risk of existing documented vulnerabilities. Since the CVSS offers a way to collect the main properties of a vulnerability and generate a numbered score that indicates its severity, we design a vulnerability analysis framework (VAF) to process 1363 vulnerabilities corresponding to the thirty-three protocols, standards, and buses to determine that both the severity and impact of a vulnerability are higher if the asset is part of an Operational Technology (OT) environment than of an Information Technology (IT) environment.

Considering RFID as one of the base protocols for us (together with Modbus) and as part of the contribution (1B) of our roadmap (Fig. 1), we analyzed in detail the documented vulnerabilities, security risks and threats for RFID environments, generating proofs of concept for several of them from both open source and open hardware tools such as Proxmark3 [2].

In both contributions, as part of the countermeasures section, access control systems (ACS) have been proposed as a vulnerability mitigation mechanism. However, the effectiveness of these solutions requires a comprehensive analysis that includes, based on experimental results, benefits, and limitations of their implementation in IIoT environments, which constitutes the purpose of this manuscript. The following sections analyze the contributions associated with each and every one of the access control solutions proposed.

II. Access Control Systems

This section describes the different proposed ACS architectures by focusing on their evolution. For each contribution, benefits and limitations achieved from experimental results are included.

A. RBAC on centralized environment

Based on the security recommendations established by the Modbus organization, we have proposed a role based access control (RBAC), in order to authenticate and authorize Modbus IIoT endpoints (contribution 2A in Fig. 1) [3]. The mutual authentication is provided by TLS. The authorization process includes the entity authorization performed through entity roles (RBAC), which are included as arbitrary extension in the X.509v3 certificates. The roles are validated from values stored in a secure database. Comprehensive performance tests demonstrate the feasibility of RBAC in a centralized environment. In this sense, latency measurements when applying the RBAC policy over a secure channel, i.e., Modbus TLS, are compared, using as benchmarks latencies of both Modbus TCP transactions and, latencies of industrial processes standard. Despite these benefits the proposal includes limitations around (1) the lack of flexibility of RBAC compared to, e.g., ABAC; (2) certificate-based identity management and PKI infrastructure, with issues around scalability; (3) centralized execution of RBAC authorization policy; and (4) a single point of failure for role storage.

B. ABAC on public blockchain

Based on the limitations established for II.A, contribution 3A (Fig. 1) focuses on the development of an ABAC access control system over an IIoT-RFID environment, which is more flexible and scalable than RBAC [4]. To solve the centralization issues, the ABAC policy, i.e., authorization, is executed on a decentralized application (DApp), where each of the endpoints running the DApp (RFID Reader) have control of the keys according to the Ethereum (ETH) public network rules, which decentralizes the identity significantly. Proof-of-concepts conducted over the ETH Ropsten test network demonstrated the technical feasibility of the proposal with acceptable latency levels over the test network compared to the latency levels achieved when deploying local ETH nodes: “geth”. However, the performance achieved are not acceptable

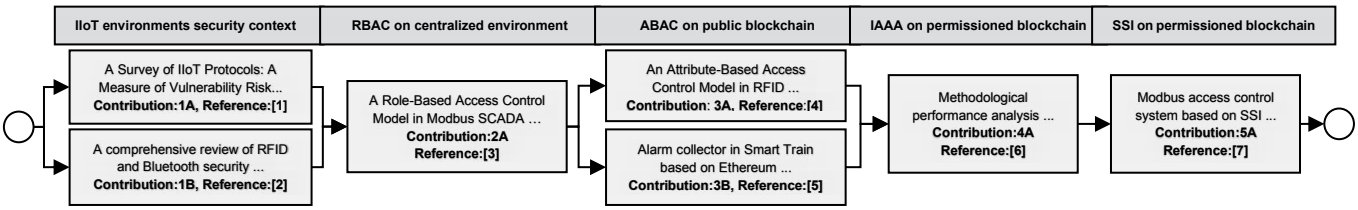


Fig. 1. Access Control Systems roadmap.

for constrained IIoT environments, due to delays introduced by ETH networks. Other limitations of the proposal are, on the one hand, the fact that ETH identity introduces pseudo-anonymization, and on the other hand, the fact that Ethereum identity is not associated for RFID tags. Additionally, the authorization policy is executed off-chain, i.e., on the DApp, to minimize transaction costs in ETH, i.e., the blockchain is used only as verifiable data registry (VDR).

Considering transaction costs as one of the limitations of contribution 3A, contribution 3B (Fig. 1) focuses on providing a mechanism to improve this issue. In this regard, the 3B contribution designs an alarm collection system based on the ETH event-log emission, which integrates both private data collection and alarm collection [5]. The results justify the use of events-log as the most efficient technology in terms of gas costs, as well as determine the capacity of our system to manage a high number of concurrent alarms. However, these results imply maintaining a private data collection (private data local management), i.e., a centralized external storage to the blockchain, so that the smart contract is only used to emit a key-value pair as part events-log, enabling traceability by leaving a mark on the blockchain. Despite this, the proposal is not able to ensure real-time traceability due to the delays introduced by the ETH network, consequently, the performance issues mentioned in 3A are not solved either.

C. *IAAA on permissioned blockchain*

In order to overcome the limitations of contributions 3A and 3B, as part of contribution 4A (Fig. 1) an on-chain ACS was designed on a RFID-IIoT environment that includes all phases of access control: Identification, Authentication, Authorization, Auditing and Accountability [6]. The proposal was designed over a Hyperledger Fabric Blockchain (HFB) network to achieve high performance levels. In this regard, the registration, authentication, and attributed-based authorization (ABAC) phases are on-chain, i.e., blockchain is more than a VDR. From the performance analysis of the HFB network, based on a methodological framework, we have demonstrated the feasibility of the ACS in a performance constrained IIoT environment such as an engine assembly line. In addition, we have designed the registration phase of our ACS based on HFB private data collection, which promotes a novel reliable data privacy model applied to an ACS. Additionally, we have demonstrated the feasibility of using HFB's private data collection over a private data local management, comparable to the one used in contribution 3B, and finally, we have selected the most suitable combination of network elements and resources for an optimal deployment of the HFB network associated to our use case. The most important limitation of the proposal is that identity management is based on HFB, which adopts traditional Public Key Infrastructure (PKI) hierarchical model, which introduces a certain degree of centralization, since all endpoints participating in the network must be registered in the HFB Certificate Authority (CA), which can be a limitation from a scalability point of view.

D. *SSI on permissioned blockchain*

To address the limitations of 4A, the 5A contribution (Fig. 1) design an Self-Sovereign Identity (SSI) system, over Modbus IIoT, that while maintaining the benefits achieved in 4A, solves the identity problems [7]. Therefore, 5A proposes an ACS based on SSI over HFB that, through a decentralized identity system, compliant with both DID standard and verifiable credentials, promotes, at the chaincode level, not only on-chain authentication, and authorization but also advanced operations such as signature verification, i.e., blockchain is again more than a VDR, it is an identity and access management system itself, which includes an ACS. The 5A contribution provides security to Modbus connections and ensures scalability in environments with more than one organization. The latency and throughput levels achieved for a network of up to 32 clients and one server demonstrate the feasibility of the 5A contribution to secure the channel for Modbus transactions while ensuring simplicity, compatibility, and interoperability.

III. CONCLUSIONS

This manuscript summarizes the research carried out about access control architectures in IIoT environments suitable for different scenarios (from traditional scenarios to more complex decentralized scenarios), emphasizing how the adoption of blockchain technology ensures properties not achievable in centralized environments. Smart contracts are the core of the ACS evolution, allowing blockchain to be used from a VDR to an identity management system, which integrates an ACS.

REFERENCES

[1] S. Figueroa-Lorenzo, J. Añorga, and S. Arrizabalaga, "A Survey of IIoT Protocols: A Measure of Vulnerability Risk Analysis Based on CVSS," *ACM Comput. Surv.*, vol. 53, no. 2, 2020, doi: 10.1145/3381038.

[2] S. Figueroa Lorenzo, J. Añorga Benito, P. García Cardarelli, J. Alberdi Garaia, and S. Arrizabalaga Juaristi, "A Comprehensive Review of RFID and Bluetooth Security: Practical Analysis," *Technologies*, vol. 7, no. 1, p. 15, 2019, doi: 10.3390/technologies7010015.

[3] S. Figueroa-Lorenzo, J. Añorga, and S. Arrizabalaga, "A Role-Based Access Control Model in Modbus SCADA Systems. A Centralized Model Approach," *Sensors* , vol. 19, no. 20. 2019, doi: 10.3390/s19204455.

[4] S. Figueroa, J. Añorga, and S. Arrizabalaga, "An Attribute-Based Access Control Model in RFID Systems Based on Blockchain Decentralized Applications for Healthcare Environments," *Computers*, vol. 8, no. 3, 2019, doi: 10.3390/computers8030057.

[5] S. Figueroa-Lorenzo, Santiago; Goya, Jon; Añorga, Javier; Adin, Iñigo; Mendizabal, Jaizki; Arrizabalaga, "Alarm collector in Smart Train based on Ethereum blockchain events-log." 2020, doi: doi.org/10.1109/JIOT.2021.3065631.

[6] S. Figueroa-Lorenzo, J. Añorga, and S. Arrizabalaga, "Methodological performance analysis applied to a novel IIoT access control system based on permissioned blockchain," *Inf. Process. Manag.*, vol. 58, no. 4, p. 102558, 2021, doi: https://doi.org/10.1016/j.ipm.2021.102558.

[7] S. Figueroa-Lorenzo, J. Añorga, and S. Arrizabalaga, "Modbus access control system based on SSI over Hyperledger Fabric Blockchain," *Sensors (Basel, Switzerland)2*, 2021 (*under review*).

Sesión de Investigación A5: Ciberseguridad industrial y aplicaciones
http://doi.org/10.18239/jornadas_2021.34.34

Protocolo de comunicación entre vehículos basado en el número de bastidor digital

Pablo Escapa
Research Institute of Applied
Sciences to Cybersecurity
Universidad de León
ORCID: 0000-0002-8066-0396
pescg@unileon.es

Adriana Suárez Corona
Research Institute of Applied
Sciences to Cybersecurity
Universidad de León
ORCID: 0000-0002-8252-8620
asuac@unileon.es

***Resumen*—Establecer una comunicación vehicular segura es esencial para el desarrollo de la conducción autónoma y de nuevas funcionalidades como la gestión vial inteligente o los nuevos sistemas destinados a evitar o mitigar los accidentes de tráfico. En este escenario, donde puede ser difícil el despliegue de una infraestructura de clave pública, la criptografía basada en identidades presenta una buena alternativa. Este artículo presenta un protocolo de comunicación aplicable entre vehículos V2V (Vehículo a Vehículo) y generalizable a entre éstos y su entorno V2X (Vehículo a todo) sustituyendo la infraestructura de clave pública por esquemas criptográficos basados en identidades usando el número de bastidor en formato digital como identificador inequívoco de los vehículos.**

***Index Terms*—Ciberseguridad, V2V, V2X, comunicaciones vehiculares, criptografía basada en identidades.**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

Desde hace años los vehículos han dejado de estar formados exclusivamente por elementos hidráulicos y mecánicos para dar paso a la introducción de componentes eléctricos y electrónicos, los cuales les hacen dotarse de sistemas ADAS (Advance Driver Assistance Systems – Sistemas Avanzados de Ayudas a la Conducción)[1], así como de conectividad. Estas nuevas características basadas en IoT (Internet of Things - Internet de las Cosas) agregan un valor añadido que mejora la experiencia de los conductores y/o acompañantes e incluso pueden ser tecnologías habilitadoras de la conducción autónoma. En este sentido, se puede considerar a los vehículos como una plataforma hiper sensorizada y conectada, que forma parte del mundo del IoT abriéndose un nuevo escenario cuyos tres principales objetivos son lograr: 0 accidentes, 0 contaminación y 0 atascos.

La conectividad y la seguridad son dos conceptos de aplicación imprescindibles para la llegada del coche conectado y autónomo. Al igual que en otros campos la aparición de internet y la comunicación vehicular supone una gran revolución dado que, gracias a la información compartida entre los vehículos, la infraestructura y el resto de actores, se podrán reducir los 1,35 millones de muertes en todo el mundo causados por los accidentes de tráfico [2] y las cerca de 7 millones causadas por la contaminación [3]. Con estos retos adquiridos, la industria automotriz necesita tanto de la creación de canales de comunicación, como de protocolos que garanticen la seguridad de esas comunicaciones. Debemos recordar la importancia de este hecho dado que la acción de conducir es crítica, y puede causar tanto daños físicos como materiales.

La criptografía basada en identidades, introducida por Shamir [4], proporciona una alternativa de criptografía de clave pública sin necesidad del uso de certificados y de la gestión y actualización de las listas de revocación de éstos (CRL). En este caso, cadenas de caracteres arbitrarias, llamadas identidades, pueden actuar como claves públicas. Éstas deben determinar de forma unívoca a los usuarios del sistema, como el número de pasaporte, la dirección de correo electrónico o el número de la seguridad social. En el contexto de la comunicación entre vehículos, el número de bastidor en formato digital, al ser este único, parece el más apropiado para este fin.

A partir de las identidades, cualquiera puede calcular las claves públicas de los usuarios y el Centro Generador de Claves (KGC), a partir de una identidad y su clave maestra privada, puede calcular la clave privada correspondiente. En el caso de los vehículos, este proceso podría ser responsabilidad de los organismos de normalización internacionales, quienes, a su vez, son los encargados de otorgar los WMI (World Manufacturer Identifier - Identificador Mundial de Constructores) de los números de bastidor tradicionales o cada fabricante, bajo un gestor universal, podría tramitar la generación de las claves, de forma similar a las destinadas a los sistemas de arranque y apertura o codificación. Un ejemplo similar podría ser el procedimiento de generación del SCN (Software Calibration Number) de Mercedes [13], el cual es totalmente telemático e independiente del concesionario o taller autorizado, siendo todo gestionado desde los servidores centrales de la marca.

Numerosos esquemas de cifrado, firma o establecimiento de claves [4], [5], [6], [7], [8] han sido publicados hasta la fecha. Muchos de ellos hacen uso de aplicaciones bilineales [9], y algunos evitan su uso para conseguir protocolos más eficientes [10] o protocolos resistentes a ataques cuánticos [11].

II. Antecedentes

Las comunicaciones vehiculares se pueden clasificar en dos tipos: *internas*, que afectan a las transmisiones entre las diferentes ECU (Unidades de Control), basadas en el protocolo CAN Bus [12], y las *externas*, denominadas V2X (Vehículos conectado a todo) [14], que relacionan a los vehículos con el

resto del entorno bajo conexiones por radiofrecuencia WiFi o 5G, y que están definidas principalmente por los estándares IEEE 802.11P [15] y 5G [16]. Su uso permite la creación de aplicaciones de ITS (Sistemas Inteligentes de Transporte) y habilitan, entre otras, las siguientes conexiones (Figura 1):

- **V2V**: conexiones entre vehículos para, por ejemplo, prevenir colisiones o interactuar entre ellos.
- **V2P**: conexiones entre vehículos y personas para, por ejemplo, evitar atropellos.
- **V2I:** conexiones entre vehículos y las redes viarias o infraestructuras.
- **V2N**: conexiones de vehículos con la nube para enrutar el tráfico.
- **V2H**: conexiones entre vehículos y hogares para implementar, por ejemplo,la apertura de la puerta de casa una vez se aparque.

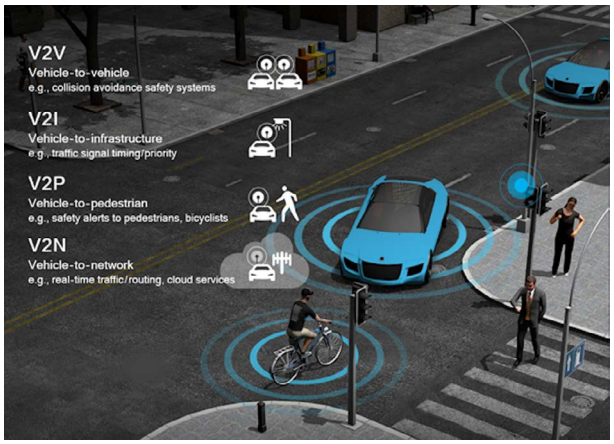


Figura 1. Comunicaciones V2X [17].

II-A. Comunicaciones WiFi vs 5G

Las conexiones vehiculares son un tipo de red emergente cuyas principales tecnologías son: **WiFi** bajo el estándar 802.11P (ver Cuadro I), también denominado WAVE (Wireless Access for the Vehicular Environment), que supone una evolución sobre el estándar IEEE 802.11A, contando con mejoras a nivel físico y de la capa de control de acceso al medio. Su importancia reside en que debería ser la base de las comunicaciones directas V2V (vehículo a vehículo). Este tipo de comunicaciones se conocen como DSRC (Dedicated Short Range Communications) y son ideales para dotar de soporte en tiempo real a los sistemas cooperativos [18].

Cellular C-V2X: Es una tecnología de comunicación de mayor alcance, que aprovecha la cobertura de las redes existentes de telefonía móvil que usan el espectro con bandas licenciadas conectando zonas amplias, siendo utilizadas para servicios V2I en los que el tiempo es menos decisivo (las tecnologías 3G y 4G LTE Release 14). Se unirá la tecnología 5G (3GPP Release 16). Ver cuadro II.

II-B. Comunicaciones V2V seguras usando criptografía basada en identidades

El uso de criptografía basada en identidades para la comunicación intervehicular ha sido considerado previamente [19], [20], así como en el contexto de las VANETS (Vehicle Adhoc Networks) [19] y, en situaciones que requieren privacidad,

Estándar	802.11P
Alcance	De 100 a 1000 metros
Banda	5.9 GHZ
Canales	7 de longitud 10 MHZ
Modulación	OFDM
Tasa de transmisión	De 3, 4, 5, 6, 9, 12, 18, 24 y 25 Mbps en canales de 10 Mhz. Podría llegar a 54 Mbps con canales de 20 Mhz
Sub portadoras	52 bajo BPSK, QPSK, 16-QAM o 64-QAM
Canales especiales	6 de servicio y uno de control

Cuadro I
CARACTERÍSTICAS 802.11P.

Estándar	3GPP 5G
Alcance	Mínimo 1000 metros
Banda en España	700 MHz, 3,6, 26 GHz
Canales	de longitud 100 MHZ por operador
Modulación	OFDM
Tasa de transmisión	100 mbps, 1, 3 o 10 Gbps dependiendo del ancho de banda

Cuadro II
CARACTERÍSTICAS C-V2X.

se han propuesto esquemas donde la autenticación se basa en pseudónimos (identificadores aleatorios) [21] o se utilizan esquemas de firma de grupo [22].

La mayoría de estas propuestas no especifican qué información usar como identidad. En [20] se propone utilizar como identidades las matrículas de los coches, previamente reconocidas a través de cámaras de teléfonos inteligentes disponibles en los vehículos. Sin embargo, la cadena de caracteres que suele usarse en la identificación de vehículos es el número de bastidor dado que es único, a diferencia de la matricula, que puede variar por cambio de país o por propia re-matriculación del vehículo.

III. EL NÚMERO DE BASTIDOR DIGITAL

El número de chasis o bastidor también conocido como VIN (Vehicle Identification Number) es la forma habitual de identificar inequívocamente a un vehículo. Sería el equivalente en vehículos del DNI (Documento Nacional de Identidad) o el pasaporte para las personas. Su formato actual consta de 17 dígitos y está estandarizado desde el año 1983 por la ISO 3779 [23], en la cual se define su composición. Existen diferencias entre el formato europeo y el americano y está formado por números y letras (excepto la I, O, Q y Ñ). Debe ser fácilmente legible, así como estar en un lugar seguro y que sea difícil de manipular: habitualmente se encuentra bajo el capó o en el interior del vehículo troquelado directamente sobre el chasis. Al igual que el número físico, también podemos obtenerlo en formato digital a través de la toma de diagnosis OBD2 que equipan los automóviles bajo el estándar SAE J1979 [24] y su correspondiente europeo ISO 15031 [25]. Esto se puede llevar a cabo mediante la utilización de los PIDs (parameter IDs), más exactamente, dentro del modo 09 con el PID 0x02, esta instrucción ejecutada en cualquier vehículo nos devuelve su número de bastidor tomado directamente de su gateway o ECU en formato ASCII o hexadecimal. La importancia de poder identificar a los automóviles inequívocamente reside en

que sin ello no sería posible establecer algoritmos de gestión del tráfico o coordinación de incidentes, así como implementar nuevas utilidades y experiencias para los conductores y los acompañantes, etc. Esto puede suponer el paso previo e imprescindible para crear la identidad digital de los vehículos.

IV. Protocolo de comunicación entre vehículos usando criptografía basada en identidades

En esta sección se presenta un protocolo genérico de comunicación entre dos vehículos que proporciona confidencialidad, autenticación e integridad. Éste partirá de un esquema de intercambio de clave de una ronda de comunicación con autenticación basada en identidades para el establecimiento de clave de sesión, y, posteriormente, un esquema de cifrado autenticado (AEAD) [27] para intercambiar los mensajes entre los vehículos garantizando la confidencialidad, autenticación e integridad.

Existirá una fase previa de configuración en la que, por un lado se establecen los parámetros públicos del sistema y la clave privada del centro generador de claves (KGC). Éste, a partir del número de bastidor de cada vehículo, VIN_{veh} y su información secreta, calcula la clave privada de cada vehículo $SK_{VIN_{veh}}$, y se la proporciona.

Después se produce el intercambio de mensajes entre los dos vehículos, con autenticación basada en sus números de bastidor digital, añadiendo al mensaje enviado por B un MAC (Message Authentication Code) calculado con la clave común generada sobre el mensaje enviado por A y el intercambio de información termina con el envío del vehículo A al vehículo B de un MAC calculado con la clave común generada sobre todos los mensajes intercambiados, a modo de confirmación de clave, como puede verse en la Figura 2.

Una vez los vehículos han compartido la clave común, pueden comenzar el intercambio de información usando un esquema de cifrado autenticado.

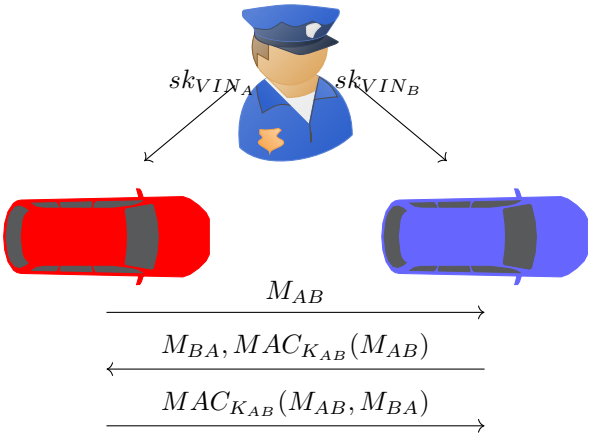


Figura 2. Protocolo de comunicación V2V

Para ejemplificar la aplicación del protocolo, usaremos el esquema propuesto en [10], representado en la Figura 3, donde q es un número primo, G es un grupo de orden q y $g \in \mathbb{G}$, $x, k_A, k_B, t_A, t_B \in \mathbb{Z}_q$, H_1 y H_2 son funciones hash y el par $r_A = g^{K_A}, s_A = k_A + xH_1(VIN_A, r_A)$ son las claves privadas de A, calculadas como una firma de Schnorr [28] de la cadena de caracteres VIN_A (resp. para B).

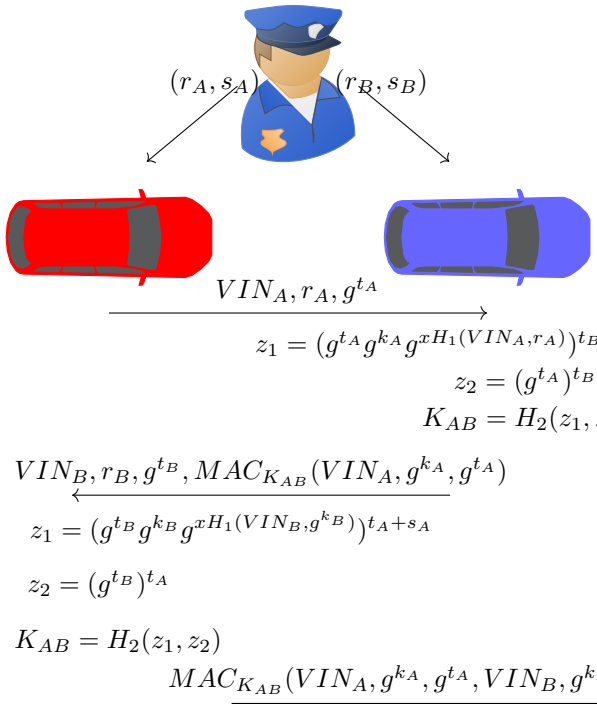


Figura 3. Protocolo de comunicación V2V basado en [10]

V. Aplicaciones

Entre las muchas posibles aplicaciones del protocolo de comunicación propuesto, y su generalización a comunicaciones V2X, se encuentran:

- **Documentación digital:** podría permitir autenticar nuestro vehículo frente a FFyCC (fuerzas y cuerpos de seguridad). Con alguna modificación, se podría añadir la funcionalidad de servir como historial del vehículo incluyendo datos relativos a sus reparaciones, propietarios, revisiones de ITV (Inspección Técnica de Vehículos), etc. Esto funcionaría con la parte privada y se crearía el concepto de identidad digital de un vehículo.
- **Peajes:** funcionará con la zona pública mediante la cual nos identificaremos frente a estas infraestructuras siendo más ágil el pago y no requiriendo de hardware adicional. Sería un sistema universal independiente del gestor de la infraestructura viaria.
- **Método de pago:** nuestra identificación inequívoca puede proporcionar que nuestros automóviles se conviertan en una pasarela de pago, siendo posible su aplicación para gasolineras, servicios de recogida de comida rápida desde el coche, etc.
- **Identificación V2X:** la propia información que contiene el número de bastidor, permite que mediante su decodificación nos aporte las características de un vehículo. Su identificación frente al resto componentes es útil para el intercambio de información. Sus aplicaciones son múltiples: por ejemplo, conocer la longitud de un

vehículo, saber dónde está situado un peatón, crear señalización digital, soluciones para evitar la congestión, etc.

- **Talleres:** en la actualidad el proceso de llevar un vehículo a un a taller es tedioso por el tiempo que se tarda en realizar una orden de reparación. Estos tiempos se podrían acortar, ya que una vez se ingrese un coche, éste será reconocido rápidamente en la zona de admisión, además de comunicar reportes y datos acerca de averías o mantenimientos.
- **Gasolineras**: el sistema evitara que repostemos erróneamente combustible además de agilizar rápidamente su pago.

VI. Conclusiones

El empleo de la criptografía basada en identidades demuestra ser una solución idónea para las comunicaciones vehiculares eliminando la tediosa tarea que supone la creación de certificados de sistemas basados en criptografía de clave pública (PKI). El modelo propuesto, basado en el VIN (Vehicle Identification Number) o bastidor digital, supone un avance logrando identificar a los vehículos inequívocamente al igual que su homólogo en formato tradicional. Esto proporciona y habilita diversas aplicaciones de gran utilidad.

Agradecimientos

Adriana Suárez Corona ha sido parcialmente financiada por el proyecto de investigación MTM2017-83506-C2-2-P, financiado por el Ministerio de Economía y Competitividad.

Referencias

[1] L. González, M. Vaca, R. Lattarulo, I. Calvo, J. Pérez y A. Ruiz. "Análisis de riesgos de ciberseguridad en arquitectura de vehículos automatizados", en *XXXIX Jornadas de Automática*, pp. 5–7, 2018.

[2] OMS. "10 datos sobre la seguridad vial en el mundo", 2017. [Online]. Disponible: http://www.who.int/features/factfiles/roadsafety/es/. Último acceso: Mar. 08, 2021.

[3] OPS/OMS. "OMS estima que 7 millones de muertes ocurren cada año debido a la contaminación atmosférica", 2017. [Online]. Disponible: https://www.paho.org/hq/index.php?option=com_content&view=article&id=9406:2014-7-million-deaths-annually-linked-air-pollution&Itemid=135&lang=es. Último acceso: Mar. 08, 2021.

[4] A. Shamir. " Identity-based cryptosystems and signature schemes", en *Workshop on the theory and application of cryptographic techniques*, vol. 3, pp. 47–53, Springer, 1984.

[5] D. Boneh, M. Franklin. "Identity-Based encryption from the Weil pairing", en *SIAM Journal of Computing*, vol. 32, nº 3, pp. 586—615, 2003.

[6] C. Gentry y A. Silverberg. "Hierarchical ID-based cryptography", en *Advances in Cryptology – ASIACRYPT 2002*, vol. 2501 de LNCS, pp. 548–566, 2002.

[7] R. Sakai, K. Ohgishi y M. Kasahara. "Cryptosystems based on pairing", en *Symposium on Cryptography and Information Security*, Okinawa, Japan, 2000.

[8] C. Boyd, K. R. Choo. "Security of Two-Party Identity-Based Key Agreement", en *Mycrypt 2005*, pp. 229-243, 2005.

[9] L. Chen, Z. Cheng, N.P. Smart. "Identity-based key agreement protocols from pairings", en *International Journal of Information Security*, vol. 6, pp. 213–241, 2007.

[10] D. Fiore, R. Gennaro. " Identity-based key exchange protocols without pairings", en *Trans. Computational Science*, pp. 42–77, 2010.

[11] C. Peng, J. Chen, L. Zhou, K. R. Choo, D. He, " CsiIBS: A post-quantum identity-based signature scheme based on isogenies", en *Journal of Information Security and Applications*, vol. 54, pp.1-13, 2020.

[12] Canbus. "What is CAN Bus?", 2015. [Online]. Disponible: http://www.who.int/features/factfiles/roadsafety/es/. Último acceso: Apr. 25, 2018.

[13] SCN "Mercedes-Benz B2B Connect", 2021. [Online]. Disponible: https://b2bconnect.daimler.com/ES/workshop-solutions/diagnosis/xentry-flash. Último acceso: Apr. 25, 2021.

[14] SNS Telecom & IT. "The V2X (Vehicle-to-Everything) Communications Ecosystem: 2019 – 2030 – Opportunities, Challenges, Strategies & Forecasts", 2019. [Online]. Disponible: https://www.snstelecom.com/v2x. Último acceso: Apr. 25, 2018.

[15] K. Technologies. "Solutions for 802.11p Wireless Access in Vehicular Environments (WAVE) Measurements", 2016. [Online]. Disponible: https://innovation-destination.com/2018/01/29/adas-v2x-trumps-self-driving-2/. Último acceso: Nov. 17, 2019

[16] G. Naik, B. Choudhury, and J. M. Park. "IEEE 802.11bd 5G NR V2X: Evolution of Radio Access Technologies for V2X Communications", en *IEEE Access*, vol. 7, pp. 70169–70184, 2019.

[17] "Qué es C-V2X, la conexión inalámbrica sólo para coches compatible con 5G Motor", en *ComputerHoy.com*, 2019. [Online]. Available:https://computerhoy.com/noticias/motor/c-v2x-conexion-inalambrica-solo-coches-compatible-5g-362505. Último acceso: Mar. 10, 2021.

[18] V. Vibin, P. Sivraj, and V. Vanitha. "Implementation of In-Vehicle and V2V Communication with Basic Safety Message Format",en *Proc. Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2018*, pp. 637–642, 2018.

[19] N. I. Shuhaimi, T. Juhana. "Security in Vehicular Ad-Hoc Network with Identity-Based Cryptography Approach: A survey", en *International Conference on Telecommunication Systems, Services and Applications (TSSA), pp. 276-279*, 2012.

[20] T. Andreica, B. Groza. "Secure V2V Communication with Identity-based Cryptography from License Plate Recognition", en *Sixth International Conference on Internet of Things, Management and Security (IOTSMS)*, pp. 366–373, 2019.

[21] M. A. Al-shareeda, M.Anbar, S. Manickam, I.H. Hasbullah "An Efficient Identity-Based Conditional Privacy-Preserving Authentication Scheme for Secure Communication in a Vehicular Ad Hoc Network", en *Symmetry*, vol. 12, pp. 1687, 2020.

[22] A. Wasef, X. Shen. "Efficient group signature scheme supporting batch verification for securing vehicular networks", en *Proceedings of the 2010 IEEE International Conference on Communications*, pp. 1–5, 2010.

[23] UNE-ISO "3779:2011 Vehículos de carretera. Número de identificación de los vehículos (VIN). Contenido y estructura.", 2017. [Online]. Disponible: https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0047944. Último acceso: Mar. 08, 2021

[24] SAE International, "J1979: E/E Diagnostic Test Modes", 2012. [Online]. Disponible: https://www.sae.org/standards/content/j1979_201202/. Último acceso: Mar. 10, 2021.

[25] ISO "15031-3:2016 - Road vehicles — Communication between vehicle and external equipment for emissions-related diagnostics — Part 3: Diagnostic connector and related electrical circuits: Specification and use", 2016. [Online]. Disponible: https://www.iso.org/standard/64636.html. Último acceso: Mar. 08, 2021

[26] J. Vanier. "App-vin-reader", en *GitHub*, 2016. [Online]. Disponible: https://github.com/carloop/app-vin-reader/blob/master/src/app-vin-reader.cpp. Último acceso: Mar. 08, 2021.

[27] " CAESAR: Competition for Authenticated Encryption: Security, Applicability, and Robustness", 2012. [Online]. Disponible en http://competitions.cr.yp.to/caesar.html. Último acceso: Mar. 10, 2021.

[28] C.P. Schnorr. "Efficient identification and signatures for smart cards", en *Advances in Cryptology—Crypto '89*, vol. 435 de LNCS pp. 239-252, Springer-Verlag, 1990.

Sesión de Investigación A5: Ciberseguridad industrial y aplicaciones
http://doi.org/10.18239/jornadas_2021.34.35

Classifying Screenshots of Industrial Control System Using Transfer Learning and Fine-Tuning

Roberto A. Vasco-Carofilis *†, Pablo Blanco-Medina *†, Franciso Jáñez-Martino *†,
Guru Swaroop Bennabhaktula *‡, Eduardo Fidalgo *†, Alejandro Prieto Castro †, and Víctor Fidalgo †

*Department of Electrical, Systems and Automation, Universidad de León, León, ES
†Researcher at INCIBE (Spanish National Cybersecurity Institute), León, ES
‡Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Groningen, NL
Email:{andres.vasco, pablo.blanco, francisco.janez, eduardo.fidalgo}@unileon.es,
g.s.bennabhaktula@rug.nl, {alejandro.prieto, victor.fidalgo}@incibe.es

***Abstract*—Industrial control systems are heavily dependant on security and monitoring protocols. For this purpose, monitoring tools take screenshots of control panels for later analysis. Classifying these screenshots into specific groups can be a time-consuming process, but it is crucial for the security tasks performed by manual operators. To solve this problem, we propose a pipeline based on deep learning to classify snapshots of industrial control panels into three categories: Internet Technologies (IT), Operation Technologies (OT), and others. We compare the results obtained with transfer learning and fine-tuning on nine convolutional neuronal networks pre-trained with the ImageNet dataset, testing them on a custom CRitical INFrastructure dataset (CRINF-300). Inception-ResNet-V2 obtains the best learning result with an F1-score of 98.32% on CRINF-300, while MobileNet-V1 obtained the best performance-speed trade-off.**

***Index Terms*—Deep Learning, Image Classification; Transfer Learning; Industrial Control System; Fine-tuning**

Type of contribution: *Research already published – Detecting Vulnerabilities in Critical Infrastructures by Classifying Exposed Industrial Control Systems Using Deep Learning. Applied Sciences [1]*

I. Introduction

Critical infrastructures, namely healthcare, transportation or manufacturing, require constant monitoring in environments such as Industrial Control Systems (ICSs). An error on these infrastructures might cause serious consequences, such as equipment failure or information leak [1].

Supervisory Control and Data Acquisition (SCADA) systems are used to control physical equipment and ICS infrastructures. SCADA systems are commonly referred to as Operational Technology (OT) systems, which control and monitor specific devices. Other industrial systems used to control software, including management, and delivery of data, are known as Information Technology (IT) systems [2].

Law Enforcement Agencies (LEAs) use specialized tools known as Internet Metasearch Engines (IMEs) to monitor these exposed assets. For those services that include a graphical interface, IMEs take screenshots to log relevant information graphically. The classification of these screenshots might help to classify the devices, as well as discover vulnerabilities.

In our paper [1], we aimed to classify ICS screenshots automatically into IT and OT categories using deep learning. Our proposal might help monitor critical infrastructures in both performance and computational cost.

II. State of the Art

Image classification is the task of assigning a label to an image. In the last years, Convolutional Neural Networks (CNNs) have been established amongst the best algorithms for this task [3]. Several works have studied the use of transfer learning applied to CNNs in different fields [4]. However, these networks need to be trained on a large amount of data, and data gathering and annotation can be a complex, time-consuming process.

One of the most common problems in this task is the low number of available images for a specific task. To address this, transfer learning is a technique that allows taking a model trained for a specific application and apply it to a closely related task [5]. In cases where images for training a model are scarce, or the classification tasks are challenging, manually-crafted feature extraction can outperform the CNNs [6], [7].

III. Methodology and Experimental Settings

Our proposal for image classification is based on transfer learning and fine-tuning, due to the limited amount of data. Figure 1 depicts the proposed methodology.

For the transfer learning approach, we freeze the final layer from each pre-trained network and obtain features to train a logistic regression model and a Support Vector Machine (SVM) with a linear kernel. We measured their performance using 5-Fold cross-validation.

For the fine-tuning approach, we drop the last layer of each model and replace it with an average 2D pooling of 7x7, a Flatten layer, a Dense layer with 256 neurons and ReLU activation, a 0.5 dropout and an output layer with softmax activation. We used 70% of our images for training, 20% for testing, and 10% for validation. Our training parameters consisted of a batch size of 26 and a learning rate of $1e-4$ for 20 epochs. Finally, we applied data augmentation to our images, increasing their original number up to 5 times.

For training and evaluation, we used 337 ICS snapshots provided by the Spanish National Cybersecurity Institute (INCIBE). We manually labeled them into two categories: 74 IT and 263 OT images, resulting in a dataset we named **Cr**itical **Inf**rastructure (CRINF-300).

We selected nine architectures commonly used in image classification, pre-trained with the Imagenet dataset; Inception-V3 [8], MobileNet-V1, MobileNet-V2 [9],

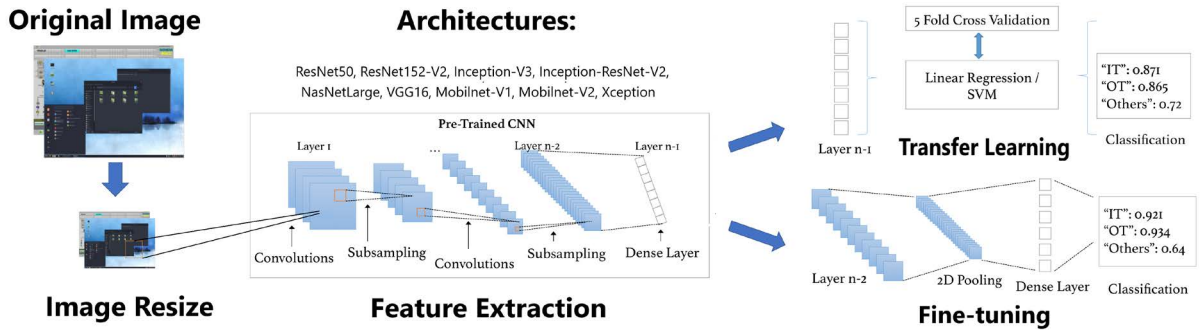


Fig. 1. Proposed pipeline for classifying ICS snapshots into three categories –IT, OT and Others– using transfer learning and fine-tuning

ResNet50, ResNet152v2 [10], VGG16 [11], NasNetLarge [12], Inception-ResNet-V2 [13], and Xception [14].

IV. Experimental Results

The F1-Score-based results of our experiments are presented in Table I. Inception-ResNet-V2 obtained the best results with an F1-score of up to 98.32% using transfer learning and logistic regression. In fine-tuning, the best architecture was VGG16, with an F1-score of 93.73%.

We found that the fastest model was VGG16, with an average of 0.03s to process each image in GPU with both SVM and logistic regression, and 0.14s with fine-tuning. CPU processing achieved the best speed of up to 0.14s with transfer learning and 0.21s with fine-tuning. On the other hand, MobileNet-V1 obtained the best performance-speed balance with a CPU speed of 0.06s and a GPU speed of 0.04s.

Although our fine-tuning additions perform well on top of ResNet50 and VGG16, they do not achieve good performance on the rest of the architectures, obtaining lower results than those of the transfer learning based models.

TABLE I
F1-Score results in the three proposed strategies

Architecture	Logistic Reg.	SVM	Fine-tuning
ResNet50	87.67 (±0.35)	87.67(±0.35)	92.16
VGG16	87.67(±0.35)	87.67(±0.35)	**93.73**
Xception	89.46(±0.67)	90.08(±1.51)	80.98
Inception-V3	97.02(±1.50)	97.20(±1.31)	82.01
Mobilenet-V1	97.58(±0.70)	97.56(±0.75)	86.51
Mobilenet-V2	97.55(±0.49)	97.92(±0.74)	48.78
NasNetLarge	96.44(±1.29)	96.78(±1.04)	81.20
Inception-ResNet-V2	**98.32**(±0.70)	**98.13**(±0.84)	76.22
ResNet152v2	97.41(±0.90)	97.59(±0.94)	71.22

V. Conclusions

This paper presented a pipeline for classifying ICS images as belonging to IT, OT or Others systems, using transfer learning and fine-tuning based approaches on CRINF-300, a custom dataset. We used nine CNN architectures pre-trained on the Imagenet dataset from two perspectives: transfer learning, with a logistic regression classifier and an SVM classifier, and fine-tuning. We chose these techniques due to our short supply of images.

The best CNN architectures to solve the proposed problem are Inception-ResNet-V2 and Mobilenet-V1. The first obtained the best performance with an F1-score of up to 98.32%, while the latter achieved the best performance speed trade-off, with an F1-score of 97.58 and a speed of 0.06s on CPU. We conclude that transfer learning with logistic regression is the best approach, based on the performance of the networks.

Acknowledgements

This work was supported by the framework agreement between the Universidad de León and INCIBE (Spanish National Cybersecurity Institute) under Addendum 01. We acknowledge NVIDIA Corporation with the donation of the TITAN Xp and Tesla K40 GPUs used for this research.

References

[1] P. Blanco-Medina, E. Fidalgo, E. Alegre, R. A. Vasco-Carofilis, F. Jañez-Martino, and V. F. Villar, "Detecting vulnerabilities in critical infrastructures by classifying exposed industrial control systems using deep learning," *Applied Sciences*, vol. 11, no. 1, 2021.

[2] W. A. Conklin, "It vs. ot security: A time to consider a change in cia to include resilienc," in *2016 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE, 2016, pp. 2642–2647.

[3] N. Sharma, V. Jain, and A. Mishra, "An analysis of convolutional neural networks for image classification," *Procedia Computer Science*, vol. 132, pp. 377–384, 2018.

[4] Z. Xiao, Y. Tan, X. Liu, and S. Yang, "Classification method of plug seedlings based on transfer learning," *Applied Sciences*, vol. 9, no. 13, 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/13/2725

[5] M. Hussain, J. J. Bird, and D. R. Faria, "A study on cnn transfer learning for image classification," in *UK Workshop on Computational Intelligence*. Springer, 2018, pp. 191–202.

[6] E. Fidalgo, E. Alegre, V. Gonzalez-Castro, and L. Fernández-Robles, "Boosting image classification through semantic attention filtering strategies," *Pattern Recognition Letters*, vol. 112, pp. 176–183, 2018.

[7] E. Fidalgo, E. Alegre, L. Fernández-Robles, and V. González-Castro, "Classifying suspicious content in tor darknet through semantic attention keypoint filtering," *Digital Investigation*, vol. 30, pp. 12–22, 2019.

[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[12] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

[13] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[14] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

Sesión de Investigación A5: Ciberseguridad industrial y aplicaciones
http://doi.org/10.18239/jornadas_2021.34.36

Retos de ciberseguridad en automoción. Cifrando el bus CAN

Estibaliz Amparan Calonge
TECNALIA, Basque Research and Technology Alliance (BRTA)
Derio
estibaliz.amparan@tecnalia.com

Alejandra Ruiz
TECNALIA, Basque Research and Technology Alliance (BRTA)
Derio
alejandra.ruiz@tecnalia.com

Resumen- **Se pretende profundizar en las necesidades de ciberseguridad en las comunicaciones intra-vehiculares. La tecnología de los vehículos conectados es tendencia, y más recientemente con la entrada de la tecnología 5G para comunicar los vehículos. Se viene innovando en la protección de las nuevas redes de comunicación entre vehículos, no obstante, este no son el único punto vulnerable de los vehículos. La principal red de comunicación interna de los sistemas críticos del vehículo, la conocida como Controller Area Network (CAN) se ha visto vulnerada en distintos ataques- Este documento recoge las vulnerabilidades del bus CAN y la aplicación de técnicas de cifrado para obtener autenticidad e integridad de los datos.**

***Index Terms*- ciberseguridad, CAN bus, automoción, cifrado, integridad, autenticidad**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

En esta última década se está produciendo una nueva revolución llamada la revolución digital en automoción. Esta revolución consiste en la comunicación entre el vehículo y otros sistemas externos a él (vehículos, infraestructuras, dispositivos móviles), abriendo las puertas al cambio de información. La información que se comparte con los demás vehículos permite conocer la situación del tráfico de nuestra ruta ayudándonos a evitar atascos e incluso a prevenir accidentes. Además, con los smartphones se puede conocer y controlar el estado de nuestro vehículo y actuar de forma remota, por ejemplo, iniciando el motor del coche, ajustando la climatización del vehículo y abriéndolos. Incluso se puede manejar el software reproduciendo música o utilizando el navegador GPS. Por desgracia, estos avances tecnológicos generan una alta superficie de ataque que tiene como objetivo el control del vehículo. Estos puntos de comunicación, hacen que un usuario externo sea capaz de llegar a obtener el control del vehículo.

En los años 60 se comenzó a sustituir partes mecánicas del vehículo por otras electrónicas con el objetivo de mejorar la fiabilidad. En los años 80 se produjo uno de los cambios más importantes a destacar, que fue el desarrollo del protocolo CAN bus. Un protocolo de comunicación para el intercambio de información entre unidades de control electrónicas. Hasta hoy, no se había contemplado la seguridad en el medio de comunicación CAN, por encontrarse prácticamente aislado del exterior. Pero el incremento de posibles puntos de conexión y pruebas realizadas con éxito de ataques en el vehículo ha provocado que se empiece a diseñar una arquitectura no solo tolerante a los fallos que se originen por su propia naturaleza sino también a las provocadas por atacantes. La **¡Error! No se encuentra el origen de la referencia.** muestra los protocolos actualmente existentes de comunicación internas de un vehículo cada uno encargado de la transmisión de ciertos datos como por ejemplo el protocolo MOST, que es el encargado de la transmisión de los datos multimedia.

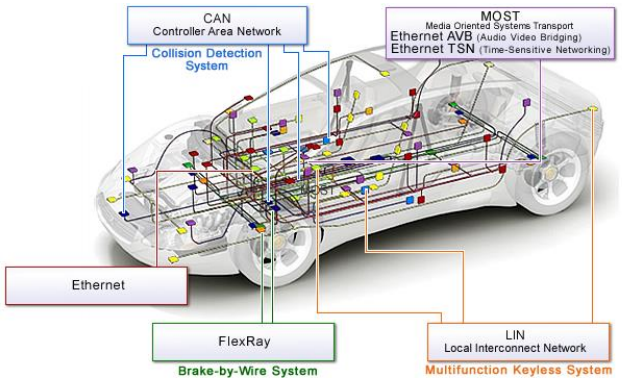


Figura 1: Comunicaciones internas del vehículo. [1]

II. Retos de la seguridad en los vehículos

El tema de la seguridad es uno de los problemas más importantes de los sistemas ciber físicos como lo son por ejemplo los aviones, los trenes y los vehículos. Asegurarse de que van a funcionar correctamente es crucial, de no ser así puede provocar grandes daños físicos en la integridad de las personas. Desgraciadamente la seguridad ante posibles ataques dentro de los automóviles es desconocida tanto por los usuarios como por los fabricantes, donde en muchas ocasiones no se toman las medidas pertinentes.

Uno de los retos más importantes del mundo de la automoción es aplicar la seguridad en los vehículos donde hasta ahora no se le había dado importancia ya que la comunicación de los vehículos con el exterior era mínima. De forma local un atacante puede acceder a la red de comunicaciones del vehículo a través del puerto estándar de diagnóstico OBD II (On Board Diagnostics) y manipularlo. Además, también se pueden realizar ataques a través de USB,

CD y DVD, pero como requieren un acceso físico este se vuelve más inaccesible.

Inevitablemente los vehículos se están volviendo más accesibles de forma remota. La superficie de ataque se extiende de forma exponencial en el momento en el que se añaden nuevos puntos de comunicaciones como, WiFi, bluetooth o redes de telefonía móvil como 4G o 5G. Además, los sistemas de automatización actuales almacenan y procesan datos del perfil de los usuarios, hábitos y comportamientos, como, por ejemplo, datos personales de navegación o datos de info-entretenimiento.

I. Ejemplos de ataques

Es bien conocido el ataque que realizaron y publicaron Charlie Miller y Chris Valasek quienes fueron capaces de conducir remotamente un Jeep in 2015, tomando control del vehículo, conducir por una autovía, dirigir al vehículo hacia una salida y finalmente frenarlo. Aprovecharon la vulnerabilidad del Sistema Uconnect, un Sistema que permite al vehículo conectarse a Internet permitiendo controlar el sistema de entretenimiento del vehículo, el sistema de navegación, hacer llamadas o incluso convertir el vehículo en un hot spot de WiFi. A través de este Sistema los atacantes enviaron comandos al Jepp a través de su sistema de entretenimiento hasta su ordenador central, girar la dirección, frenar o mandar distintas órdenes al sistema de tracción, todo ello remotamente. Como consecuencia, 1.4 millones de coches fueron llamados a revisión. [2]

En [3] se presentan un resumen de una serie de ataques potenciales en entornos cooperativos como en los que se encajan los vehículos autónomos. Petit y Shladover analizaron posibles vulnerabilidades sobre infraestructura en la vía, sensórica, mapas internos del vehículo, comunicaciones internas del vehículo (CAN principalmente) y comunicaciones con otros entes en entornos cooperativos.

Hasta ahora los expertos en seguridad se centraban en aquellas vulnerabilidades que pueden dar acceso remoto a los sistemas del vehículo, no obstante, este paradigma está cambiando con la creciente tendencia a la compartición de coches o vehículos autónomos, escenarios donde distintos usuarios tienen acceso al mismo vehículo y hay que poner especial atención en el modelo de ataque al atacante local.

Uno de los últimos informes publicados [4] menciona que sobre los ataques al puerto OBD ya suman el 10.5% de los ataques. A pesar de que es necesaria un acceso físico al puerto, una vez acceden pueden inyectar mensajes maliciosos al bus CAN, manipulando el comportamiento del vehículo. Además, el 27.62% de los ataques analizados en el informe tenían como objetivo el control no autorizado sobre los sistemas del vehículo.

En [5] proponen un ataque al CAN bus. Hasta el momento los posibles ataques selectivos DoS a dispositivos específicos conectados al bus podían ser detectados, especialmente aquellos que usaban inyección de frames. El trabajo de TrendLabs se diferencia a ataques estudiados previamente en que reúsa frames ya existentes en el bus a los que modifica un único bit (de 1 a 0) de una posición específica induciendo un error. Este ataque es paradigmático, ya que no es detectable y para hacerlo y poder bloquearlo hacen falta cambios al standard.

Investigadores de la universidad de California en San Diego, demostraron en 2015 que eran capaces de explotar las vulnerabilidades del puerto OBD a través de un dispositivo comúnmente usado por empresas de seguros y seguimiento de flotas. A través de la llave electrónica conectada al Puerto ODB y usando remotamente un SMS, explotaron la vulnerabilidad y transmitieron mensajes CAN para poder controlar el vehículo, por ejemplo, frenándolo. Entre las compañías que usaban estos dispositivos está la española Coordina, no obstante, informaron que la vulnerabilidad solo se daba en dispositivos antiguos y que estaban actualizándolos. [6]

De una manera similar en [7], presentan cómo utilizando herramientas disponibles para el mantenimiento de vehículos pesados como camiones o autobuses se puede tomar control de las comunicaciones CAN. En este caso, se aprovecha del protocolo J1939 [8], habitual en este tipo de vehículos para la integración de comunicaciones de distintos sistemas desarrollados por distintos proveedores, para inducir a comportamientos no deseados de los sistemas de control.

En 2018 Keen Security Labs reveló tras su estudio de vulnerabilidades sobre vehículos BMW que eran susceptibles de ser atacados y el autor tomar control remoto de los buses CAN [9].

III. Bus CAN

El Can-Bus [10] es un protocolo de comunicación en serie desarrollado por Bosch para el intercambio de información entre unidades de control electrónicas del automóvil. Gracias a este protocolo de comunicación se produce una gran reducción en el número de sensores y en el cableado que componen la instalación eléctrica. El bus CAN transmite información de diferentes funciones como el control del motor, ABS, la dirección entre otros.

Una de las desventajas más destacables es la seguridad donde hasta ahora no ha sido necesaria ya que se encontraba aislada con el exterior, excepto por una conexión física que se emplea para el diagnóstico del vehículo. Por desgracia, la integridad, confidencialidad y disponibilidad de la información se ve amenazada como resultado de la interconectividad. Por lo tanto, se deben tomar contramedidas de seguridad para prevenir posibles ataques o para minimizar los riesgos de dichos ataques. Ninguna contramedida por si sola tiene éxito para evitar todos los ataques, por lo que normalmente se aboga por realizar un “defensa profunda” que no es nada menos que intentar asegurar el sistema a través de una serie de capas, utilizando diferentes estrategias.

La criptografía es una técnica que realiza una conversión de los datos en un formato secreto donde se pueden cumplir requisitos de autenticidad, confidencialidad, integridad y no-repudiación. En este trabajo se quiere estudiar esta técnica para mejorar la seguridad del CAN bus pero hay que tener ciertas limitaciones en cuenta del Bus CAN a la hora de aplicar los algoritmos de criptografía.

- La estructura de mensaje: Esta es la primera y más importante cuestión a tener en cuenta. Debemos saber cómo está definido un mensaje CAN para saber si es posible o no encriptar este mensaje, es decir, saber cómo está estructurada su trama [11]. En la **¡Error! No se encuentra el origen de la referencia.** se muestra como es la trama del protocolo CAN de forma simplificada. Aunque hay dos formatos de mensaje de dos longitudes diferentes ambos se dividen en los siguientes 3 campos más destacables: a) El *campo identificador*, este es un valor numérico que controla su prioridad del mensaje en el bus, por lo que no hay una dirección explicita en el mensaje. b) El *campo de control* especifica el tamaño del campo de datos, y por último, c) el *Campo de datos* donde se almacena la información que se quiere transmitir.

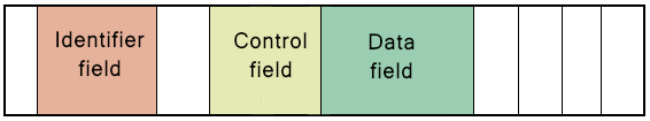


Figura 2: Trama simplificada del protocolo CAN

Este protocolo de comunicación está orientado al mensaje y no al destinatario, es decir, que todos los nodos son transmisores y receptores. Todos reciben el mensaje, lo filtran y solo emplean dicho dato los nodos que lo necesitan. Los receptores sabrán si necesitan dicha información a través del campo identificador. Si se da el caso en que dos o más emisores quieren introducir un mensaje en el bus al mismo tiempo lo harán a través del número de prioridad donde el que tenga mayor prioridad, mandará su mensaje primero.

- Tamaño del mensaje: Una de las limitaciones a la hora de asegurar el CAN bus es el ancho de banda. El campo de datos del protocolo envía 8 bytes de información en cada trama dificultando la aplicación de algoritmos.

- Tiempo de cómputo: Debido a las limitaciones de la potencia computacional de una ECU (Engine Control nit) la ejecución de instrucciones complejas necesita un mayor tiempo. Además, el software en automoción tiene restricciones de tiempo real, en particular a las aplicaciones que deben ejecutarse en un instante determinado para garantizar la seguridad del vehículo y sus pasajeros. Por lo tanto, cualquier algoritmo de criptografía no puede afectar de forma significativa al rendimiento de las funciones críticas.

- Ciclo de vida del vehículo: El ciclo de vida de un vehículo es mucho más largo, alrededor de 20 años, que un ordenador. Por lo tanto, el aseguramiento de sistema debe ser suficientemente efectivo a lo largo del tiempo.

Algunas de las características que se exaltan como ventajas del bus CAN hacen que, por otro lado, desde el punto de vista de la seguridad, sean un gran reto.

- Por la propia naturaleza de transmisión tipo broadcast, todos los nodos reciben los mensajes y son capaces de leerlo, porque esto significa que es posible escuchar a escondidas la información que circula por el nodo, es decir, no hay confidencialidad entre de la red de comunicación.

- No hay mecanismos de autenticidad, ya que no hay ningún modo de identificar el origen del mensaje. El campo de identificación no da información de quien ha enviado el mensaje, por lo tanto, los atacantes pueden enviar mensajes falsos con un alto nivel de prioridad comprometiendo el bus o causar problemas enviando errores de trama falsos.

- Un nodo malicioso puede comprometer al bus enviando mensajes constantemente de una alta prioridad provocando que otros nodos no puedan mandar mensajes, es decir, que un atacante puede implementar una denegación de servicio.

- Actualmente no hay ninguna forma de que un nodo demuestre que no ha recibido o enviado un mensaje en particular (no repudio).

- Las limitaciones que se producen por las estrictas limitaciones del tiempo real. Los plazos de respuesta deben de respetarse siempre de forma estricta y una sola repuesta fuera del intervalo de tiempo especificado puede tener consecuencias fatales para el sistema.

- La criticidad de la información que se envía a través de este bus normalmente es muy alta y puede ser información de funciones críticas de seguridad funcional (functional safety), es decir, cuando un sistema tiene funciones que se dedican a asegurar la integridad de un proceso que contiene riesgos potenciales que podrían desencadenar en un accidente con implicaciones graves para los seres humanos y su entorno.

- Por último, la limitación en la cantidad de información que puede enviarse en cada mensaje. Un mensaje puede contener entre 1 y 8 bytes de datos. Cualquier dato extra que se vaya a necesitar añadir en el mensaje debe estar incluido en él.

Teniendo en cuenta todas estas desventajas del protocolo CAN, se va a estudiar cuales de ellas pueden ser solventadas y cuáles no.

I. Seguridad criptográfica en CAN

Con el objetivo de cubrir las nuevas necesidades de seguridad, se ha optado por analizar la criptografía con el objetivo de conseguir la integridad y autentificación de los datos que circulan por el CAN. Se da más importancia a estas dos características que a la confidencialidad, ya que normalmente la información que circula no es información sensible del usuario.

Teniendo en cuenta los retos que se han mostrado en el apartado anterior la trama del mensaje no puede ser cifrada por completo, ya que por la propia naturaleza del mensaje CAN es necesario que el campo identificador esté visible para los demás nodos. Ya que como se ha mencionado

anteriormente es necesario saber la prioridad de los mensajes que se quieren enviar en un momento determinado para enviarlos de forma secuencial por su orden de prioridad. De la misma forma cuando el mensaje ha sido enviado todos los nodos receptores deben ser capaces de leer el campo identificador, y de esta forma el receptor será capaz de saber si debe o no procesar la información. Por lo tanto, se concluye que el campo que puede ir cifrado es el campo de datos. Todos los demás campos realizan una función que necesitan estar en claro.

La autentificación del mensaje es el objetivo principal. De esta forma el receptor sabrá que el mensaje recibido es legítimo Para ello se ha propuesto usar código de autentificación de mensaje, a menudo llamado por su sigla MAC. Teniendo en cuenta las altas restricciones del tiempo real, hay que emplear un algoritmo MAC que sea rápido usando un tamaño pequeño de clave. También hay que tener en cuenta que esa MAC va a ser incluida dentro del campo de datos del mensaje CAN por lo que es necesario que ocupe poco espacio. Según Hiroshi y Ryo [12] un algoritmo que satisface estas necesidades es el Código de Autenticación de Mensajes Hash con clave (HMAC). En criptografía, un código HMAC implica una función hash en combinación con una clave criptográfica secreta. Puede usarse cualquier función hash criptográfica como MD5 oSHA-1, puede usarse en el cálculo de un HMAC. La fuerza del HMAC depende de la función hash, del tamaño de su salida hash, y del tamaño y cualidad de la clave. La MAC será almacenada dentro del campo de datos junto con la información que será enviada.

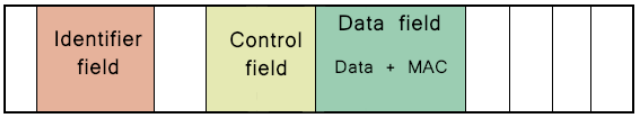


Figura 3: Campo de datos con MAC encriptado.

Si se aplica la criptografía como se ha definido anteriormente, se consigue obtener integridad y autenticidad de la información. Pero para conseguir la seguridad completa falta asegurar la disponibilidad, es decir, que la información o el servicio debe estar disponible cuando sea necesario. ¿Cómo se puede evitar un ataque de denegación de servicio en este protocolo? A través de la criptografía no se consigue evitar, ya que un atacante solo necesita inyectar continuamente un mensaje con una alta prioridad para bloquear los demás mensajes.

IV. Próximos pasos y Conclusiones

La exposición de los vehículos a nuevas redes de comunicación exterior requiere que se aseguren las comunicaciones internas de cada vehículo. Por desgracia, implementar la criptografía genera un coste adicional en el procesamiento de los datos y en el retardo en los plazos de respuesta provocando una gran limitación en la seguridad que se podría añadir al bus

Uno de los siguientes pasos es recopilar el estado del arte de todas las metodologías que se han ido desarrollando con el objetivo de obtener la autenticidad en el CAN y posteriormente aplicar los algoritmos criptográficos más interesantes, y comparar su rendimiento en tiempo y coste computacional. Otro paso que se va a estudiar del que no se ha hecho referencia en este artículo es el intercambio y actualización de las claves de autentificación. Estudiar cada cuanto tiempo la clave debe ser actualizada y saber cómo evitar la carga del bus debido a este intercambio.

Referencias

[1] In-Vehicle Networking Solutions [Online] https://www.renesas.com/us/en/application/automotive/in-vehicle-networking-application

[2] E. Stenberg, “Securing the Connected Car” Mender.io, pp. 9,https://elinux.org/images/0/01/Securing_the_Connected_Car.pdf

[3] J. Petit and S. E. Shladover, “Potential Cyberattacks on Automated Vehicles, ”IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 546–556, 2015.

[4] ENISA: “Cyber Security and Resilience of smart cars. Good practices and recommendations” December 2016; link: https://www.enisa.europa.eu/publications/cyber-security-and-resilience-of-smart-cars

[5] Upstream Security Global Automotive cybersecurity report 2019 https://documents.trendmicro.com/assets/A-Vulnerability-in-Modern-Automotive-Standards-and-How-We-Exploited-It.pdf

[6] Hackers Cut a Corvette's Brakes Via a Common Car Gadget [Online] https://www.wired.com/2015/08/hackers-cut-corvettes-brakes-via-common-car-gadget/

[7] Burakova, Yelizaveta et al. “Truck Hacking: An Experimental Analysis of the SAE J1939 Standard.” WOOT(2016). https://keenlab.tencent.com/en/Experimental_Security_Assessment_of_BMW_Cars_by_KeenLab.pdf

[8] SAE J1939 Standards Collection, https://www.sae.org/publications/collections/content/j1939_dl/

[9] Keen Security Lab, “Experimental Security Assessment of BMW Cars: A Summary Report” 2018, https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Assessment_of_BMW_Cars_by_KeenLab.pdf

[10] ISO 11898 Road vehicles - Controller area network (CAN)[Online] https://www.iso.org/standard/63648.html

[11] S.Corrigan, “Introduction to the Controller Area Network (CAN)”, Texas Instruments Report May 2016

[12] U. Hiroshi, “Security Authentication System for In-Vehicle Network”, SEI Technical Review October 2015

Sesión de Investigación B1:
Inteligencia Artificial en ciberseguridad

Sesión de Investigación B1: Inteligencia Artificial en ciberseguridad
http://doi.org/10.18239/jornadas_2021.34.37

Estado del arte en generación y detección de contenido sintético. Limitaciones y oportunidades

Miguel Hernández Boza
BBVA Next Technologies
Av. Manoteras 44. Planta 5. 28050 Madrid
https://orcid.org/0000-0001-8610-6225

José Ignacio Escribano Pablos
BBVA Next Technologies
Av. Manoteras 44. Planta 5. 28050 Madrid
https://orcid.org/0000-0002-0079-642X

***Resumen*—Los avances en nuevas tecnologías, sobre todo las relacionadas con inteligencia artificial y el auge del *deep learning*, han puesto en tela de juicio los mecanismos tradicionales para determinar si una información es legítima o no. Herramientas como *StyleGAN2*, *GROVER* o *Deepfake* son buenos ejemplos de cómo es posible generar información manipulada de alta calidad. En este artículo, se detallan las técnicas, algoritmos y herramientas que permiten generar y detectar contenido falso, profundizando en cada área (audio, texto, imagen y vídeo) haciendo énfasis en las limitaciones que existen en su uso para generar identidades falsas para cometer cualquier tipo de fraude.**

***Index Terms*—Contenido Sintético, Deep Learning, Inteligencia Artificial, redes GAN, Seguridad Defensiva, Seguridad Ofensiva, Fake News.**

Tipo de contribución: *Investigación original*

I. Introducción

Cada vez más, se están desarrollando técnicas de aprendizaje automático (*machine learning*, en inglés) relacionadas con la seguridad de la información, evolucionando y aplicándose en diversos ámbitos. Desde la generación y detección de archivos maliciosos basados en mutaciones automáticas, reconocimiento masivo de personas a través de visión computacional hasta el tema que nos ocupa en este artículo con la generación de perfiles sintéticos que permitan a un atacante engañar fácilmente a usuarios o máquinas. Debido a esto, es necesario analizar las medidas defensivas a las que podemos acudir para poder protegerse frente a este nuevo vector de ataque.

I-A. Machine Learning y ciberseguridad

En los últimos años, el *machine learning* [1], y en particular, el *deep learning* [2] se han convertido en potentes herramientas con las que automatizar una gran cantidad de procesos, mejorando en muchos casos la detección de patrones en grandes cantidades de datos. Se ha aplicado con éxito en multitud de ámbitos como medicina o astronomía. Un área donde el *machine learning* ha cobrado una notable relevancia es la seguridad de la información, ya que es posible aplicarlo en múltiples escenarios como pueden ser la detección de anomalías en tráfico de red [3–5], la clasificación de malware automática [6–8], la detección de vulnerabilidades [9–11] o la estenografía [12–14] entre otros.

Además de estos casos, ha aparecido una nueva rama de investigación relacionada con la generación de datos a partir de estos algoritmos de *deep learning*. Estos nuevos algoritmos están basados en redes GAN [15], dos redes neuronales contrapuestas donde aparece un generador y un detector que aprenden el uno del otro hasta llegar a un equilibrio. Este nuevo paradigma también ha sido aplicado en ciberseguridad [16] con gran éxito como por ejemplo en la destrucción de contenido oculto con estenografía [17]. Un análisis detallado de las áreas de aplicación en ciberseguridad se pueden encontrar en [18–22].

Organización. La Sección II contiene algunas de las manipulaciones clásicas de información que se han dado a lo largo de la historia y la Sección III todo lo relacionado con contenido sintético y las amenazas en ciberseguridad. Las Secciones IV y V se centran en la generación y detección de contenido en audio, texto, imagen y vídeo, respectivamente, poniendo énfasis en las herramientas *open source*. La Sección VI contiene las próximas tendencias y amenazas que pueden tener lugar en el futuro. Por último, la Sección VII contiene las conclusiones de la investigación realizada.

II. Manipulaciones clásicas

Siempre ha existido la modificación o manipulación de información [23], [24], pero con el paso del tiempo ha ido evolucionando. Comenzó con casos muy rudimentarios, manipulaciones caseras o muy burdas pero efectivas que pretendían causar un gran impacto social (manipulaciones en carteles de manifestaciones) o provocar una reacción en la gente, como el vandalismo. Tradicionalmente la manipulación se ha dado en imágenes y se tiene constancia que desde 1860 se han empleado técnicas en manipulación (*Fig.* 1) que llegan hasta nuestros días. Otras manipulaciones clásicas de imágenes se pueden ver en [25].

Figura 1: Foto manipulada de Abraham Lincoln de alrededor de 1860. A la izquierda se muestra la imagen manipulada y a la derecha la imagen original. Fuente: [25].

Otros ejemplos más recientes vienen dados por el uso de herramientas como Photoshop, una herramienta de retoque y

edición de fotografías, que desde versiones iniciales permitía la manipulación de imágenes con una gran calidad de salida.

A raíz del crecimiento exponencial del *deep learning* y las redes GAN, la generación de contenidos multimedia de forma automática y masiva se ha vuelto una realidad. En las *Figs.* 2 y 4 se puede ver algunas de las técnicas más sofisticadas en cuanto a generación y detección, respectivamente.

III. Contenido sintético

Contenido sintético es un término general para la producción, manipulación y modificación artificial de datos y por medios automatizados, especialmente mediante el uso de algoritmos de inteligencia artificial, por ejemplo, con el fin de engañar a las personas [26]. En este artículo, este contenido sintético estará centrado en el contenido multimedia que trata de ser lo más verídico posible tratando de engañar a personas o máquinas.

El contenido de los medios sintéticos ha crecido rápidamente desde la creación de las redes GAN, principalmente mediante el aumento de las de las *fake news* [27], así como, a la generación de caras hiperrealistas, la síntesis de voz y los *deepfakes*.

III-A. Amenazas en ciberseguridad

El aumento de la generación de contenido sintético supone una serie de amenazas en ciberseguridad. En el caso de audio, las técnicas actuales permiten crear voces humanas que puedan aparentar ser de una persona real (clonada)[1] o de una persona que no exista (basada en una fuente de datos)[2]. Esto podría permitir la suplantación de personas. En imágenes, aparece la capacidad de crear rostros realistas de personas que no existen[3], lo que implica poder crear identidades falsas. En el caso del texto, da la posibilidad de crear textos coherentes falsos[4] con el fin de mejorar la creación de *fake news*, y por último en vídeo, es posible insertar caras en cuerpos que no se corresponden con la persona original[5].

Los medios sintéticos incluyen el potencial de crear noticias falsas, propagar la desinformación, aumentar la desconfianza en la realidad y automatizar de forma masiva de los trabajos creativos y periodísticos.

IV. Generación de contenido sintético

En esta sección se desgranarán los métodos y modelos que se utilizan actualmente para realizar contenido sintético, centrándose en el contenido multimedia con el objetivo de ser lo más similar posible al contenido real. Para ello, las herramientas hacen uso de las técnicas de *deep learning* que se han mencionado en la Sección I-A. En la *Fig.* 2 se puede ver la evolución de las técnicas en la generación de contenido.

IV-A. Audio

En los últimos años, distintas empresas como Lyrebird[6] o Resemble.ai[7] han aparecido en el panorama con el objetivo de clonar voz usando inteligencia artificial. Lyrebird prometía clonar una voz empleando sólo 30 frases predefinidas de antemano. Con estas frases, la inteligencia artificial era capaz de entrenar los audios con las frases grabadas y obtener un modelo de la voz. La limitación de este sistema es que está pensado para clonar voces en inglés. De igual forma, Resemble.ai es una plataforma similar a Lyrebird cuyo objetivo es clonar voces. En este caso, requiere 50 frases prefijdas (en inglés) para clonar una voz. La novedad que presenta esta plataforma es que dispone de distintas voces preentrenadas, incluyendo distintos acentos y voces masculinas y femeninas. En particular, incluye voces en español tanto en español de España como de Sudámerica.

Dejando a un lado las dos herramientas anteriores, Real Time Voice Cloning [28] es la herramienta *open source* más potente de clonado de voz, cuyo código está disponible en GitHub [29]. Está compuesto de cuatro partes: SV2TTS [30], WaveRNN [31] como *vocoder*, Tacotron 2 [32] como sintetizador y GE2E [33] como *encoder*. El objetivo de la herramienta es clonar una voz usando sólo 5 segundos de audio. Después de realizar distintas experimentos con esta herramienta siguiendo la configuración descrita en su manual, no hemos conseguido obtener una voz realista usando sólo 5 segundos. Esta herramienta sólo requiere un audio de la voz que se requiera clonar y el texto que dirá la voz clonada. Los resultados más prometedores se han obtenido al emplear unos 10 minutos de audio. A modo de ejemplo, usando 10 minutos de un discurso de Donald Trump[8] se pueden obtener resultados realistas, aunque lejos de ser perfectos, ya que es posible escuchar cierto tono robótico. De hecho, si se emplean textos con palabras complejas, como pueden ser textos de Shakespeare[9], se pueden ver las limitaciones del modelo entrenado: no es capaz de generalizar palabras que no ha visto con anterioridad. Esto pone en entredicho que esta herramienta se pueda emplear para clonar todas las voces y convertir a voz cualquier texto. Esta limitación se suma a la comentada previamente sobre el idioma: solamente genera voces en inglés. Para generar voces en otros idiomas, es necesario disponer de un conjunto de datos de alrededor de 300 horas en el idioma deseado de más de 100 personas distintas[10]. Los proyectos anteriormente mencionados requieren de datos que son costosos de obtener, sin embargo, los proyectos de Common Voice de Mozilla[11] o M-AILABS[12] proporcionan una serie de datos interesantes. El primero dispone de audios en muchos idiomas, incluyendo el español, aunque la calidad no es demasiado buena, debido a que los micrófonos con los que ha sido grabados no tienen la suficiente calidad. El segundo, M-AILABS, dispone también de audio en distintos idiomas; este conjunto de datos está pensado para ser empleado en algoritmos de inteligencia artificial, aunque, no todos los audios disponibles de este conjunto de datos disponen de las horas suficientes para hacer funcionar Real Time Voice Cloning.

Con todas estas aplicaciones queda claro que es posible llegar a clonar una voz de manera sintética con un alto grado

[1] https://www.pandasecurity.com/mediacenter/news/deepfake-voice-fraud/

[2] https://youtu.be/s0UJCb5ZjJ0

[3] https://www.wired.com/story/facebook-removes-accounts-ai-generated-photos

[4] https://talktotransformer.com/

[5] https://youtu.be/GTh2tRAE2w4

[6] Ahora forma parte de Descript, una empresa de transcripción de audio a texto de forma automática. https://www.descript.com/lyrebird-ai

[7] https://www.resemble.ai

[8] https://youtu.be/KWcmZ8hozvU

[9] https://tinyurl.com/zxg8qle

[10] https://git.io/JvEbj

[11] https://voice.mozilla.org

[12] https://www.caito.de/2019/01/the-m-ailabs-speech-dataset

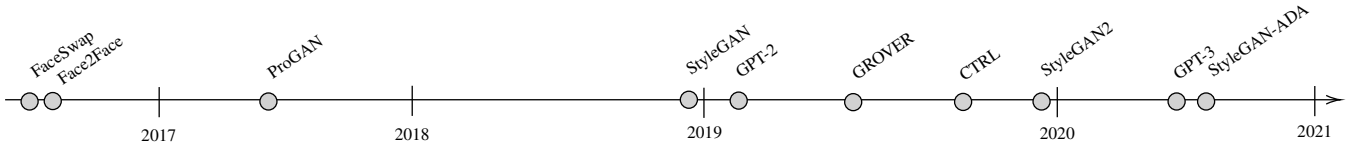


Figura 2: Principales hitos en la generación de contenido sintético.

de fiabilidad si se tienen los suficientes datos disponibles. Si se quisiera realizar un ataque real, es posible realizar inyecciones de audio clonado para suplantar a otra persona [34].

IV-B. Texto

Siempre ha existido un gran interés por la redacción de textos automática, al igual que un gran conocimiento sobre cómo puede tratar la máquina con los textos. Este tema sería una extensión del conocido como procesamiento natural del lenguaje, que permite un análisis de un texto sintética y semánticamente.

Cuando hablamos de generar un texto sintético empleando *deep learning*, los sistemas actuales están basados en la premisa de predecir la siguiente palabra dado un mensaje o contenido inicial, ya que por sí solos no son capaces de generarlo. A partir de este concepto, diversos investigadores comenzaron a buscar fuentes de datos que fueran lo suficiente ricos en contenido para que los modelos pudieran aprender. OpenAI liberó el modelo GPT [35] y tras unos años de controversia sobre los usos éticos del generador también liberó su evolución llamada GPT-2 [36], un modelo compuesto de 1 500 millones de parámetros. A raíz de su publicación, se han desarrollado distintas aplicaciones *online* donde probar este modelo. Un ejemplo de esto es *Talk to Transformer*[13]. La versión más reciente es GPT-3 [37], que permite interpretar textos. Este modelo no es *open source* y se requiere de licencia para usarse.

Este no es el único modelo, existen muchos otros como CTRL [38] o GROVER [39] que mantienen esta premisa, en el caso de GROVER permite tanto la generación como detección de texto.

Usando GPT-2 o GROVER es posible ver las limitaciones en cuanto al idioma que tienen estos sistemas, ya que al ser entrenados en inglés están forzados a utilizar ese lenguaje para funcionar correctamente. Aún así, existen casos de éxito a la hora de simular un comportamiento humano. Por ejemplo, esto sucede en redes sociales como Twitter que además tienen la limitación de cantidad de caracteres: muchos *bots* envían mensajes para evitar ser detectados por los sistemas anti-automatización. En STIC CCN-CERT se publicó[14] que algunos simpatizantes con el Brexit eran *bots* controlados por los ponentes.

Por lo tanto, siempre que el idioma no sea un impedimento, empiezan a aparecer usos maliciosos que ya preveía OpenAI[15] antes de liberar su sistema y es necesario combatirlo de manera efectiva. En la sección V-A de detección de audio profundizaremos en ello.

[13]https://talktotransformer.com

[14]https://youtu.be/wzctZgha1yw

[15]https://www.theverge.com/2019/11/7/20953040/openai-text-generation-ai-gpt-2-full-model-release-1-5b-parameters

IV-C. Imagen

StyleGAN [40] es la herramienta estado del arte en generación de imágenes de caras hiperrrealistas, desarrollada por NVIDIA Labs. Tanto el código como los datos de entrenamiento han sido liberados en GitHub [41], [42], junto con los modelos preentrenados disponibles en Google Drive[16]. StyleGAN emplea una modificación de la red GAN [15] como arquitectura y conseguir resultados hiperrealistas. Algunas de las imágenes generadas con StyleGAN se puede ver en la *Fig.* 3. Se pueden observar algunas de las limitaciones de este sistema: una oreja sobresale por encima del pelo, los pendientes son de distinto tipo colgando en el aire y los dientes no están alineados con respecto al centro de la cara (los dientes se muestran como si la cara estuviese siempre de frente). Para solucionar estos problemas, a finales de 2019, se liberó una nueva versión, llamada StyleGAN2 [43]. En esta segunda versión se han empleado los mismos datos de entrenamiento que en la primera versión, pero la arquitectura de la red GAN se ha modificado para que corrija los problemas descritos anteriormente. Como en la versión previa, todo el código y modelos preentrenados han sido publicados [44]. Además, esta nueva versión introduce el concepto de *proyección* que, dada una imagen objetivo, se itera para obtener la imagen más similar que se podría obtener con StyleGAN2. Si la imagen ha sido generada por este sistema, la imagen de la proyección converge a una imagen muy similar de la imagen objetivo. Sin embargo, si la imagen objetivo es una imagen real las proyecciones convergerán a una imagen muy distinta a la objetivo (ver *Fig.* 5 para un ejemplo). StyleGAN2 es el modelo empleado en la página más famosa de generación de caras falsas: https://www.thispersondoesnotexist.com. La versión más reciente es StyleGAN-ADA [45], que necesita menos datos de entrenamiento obteniendo resultados similares a StyleGAN2.

Los datos de entrenamiento de StyleGAN también han sido también publicados en GitHub [42] bajo el nombre de FFHQ. Los datos constan de 70 000 imágenes de caras de tamaño 1024x1024 extraídas de Flickr[17] con licencia de libre uso y redistribución fijada por los autores de las fotos. Este conjunto de datos tiene una mayor variabilidad en cuanto a edad, grupos étnicos, accesorios (gafas, gafas de sol, gorros, etc.) y fondos de imágenes que conjuntos de datos previos como CELEBA-HQ [46], aunque el conjunto de datos hereda todos los sesgos (*bias*) presentes en Flickr.

IV-D. Vídeo

Los avances en la digitalización de los rostros humanos se han convertido en la base de las modernas herramientas de edición de imágenes faciales. Las herramientas de edición de

[16]https://tinyurl.com/t5ys8dq

[17]https://www.flickr.com

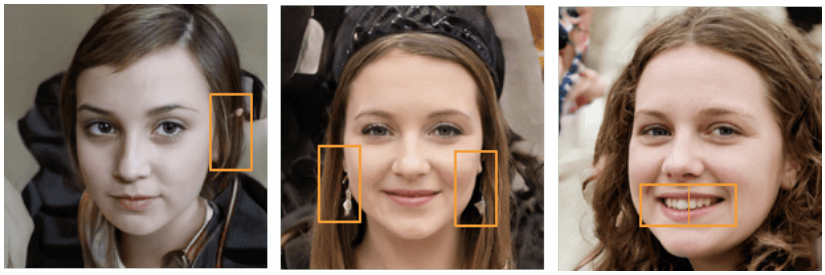


Figura 3: Algunas de las limitaciones en la generación de StyleGAN. Imágenes generadas usando el código oficial disponible en GitHub [41].

vídeo pueden dividirse en dos categorías principales: modificación de la identidad, con herramientas como FaceSwap [47], y modificación de la expresión, con Face2Face [48]. Pero estas no son las únicas herramientas que existen. En la Tabla I se enumeran las más reconocidas cuyo código está disponible. Estas herramientas están diseñadas para generar *deepfakes*, vídeos modificados en los que una cara se ha introducido en un vídeo objetivo. El estado del arte sobre la generación de *deepfakes* se puede encontrar en [49].

Tabla I: Algunas herramientas *open source* para crear *deepfakes*.

Herramienta	URL
Deepfakes	https://github.com/deepfakes/faceswap
Faceswap-GAN	https://github.com/shaoanlu/faceswap-GAN
DFaker	https://github.com/dfaker/df
DeepFaceLab	https://github.com/iperov/DeepFaceLab

En nuestros experimentos, hemos empleado la herramienta Deepfakes [50] para crear *deepfakes*, aunque las demás herramientas de la Tabla I tienen un funcionamiento similar. Esta herramienta se encarga de generar *deepfakes* en tres fases: extraer caras, entrenar el modelo e insertar las cara. La extracción de caras se puede realizar a través de distintos extractores disponibles en la herramienta. El entrenamiento del modelo se realiza a partir de las caras extraídas en el paso anterior, produciendo un modelo que es capaz de transformar una cara en otra. El tercer y último paso, a partir de un vídeo objetivo, transforma la caras aplicando del modelo entrenado previamente, fotograma a fotograma. Para obtener resultados aceptables con esta herramienta es necesario disponer de una GPU. En nuestro caso, empleamos una NVIDIA GTX 1080 con 8GB de RAM. Además, es imprescindible contar con una gran cantidad de vídeos de los que extraer las caras que cuenten con distintos ángulos e iluminaciones. En la medida de lo posible, estos vídeos deben tener la mejor resolución posible. En nuestro caso, empleamos vídeos en calidad 1920x1080 (*Full HD*). A la hora de extraer las caras de los vídeos, se debe comprobar manualmente que las caras extraídas se corresponden con las caras que queremos extraer y eliminar falsos positivos para evitar que el modelo entrenado sea de la mejor calidad posible. En nuestras pruebas empleamos entre 10 y 40 minutos de vídeo, quedando entre 3000 y 15 000 caras después de eliminar los falsos positivos del detector de caras. El entrenamiento del modelo, en nuestras pruebas con el *hardware* descrito anteriormente, tardó entre 1.5 y 2 días, no produciendo mejores resultados dejando continuar el entrenamiento de forma indefinida[18]. Los vídeos[19] tienen ciertas limitaciones:

- El detector de caras sólo detecta una cara: si se aplica a un fotograma con más de una cara sólo detecta la que sea más grande o la que encuentre primero.
- Los fotogramas con caras lejanas o con determinadas posiciones de la cara tampoco son detectadas.
- Si se detecta una cara de persona con la que ha sido entrenada se producen resultados no esperados que son fácilmente detectados.
- Si se aplica a caras que tengan gafas, barbas o similares se producen resultados extraños, siendo el modelo incapaz de generalizar de forma correcta.

V. Detección de contenido sintético

El fuerte impacto que tienen los *deepfakes* en la sociedad, hizo que DeepTraceLab se preguntara si la generación de contenido sintético es un problema real. Las conclusiones fueron claras, los *deepfakes* estaban centrados en el contenido adulto y no en política [51]. Por otro lado, Danielle Citron en una charla en TED[20] explica qué poder hacer para verificar una información real, ya que podemos empezar a dudar de ella también. Con todo esto, es necesario poner énfasis en la búsqueda de soluciones para protegernos, ya que la exposición a este tipo de material es inevitable. En la siguiente sección se describen herramientas o modelos que permiten la detección del contenido sintético.

V-A. Audio

La voz nos identifica y en muchos casos nos autentica en un sistema, por lo que es necesario protegerla y seguramente ni nos demos cuenta de la exposición que le damos cada día. En la sección de generación se ha demostrado que con los suficientes datos es posible replicar una voz con gran precisión, difícil de detectar de manera manual. Por lo tanto, debido al posible uso fraudulento de estas nuevas técnicas, se ha vuelto urgente encontrar métodos de detección de voces clonadas.

En el evento sobre procesamiento de voz, Interspeech 2019[21], se trató el problema de detectar voces sintéticas, después de la generación que tuvo lugar en años anteriores.

[18]Deepfakes fija el número de iteraciones a 1 millón, aunque este parámetro es personalizable por el usuario.

[19]Deepfakes no genera el vídeo propiamente dicho sino que tranforma las caras fotograma a fotograma, aunque proporciona herramientas para unir todos los fotogramas y formar el vídeo.

[20]https://www.daniellecitron.com/ted-talk

[21]https://www.interspeech2019.org

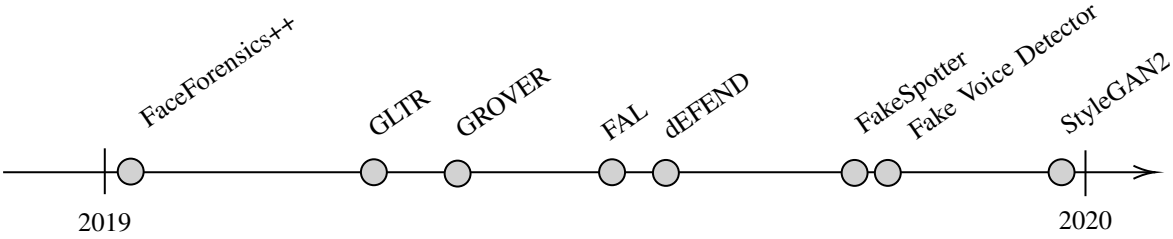


Figura 4: Principales hitos en la detección de contenido sintético.

Dentro de este evento se desarrolló el reto ASVspoof 2019 donde diversas universidades trataron de crear un modelo o herramienta que, dados unos datos etiquetados como reales o falsos, fuera capaz de identificar cómo de falsa era una voz introducida [52]. Una de las soluciones, llamada *Fake Audio Detector* publicó el código [53] en GitHub. En el repositorio se proporciona el código y el modelo preentrenado sobre el que realizar pruebas y verificar si una voz es real o no.

Esta herramienta se probó en distintos experimentos con diversos tipos de audio: muestras de Real Time Voice Cloning, Resemble.ai y Lyrebird. Los resultados obtenidos fueron inconsistentes, puesto que al introducir tanto audios falsos como reales, la probabilidad que devolvía la herramienta de ser reales en ambos casos era muy baja. Por lo que se puede decir que no son unos resultados lo suficientemente robustos como para valorar esta solución como algo fiable.

V-B. Texto

La detección de texto se está centrando en la detección de *fake news*. Los investigadores están proponiendo en la actualidad distintas técnicas para detectar este tipo de noticias manipuladas. Los grafos [54], [55] y *Capsule Neural Networks* [56] son alguna de las técnicas más novedosas para detectar este tipo de noticias. En paralelo, también se están centrando los esfuerzos en crear sistemas que recopilen, analicen y detecten *fake news* de forma automática. Ejemplos de esto son *Kauwa-Kaate Fake News Detection System* [57], *FakeNewsTracker* [58] o *dEFEND* [59] cuyo objetivo es detectar y explicar por qué las noticias detectadas son falsas.

GROVER [39] es una de las herramientas más conocidas en cuanto a la detección de *fake news*[22]. Los creadores de GROVER han liberado los modelos preentrenados y el código en GitHub [60] y se puede probar este sistema a través de su demo online[23]. Este sistema se ha creado con el objetivo de que los atacantes que generan *fake news* para difundir propaganda o hacer *clickbait* sean detectados fácilmente. Emplea la misma arquitectura que GPT-2 [36], aunque para entrenarlo se ha empleado un conjunto de datos llamado REALNEWS[24]. Este conjunto de datos ha sido formado por noticias procedentes de *Common Crawl*[25]. Las noticias extraídas para el entrenamiento tienen fechas comprendidas entre diciembre de 2016 y abril de 2019 para el entrenamiento y noticias desde abril de 2019 para la evaluación del modelo. Este conjunto de datos tiene un tamaño de 120 GB sin comprimir.

En cuanto a detectores de propósito general, GLTR [61] es uno de las más empleados. Está disponible en GitHub [62] y se puede probar de forma *online*[26]. Permite detectar textos generados con GPT-2 [36] y la detección se basa en cómo funciona este modelo, es decir, prediciendo la siguiente palabra. Así pues, GLTR va iterando por el texto y va comprobando si la siguiente palabra que aparece en el texto que queremos comprobar si es real o falso, se comprueba si aparece en las k primeras palabras de GPT-2. De esta forma se puede dar una estimación de cómo de sintético es un texto dado.

V-C. Imagen

Existe un gran número de trabajos que tratan de verificar la integridad de imágenes [63]. Por otro lado, también se está trabajando en la detección de imágenes generadas con herramientas tradicionales como Photoshop. Este es el caso de FAL [64], que permite detectar si una foto ha sido modificada con Photoshop a través de *Face-Aware Liquify*[27] disponible en Photoshop. *Face-Aware Liquify* permite cambiar expresiones faciales de una cara en unos pocos *clicks*.

En cuanto a detección de imágenes generadas artificialmente, se han propuesto numerosas alternativas. GAN Fingerprints [65] permite la detección de distintas redes GAN por el ruido introducido en la generación de la imagen. Las redes GAN que detecta son: ProGAN [46], SNGAN [66], CramerGAN [67] y MMDGAN [68]. Otra aproximación es *FakeSpotter* [69], que se basa en monitorizar el comportamiento de las neuronas de un modelo para detectar las imágenes generadas usando un clasificador binario, siendo empleado con éxito en distintos métodos de generación: ProGAN [46], StyleGAN2 [43], STGAN [70].

La detección de StyleGAN [40] y StyleGAN2 [43] se puede realizar a través de las proyecciones presentadas en StyleGAN2 y descritas en la Sección IV-C. Si empleamos como imagen objetivo (la imagen que queremos comprobar si es sintética o no) una imagen generada con cualquiera de las versiones de StyleGAN, las proyecciones convergerán a una imagen muy similar al objetivo. Sin embargo, empleando una imagen no sintética, las proyecciones no convergen al objetivo sino a una imagen que se diferencia significativamente del objetivo. Un ejemplo de detección de imágenes con este método se puede ver en la *Fig.* 5.

V-D. Vídeo

Detectar los *deepfakes* [71] se ha vuelto una necesidad debido al uso malicioso que pueden tener en áreas como la

[22]GROVER también permite generación de *fake news*, aunque sólo nos centraremos en la detección.

[23]https://grover.allenai.org

[24]https://github.com/rowanz/grover/tree/master/realnews

[25]https://commoncrawl.org

[26]http://gltr.io/dist/index.html

[27]https://helpx.adobe.com/es/photoshop/how-to/face-aware-liquify.html

(a) Imagen generada por StyleGAN

(b) Imagen generada por StyleGAN2

(c) Imagen no sintética

Figura 5: Proyecciones de StyleGAN2 usando una imagen generada por StyleGAN (a), por StyleGAN2 (b) y una no sintética (c). A la izquierda se muestra la imagen objetivo y a la derecha el resultado de aplicar las proyecciones después de 1000 iteraciones. Para generar las proyecciones se ha utilizado el código oficial de StyleGAN2 publicado en GitHub [44].

política. Han aparecido ya trabajos que permiten reconocer incoherencias en el parpadeo [65], sincronización del movimiento de los labios [72], [73] u otros [72], [74], [75]. Es necesario disponer de sistemas automáticos que no necesiten de un procesado muy grande para poder verificar un vídeo. Es por ello que han aparecido retos como DFDC [76] y bases de datos como CELEB-DF [77] para tratar de generar modelos de detección de *deepfakes*.

FaceForensics++ [78] es la herramienta, de código libre [79], más avanzada en cuanto a la detección de *deepfakes*. Esta herramienta permite detectar si un vídeo ha sido manipulado. La forma de funcionar es relativamente sencilla: el vídeo es dividido en fotogramas y por cada uno de ellos, se calcula la probabilidad de que el fotograma sea falso usando un modelo preentrenado. Este es un modelo de *deep learning* entrenado con vídeos extraídos de YouTube y estos mismos vídeos generados como *deepfakes* usando cuatro métodos de generación: DeepFakes [50], Face2Face [48], FaceSwap [80] y NeuralTextures [81].

Después de realizar distintos experimientos con FaceForenics++ (ver *Fig.* 6), es capaz de detectar *deepfakes* aunque tiene ciertas limitaciones que enumeramos a continuación:

- Sólo es capaz de detectar *deepfakes* que hayan sido creados con uno de los métodos con los que ha sido entrenado.
- En caso de que un mismo fotograma aparezca más de una cara, FaceForensics++ sólo aplica la detección a la cara más grande que detecte. Esto hace que se pueda evadir la detección si se inserta el *deepfake* en alguna de las caras más alejadas, aunque esta limitación sólo aplica a vídeos en los que aparece más de una cara.
- Falsos positivos. Normalmente estos positivos vienen dados por reconocimiento erróneo de caras, que en este caso se heredan de la librería de reconocimiento facial `dlib` [82] que emplea FaceForensics++ de forma interna.
- Difícil de instalar. La falta de información en el repositorio [79] hace que sea difícil de instalar, teniendo que cambiar rutas del código de los modelos preentrenados que se encuentran con rutas absolutas y no relativas, lo que hace que esta herramienta no sea apta para personas no técnicas.

FaceForensics++ puede ser evadido usando técnicas de *adversarial machine learning* [83], usando ejemplos adversarios (pequeñas perturbaciones añadidas a ejemplos legítimos imperceptibles para el ojo humano que hacen clasificar de forma incorrecta al modelo). Este es el ejemplo de *Adversarial Deepfakes* [84], que propone modificar de forma imperceptible usando pequeñas perturbaciones en los *deepfakes* creados con los métodos de generación vistos en la Sección IV-D, siendo efectivos tanto en caja negra (método de detección desconocido) como caja blanca (método de detección conocido). Además, los autores demuestran que el método propuesto es robusto frente a *códecs* de compresión de audio y vídeo [84].

VI. FUTURO

A partir de todo lo visto en este articulo, queda claro que queda mucho trabajo por hacer en métodos de detección y de adaptación de modelos en casos donde no se cuente con suficientes datos. Además, desde DARPA[28] están en fase de desarrollo de un sistema genérico que permita tener cierta certeza sobre los contenidos multimedia que aparecen por internet. Su sistema se llama MediFor[29] y contiene tres indicadores donde comprueban diferentes integridades (digital, semántica y física) de las imágenes analizadas. Esta organización también se está preparando para la generación masiva de contenido sintético que se realice como servicio y sobre un mismo contexto, como puede ser una manifestación, combinando cada uno de los tipos de multimedia que hemos visto anteriormente.

VII. CONCLUSIONES

A lo largo de este artículo se ha tratado de dar una visión global del contenido sintético, los diversos usos que se pueden llegar a dar para cada una de los tipos de multimedia y métodos que pueden ser usados para protegerse y poder verificar este tipo de contenido, poniendo foco en las herramientas disponibles en internet. Creemos que es necesario prestar atención a este tipo de tecnologías y comenzar a pensar en cómo legislar sobre este contenido para evitar su uso fraudulento. Hay que tener en cuenta que es una tecnología con mucho

[28] https://www.darpa.mil

[29] https://www.darpa.mil/program/media-forensics

Figura 6: Algunos fotogramas con el resultado de aplicar FaceForensics++ [79]. Nótese que, cuando aparece más de una cara en el fotograma, FaceForensics++ solamente es capaz de detectar la cara más próxima. Vídeos extraídos del canal de YouTube FaceToFake (https://tinyurl.com/uquu7sn), que se dedica a producir *deepfakes* haciendo uso de la herramienta DeepFaceLab [85].

potencial pero tiene limitaciones: las más importantes son el cambio de idioma de un modelo preentrenado y la falta de conjuntos de datos que sean de utilidad en el entrenamiento de nuevos modelos tanto para la generación como la detección de contenido sintético.

Referencias

[1] S. Marsland, *Machine Learning - An Algorithmic Perspective*, ser. Chapman and Hall / CRC machine learning and pattern recognition series. CRC Press, 2009.

[2] I. J. Goodfellow, Y. Bengio, and A. C. Courville, "Deep Learning," *MIT Press*, 2016.

[3] D. K. Bhattacharyya and J. K. Kalita, *Network Anomaly Detection: A Machine Learning Perspective*. Chapman & Hall/CRC, 2013.

[4] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, *Network Traffic Anomaly Detection and Prevention - Concepts, Techniques, and Tools*, ser. Computer Communications and Networks. Springer, 2017.

[5] F. Iglesias and T. Zseby, "Analysis of network traffic features for anomaly detection," *Machine Learning*, vol. 101, no. 1-3, pp. 59–84, 2015.

[6] B. Cakir and E. Dogdu, "Malware classification using deep learning methods," in *ACM Southeast Regional Conference*. ACM, 2018, pp. 10:1–10:5.

[7] L. Liu, B. Wang, B. Yu, and Q. Zhong, "Automatic malware classification and new malware detection using machine learning," *Frontiers of IT & EE*, vol. 18, no. 9, pp. 1336–1347, 2017.

[8] N. Peiravian and X. Zhu, "Machine Learning for Android Malware Detection Using Permission and API Calls," in *ICTAI*. IEEE Computer Society, 2013, pp. 300–305.

[9] Z. Li, D. Zou, S. Xu, X. Ou, H. Jin, S. Wang, Z. Deng, and Y. Zhong, "VulDeePecker: A Deep Learning-Based System for Vulnerability Detection," in *NDSS*. The Internet Society, 2018.

[10] F. Wu, J. Wang, J. Liu, and W. Wang, "Vulnerability detection with deep learning," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, Dec 2017, pp. 1298–1302.

[11] G. Grieco, G. L. Grinblat, L. C. Uzal, S. Rawat, J. Feist, and L. Mounier, "Toward Large-Scale Vulnerability Discovery using Machine Learning," in *CODASPY*. ACM, 2016, pp. 85–96.

[12] H. Wu, H. Wang, and Y. Shi, "Can Machine Learn Steganography? - Implementing LSB Substitution and Matrix Coding Steganography with Feed-Forward Neural Networks," *CoRR*, vol. abs/1606.05294, 2016.

[13] D. Volkhonskiy, I. Nazarov, B. Borisenko, and E. Burnaev, "Steganographic Generative Adversarial Networks," *CoRR*, vol. abs/1703.05502, 2017.

[14] J. Hayes and G. Danezis, "Generating steganographic images via adversarial training," in *NIPS*, 2017, pp. 1954–1963.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative Adversarial Networks," *CoRR*, vol. abs/1406.2661, 2014.

[16] C. Yinka-Banjo and O.-A. Ugot, "A review of generative adversarial networks and its application in cybersecurity," *Artificial Intelligence Review*, Jun. 2019. [Online]. Available: https://doi.org/10.1007/s10462-019-09717-4

[17] I. A. Corley, J. Lwowski, and J. Hoffman, "Destruction of Image Steganography using Generative Adversarial Networks," *CoRR*, vol. abs/1912.10070, 2019.

[18] S. Mahdavifar and A. A. Ghorbani, "Application of deep learning to cybersecurity: A survey," *Neurocomputing*, vol. 347, pp. 149–176, 2019.

[19] A. Handa, A. Sharma, and S. K. Shukla, "Machine learning in cybersecurity: A review," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 4, 2019.

[20] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng, "Recent progress on generative adversarial networks (gans): A survey," *IEEE Access*, vol. 7, pp. 36 322–36 333, 2019.

[21] M. O. Khursheed, D. Saeed, and A. M. Khan, "Generative Adversarial Networks: A Survey of Techniques and Methods," in *Lecture Notes on Data Engineering and Communications Technologies*. Springer International Publishing, Aug. 2019, pp. 490–498. [Online]. Available: https://doi.org/10.1007/978-3-030-24643-3_58

[22] Y. Cao, L. Jia, Y. Chen, N. Lin, C. Yang, B. Zhang, Z. Liu, X. Li, and H. Dai, "Recent Advances of Generative Adversarial Networks in Computer Vision," *IEEE Access*, vol. 7, pp. 14 985–15 006, 2019.

[23] D. Brugioni, *Photo Fakery: The History and Techniques of Photographic Deception and Manipulation*. Brassey's, 1999. [Online]. Available: https://books.google.es/books?id=3t1TAAAAMAAJ

[24] L. Zheng, Y. Zhang, and V. L. L. Thing, "A survey on image tampering and its detection in real-world photos," *J. Visual Communication and Image Representation*, vol. 58, pp. 380–399, 2019.

[25] J. Sharma and R. Sharma, "Analysis of Key Photo Manipulation Cases and their Impact on Photography," 2019.

[26] S. Rosenbaum, "What is Synthetic Media?" https://www.mediapost.com/publications/article/341074/what-is-synthetic-media.html, 2020.

[27] M. M. Waldrop, "News Feature: The genuine problem of fake news," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 114, no. 48, pp. 12 631–12 634, 2017.

[28] Corentin Jemine, "Automatic Multispeaker Voice Cloning," Master's thesis, Université de Liège, Liège, Belgique, 2019, https://hdl.handle.net/2268.2/6801.

[29] Corentin Jemine, "Clone a voice in 5 seconds to generate arbitrary speech in real-time," https://github.com/CorentinJ/Real-Time-Voice-Cloning, 2019.

[30] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen,

P. Nguyen, R. Pang, I. Lopez-Moreno, and Y. Wu, "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech synthesis," in *NeurIPS*, 2018, pp. 4485–4495.

[31] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient Neural Audio Synthesis," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 2415–2424.

[32] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *ICASSP*. IEEE, 2018, pp. 4779–4783.

[33] L. Wan, Q. Wang, A. Papir, and I. Lopez-Moreno, "Generalized End-to-End loss for Speaker Verification," in *ICASSP*. IEEE, 2018, pp. 4879–4883.

[34] Y. Chen and H. C. Ma, "Biometric Authentication Under Threat: Liveness Detection Hacking," in *Black Hat*, 2019.

[35] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[36] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[37] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[38] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "CTRL: A Conditional Transformer Language Model for Controllable Generation," *CoRR*, vol. abs/1909.05858, 2019.

[39] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending Against Neural Fake News," in *NeurIPS*, 2019, pp. 9051–9062.

[40] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 4401–4410.

[41] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "StyleGAN - Official Tensorflow Implementation," https://github.com/NVlabs/stylegan, 2018.

[42] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Flickr-Faces-HQDataset (FFHQ)," https://github.com/NVlabs/ffhq-dataset, 2019.

[43] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," *CoRR*, vol. abs/1912.04958, 2019.

[44] ——, "StyleGAN2 - Official Tensorflow Implementation," https://github.com/NVlabs/stylegan2, 2018.

[45] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," 2020.

[46] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," in *ICLR*. OpenReview.net, 2018.

[47] D. Bitouk, N. Kumar, S. Dhillon, P. N. Belhumeur, and S. K. Nayar, "Face swapping: automatically replacing faces in photographs," *ACM Trans. Graph.*, vol. 27, no. 3, p. 39, 2008.

[48] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: real-time face capture and reenactment of RGB videos," *Commun. ACM*, vol. 62, no. 1, pp. 96–104, 2019.

[49] R. Tolosana, R. Vera-Rodríguez, J. Fiérrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," *CoRR*, vol. abs/2001.00179, 2020.

[50] deepfakes, "Deepfakes Software For All," https://github.com/deepfakes/faceswap, 2019.

[51] G. Patrini, "The State of deepfakes 2019," https://deeptracelabs.com/mapping-the-deepfake-landscape, 2019.

[52] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. W. D. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," *CoRR*, vol. abs/1904.05441, 2019.

[53] Dessa, "Using temporal convolution to detect Audio Deepfakes," https://github.com/dessa-public/fake-voice-detection, 2019.

[54] E. C. Garrido-Merchán, C. Puente, and R. Palacios, "Fake News Detection by means of Uncertainty Weighted Causal Graphs," *CoRR*, vol. abs/2002.01065, 2020.

[55] Y. Ren and J. Zhang, "HGAT: Hierarchical Graph Attention Network for Fake News Detection," *CoRR*, vol. abs/2002.04397, 2020.

[56] M. H. Goldani, S. Momtazi, and R. Safabakhsh, "Detecting Fake News with Capsule Neural Networks," *CoRR*, vol. abs/2002.01030, 2020.

[57] A. Bagade, A. Pale, S. Sheth, M. Agarwal, S. Chakrabarti, K. Chebrolu, and S. Sudarshan, "The Kauwa-Kaate Fake News Detection System: Demo," in *COMAD/CODS*. ACM, 2020, pp. 302–306.

[58] K. Shu, D. Mahudeswaran, and H. Liu, "FakeNewsTracker: a tool for fake news collection, detection, and visualization," *Computational & Mathematical Organization Theory*, vol. 25, no. 1, pp. 60–71, 2019.

[59] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "dEFEND: Explainable Fake News Detection," in *KDD*. ACM, 2019, pp. 395–405.

[60] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Code for Defending Against Neural Fake News," https://github.com/rowanz/grover, 2019.

[61] S. Gehrmann, H. Strobelt, and A. M. Rush, "GLTR: Statistical Detection and Visualization of Generated Text," in *ACL (3)*. Association for Computational Linguistics, 2019, pp. 111–116.

[62] H. Strobelt and S. Gehrmann, "Giant Language Model Test Room," https://github.com/HendrikStrobelt/detecting-fake-tex, 2019.

[63] P. Korus, "Digital image integrity - a survey of protection and verification techniques," *Digit. Signal Process.*, vol. 71, pp. 1–26, 2017.

[64] S. Wang, O. Wang, A. Owens, R. Zhang, and A. A. Efros, "Detecting Photoshopped Faces by Scripting Photoshop," *CoRR*, vol. abs/1906.05856, 2019.

[65] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," in *WIFS*. IEEE, 2018, pp. 1–7.

[66] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," in *ICLR*. OpenReview.net, 2018.

[67] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos, "The Cramer Distance as a Solution to Biased Wasserstein Gradients," *CoRR*, vol. abs/1705.10743, 2017.

[68] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *ICLR (Poster)*. OpenReview.net, 2018.

[69] R. Wang, L. Ma, F. Juefei-Xu, X. Xie, J. Wang, and Y. Liu, "FakeSpotter: A simple Baseline for Spotting AI-Synthesized Fake Faces," *CoRR*, vol. abs/1909.06122, 2019.

[70] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, "STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing," in *CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 3673–3682.

[71] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *AVSS*. IEEE, 2018, pp. 1–6.

[72] R. C. Bolles, J. B. Burns, M. Graciarena, A. Kathol, A. Lawson, M. McLaren, and T. Mensink, "Spotting Audio-Visual Inconsistencies (SAVI) in Manipulated Video," in *CVPR Workshops*. IEEE Computer Society, 2017, pp. 1907–1914.

[73] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in *ICASSP*. IEEE, 2019, pp. 8261–8265.

[74] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep Learning for Deepfakes Creation and Detection," *CoRR*, vol. abs/1909.11573, 2019.

[75] L. Verdoliva, "Media Forensics and Deepfakes: an overview," *CoRR*, vol. abs/2001.06564, 2020.

[76] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton-Ferrer, "The Deepfake Detection Challenge (DFDC) preview dataset," *CoRR*, vol. abs/1910.08854, 2019.

[77] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A New Dataset for DeepFake Forensics," *CoRR*, vol. abs/1909.12962, 2019.

[78] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," *CoRR*, vol. abs/1901.08971, 2019.

[79] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Github of the FaceForensics dataset," https://github.com/ondyari/FaceForensics, 2019.

[80] M. Kowalski, "3D face swapping implemented in Python," https://github.com/MarekKowalski/FaceSwap, 2018.

[81] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 66:1–66:12, 2019.

[82] D. E. King, "Dlib-ml: A Machine Learning Toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[83] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *AISec*. ACM, 2011, pp. 43–58.

[84] P. Neekhara, S. Hussain, M. Jere, F. Koushanfar, and J. McAuley, "Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples," 2020.

[85] iperov, "DeepFaceLab is the leading software for creating deepfakes." https://github.com/iperov/DeepFaceLab, 2019.

Sesión de Investigación B1: Inteligencia Artificial en ciberseguridad
http://doi.org/10.18239/jornadas_2021.34.38

A review of: Effect of the Sampling of a Dataset in the Hyperparameter Optimization Phase over the Efficiency of a Machine Learning Algorithm

David Escudero García
Research Institute on Applied Sciences in Cybersecurity
Campus de Vegazana s/n, 24071 León, Spain
descg@unileon.es
ORCID:0000-0002-3776-3920

Noemí DeCastro-García, Ángel Luis Muñoz Castañeda, Miguel V. Carriegos
Universidad de León
Campus de Vegazana s/n, 24071 León, Spain
{ncasg, amunc, mcarv}@unileon.es
ORCID:{0000-0002-5610-0153, 0000-0001-6993-9110, 0000-0002-6850-0277}

***Abstract*—Cybersecurity is a discipline in which artificial intelligence techniques are gaining in importance in order to obtain actionable knowledge. The selection of a fitting configuration of hyperparameters is an important factor in model performance. Several hyperparameter optimization algoritms have been developed, but their application imposes additional computational costs. Since one of the main factors in the resource consumption is the dataset size, we perform a study of the effect of using different partitions over five different cybersecurity datasets. Nonparametric inference has been used to measure the rate of change of the accuracy, time, and spatial (memory) complexity along the partition size. In addition, a level of gain is assigned to each partition allowing us to study patterns and determine the optimal partition size.**

***Index Terms*—Cybersecurity, Machine learning, Hyperparameter optimization**

Contribution type: *Published research. "Effect of the Sampling of a Dataset in the Hyperparameter Optimization Phase over the Efficiency of a Machine Learning Algorithm", Complexity*

I. Introduction

There are many applications of Machine Learning (ML) solutions to cybersecurity problems [1] in different domains such as network intrusion detection, malware classification and others. Tasks such as detection of anomalous activity in networks need to process large amounts of data and to make predictions in limited amounts of time. Therefore it is important to achieve a satisfactory predictive performance with limited computational costs. Furthermore, there is a lack of suitable datasets [2] and models should perform adequately with a limited amount of data.

Hyperparameter optimization tackles these problems because it improves model performance on available data. However, it is also a computationally expensive operation because it requires to train and evaluate several models. Therefore, to apply the model in high throughput systems it is necessary to find some method to increase the efficiency of the hyperparameter optimization process. In order to reduce the added computational costs, some studies consider applying diverse transformations on the data [3], but these transformations impose additional costs. The work presented in [4] remarks that intelligent sampling from the training data may provide results comparable to those of using the complete training set without any preprocessing. An effective model may be constructed with minimum computational cost by following this approach.

We applied the ideas in [4] and carried out an empirical statistical analysis of the effect of the used dataset size in the hyperperameter optimization phase in the performance of machine learning algorithms [5]. The underlying idea is that using a sample of the training set to optimize hyperparameters may yield satisfactory model performance and a reduction in computational costs (both time and memory) across different datasets, models and hyperparameter optimization algorithms.

We focused our research in problems which are characteristic of cybersecurity applications but the experiments sought to improve the overall efficiency of machine learning processes. The findings in this research have been applied on other areas that make use of machine learning algorithms [6], such as detection of structural health of wind turbines [7], in which a smaller partition size is used to speed up hyperparameter tuning of deep neural networks.

II. Experiment

In *Table I* we show the five different public cybersecurity datasets that have been used in the study. They tackle different problems: spam, attacks on autonomous robots, phishing, detection of network intrusions and of credit card fraud. These datasets have different size and class balance in order to study the results on different kinds of datasets.

Table I: Datasets used in the experiment

Dataset name	Instances	Features	Classes	Majority class size
Spambase	4601	57	2	60%
Robots	6422	12	3	73%
Phishing	11055	30	2	56%
Intrusion	148517	39	5	52%
Credit card	284807	30	2	99.8%

There are different kinds of hyperparameter optimization algorithms (HPO). Decision theoretic methods such as random search or the Nelder-Mead simplex only require simple model evaluations and are easily parallelizable, but may get stuck on local optima. Bayesian methods are more robust when optimizing many hyperparameters, but are more computationally expensive because they construct a model that optimizes

the selection of new configurations to test. We include six different HPO, both bayesian and decision theoretic, to test whether there are significant differences in their behaviour when using a partition of the dataset for hyperparameter tuning.

For each dataset D_i four partitions $P_j(D_i), j = 1, ...4$ (proportions correspond to $P_1 = 1/12, P_2 = 1/6, P_3 = 1/2, P_4 = 1$) are randomly sampled. For each $P_j(D_i)$ we apply an HPO and test its output configuration on D_i with a 80/20 split for training and test set respectively. This process is repeated 50 times for each combination of $P_j(D_i)$, model and HPO in order to analyse if the results are statistically significant. For each trial, we store the accuracy (Acc) of the resulting model, the time spent by the HPO (TC) and the maximum amount of memory allocated during the process (SC). First, we check the normality of the results. They are not normal, therefore we use Wilcoxon's test to determine whether there are significant differences in these metrics between the different partitions.

In order to measure the level of efficiency of using a partition of the dataset in the hyperparameter optimization phase we define different levels of gain based on the differences in Acc, TC and SC for each partition. We define 9 levels of gain based on the relationship between the increase or decrease in accuracy and its corresponding impact on execution time and memory consumption. The greater the gain level, the better the result. A gain level of 5 or greater signifies that it is more efficient to use a partition of the original dataset.

III. Results

As shown in *Table II*, there are statistically significant differences in the three metrics for all HPO and models. The differences in time and memory are to be expected, but the amount of significant differences in accuracy is relatively low, reaching only 20% with P_3. Accuracy, memory and time increase with the partition size. The magnitude and rate of the changes is significantly lower for accuracy than for memory and time.

Table II: Average rate of statistically significant differences for all HPO, D_i and model

Partitions	Accuracy	Time	Memory
P_1 vs P_2	39,9%	97.77%	100%
P_2 vs P_3	41.1%	95.55%	98.8%
P_3 vs P_4	19.99%	92.21%	100%

Given that there are significant differences, we examine the global average gain levels in *Table III*.

Table III: Average rates of gain

Partitions	HPO	Dataset
P_1 vs P_2	5.04	5.04
P_2 vs P_3	5.44	5.475
P_3 vs P_4	6.33	6.175

The average level of gain for all partitions is greater than the threshold of 5, which means that it is more efficient to use a reduced subset of the data to perform hyperparameter optimization. The same applies to the different HPOs. The improvement is greater in bayesian methods because they have lower scalability.

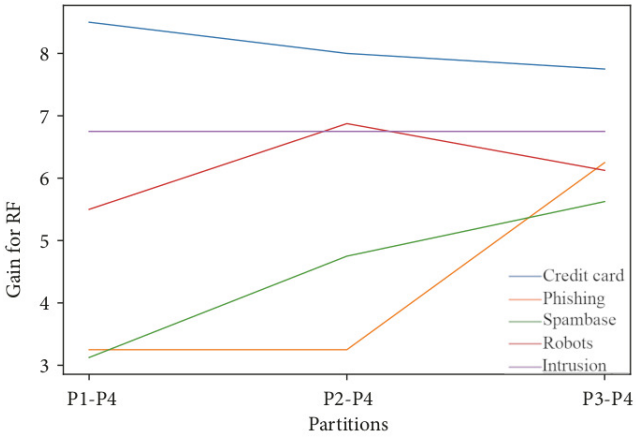


Figure 1: Levels of gain of each dataset

In *Fig. 1* it can be seen that some datasets show a worse performance than average. In these cases, the size of the partition must be higher to prevent degradation of model performance. However, a level of gain greater than 5 is achieved for all datasets when using P_3.

IV. Conclusion

In the original work, we analysed of the impact of using smaller samples of the dataset for the process of hyperparameter optimization, and its influence on the effectiveness of the model. The study was carried out over five different public cybersecurity datasets. The results let us conclude that working with smaller partitions is more efficient than performing the same process with the whole dataset. However, depending on the dataset, the size of the used partition may need to be greater to prevent the model performance from degrading.

Acknowledgements

The authors would like to thank the Spanish National Cybersecurity Institute (INCIBE), who partially supported this work under contract art. 83 key X54.

References

[1] A. Shenfield, D. Day, and A. Ayesh, "Intelligent intrusion detection systems using artificial neural networks" *ICT Express*, vol. 4, no. 2, pp. 95–99, 2018.

[2] Y. Xin et al., "Machine Learning and Deep Learning Methods for Cybersecurity", in *IEEE Access*, vol. 6, pp. 35365-35381, 2018.

[3] Z. Cheng and Z. Lu, "A Novel Efficient Feature Dimensionality Reduction Method and Its Application in Engineering", *Complexity*, vol. 2018, Article ID 2879640, 14 pages, 2018.

[4] F. Provost, D. Jensen, and T. Oates, "Efficient progressive sampling" in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 23–32, 1999.

[5] N. DeCastro-García, A.L. Muñoz Castañeda, D. Escudero García and M.V. Carriegos, "Effect of the Sampling of a Dataset in the Hyperparameter Optimization Phase over the Efficiency of a Machine Learning Algorithm", *Complexity*, vol. 2019, Article ID 6278908, 17 pages, 2019.

[6] Li Yang, Abdallah Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice", *Neurocomputing*, vol. 415, pp. 295-316, 2020.

[7] Puruncajas B, Vidal Y and Tutivén C., "Vibration-Response-Only Structural Health Monitoring for Offshore Wind Turbine Jacket Foundations via Convolutional Neural Networks". *Sensors (Basel)*, vol. 20, 19 pages, 2020.

http://doi.org/10.18239/jornadas_2021.34.39

HOUSE: Marco de trabajo modular de arquitectura escalable y desacoplada para el uso de técnicas de *fuzzing* en HPC

Francisco Borja Garnelo Del Río, Francisco J. Rodríguez Lera, Gonzalo Esteban Costales, Camino Fernández Llamas, Vicente Matellán Olivera
Universidad de León - Campus de Vegazana s/n, 24071 León (Spain)
infbgd01@estudiantes.unileon.es, {fjrodl, gestc, cferll, vmato}@unileon.es

***Resumen*—**

El *fuzzing* es una técnica de prueba automatizada que hace uso de mutaciones de entradas para ejecutar el software utilizando estas a fin de observar los valores de retorno y el estado externo del sistema; todo ello para identificar comportamientos no esperados. Por su simpleza y fácil automatización inicial, el *fuzzing* es una de las técnicas más populares para identificar vulnerabilidades en software en el mundo real. El uso de técnicas de *fuzzing* como parte del desarrollo de software en grandes compañías ha acelerado la evolución de las técnicas y mejorado las herramientas, como contrapartida muchos proyectos quedan huérfanos de sus desarrolladores originales, al ser captados por estas empresas y todas aquellas características con potencial uso comercial, son restringidas al ámbito privado o son dependientes de nubes propietarias. Esto supone una barrera inicial para nuevos desarrollos, una posible dependencia tecnológica de terceros y problemas de confidencialidad de los datos para adoptar modelos seguros de desarrollo o realizar investigaciones. Las investigaciones sobre *fuzzing* y las nuevas herramientas tienen en común partir desde cero o una base teórica y focalizarse en una característica o funcionalidad. Por todo lo anterior resulta valioso desarrollar un marco de trabajo común que de coherencia y continuidad de forma global abordando el *fuzzing* como un flujo con distintas fases, organizando todos los procesos implicados, incluyendo también aquellos con relación directa, de forma modular, abierta y agnóstica a las herramientas y técnicas. Con ello se potencia la reutilización y se elimina la dependencia a herramientas o casos de uso, a la vez que permiten sinergias con otros enfoques de seguridad en el software. El uso de una arquitectura desacoplada y escalable permite la distribución y paralelización de trabajos. Esta arquitectura es ideal para entornos de computación de alto rendimiento en adelante HPC (high-performance computing por su traducción al inglés) o basados en cargas de trabajo. El código fuente y la documentación empleada para la prueba de concepto se encuentra disponible públicamente en GitHub.

***Index Terms*—Desarrollo seguro, marco de trabajo, *fuzzing*, análisis de datos, adaptación de código, HPC, seguridad del software, automatización pruebas software, SDL, AFL++, Slurm**

I. Introducción

El *fuzzing* [1] es la automatización de generación y prueba de entradas malformadas en el software con el fin de encontrar comportamientos no esperados en el mismo, habitualmente *crashes*[1]. Este término se usó por primera vez en un estudio de la seguridad de UNIX en la década de 1990 [3], [4].

Hace poco más de una década, las herramientas de *fuzzing* eran poco más que generadores de ruido aleatorio. En la actualidad hay nuevas técnicas capaces de entender en mayor medida la semántica y los flujos de ejecución del software de forma automatizada. Algunos ejemplos son la generación de firmas [5], la ejecución simbólica dinámica [6], las pruebas de complejidad [7], la generación de casos de prueba basados en gramáticas [8] y las pruebas de comportamiento [9].

Las últimas investigaciones han avanzado mucho en las capacidades de generación de entradas y en el análisis de las estructuras internas del software; ambas con el fin de evolucionar características o funcionalidades de las herramientas de *fuzzing*. Gran parte del éxito de estas investigaciones ha venido de la mano de aplicar los últimos avances en el análisis de datos [10] y algoritmos de aprendizaje automatizado [11] o de su uso específico para la búsqueda de vulnerabilidades[2] de seguridad [13]. Como respuesta a su uso como herramienta de búsqueda de vulnerabilidades, también están surgiendo técnicas *antifuzzing* [14].

Respecto a su uso a gran escala fuera del ámbito privado, se han realizado experimentos de su uso en HPC [15] y es posible replicar experimentos parecidos en entornos cloud compatibles con las herramientas y arquitecturas HPC como AWS [16], Azure [17] o GCS [18]. No obstante, el *fuzzing* no está exento de contratiempos a la hora de su uso o aplicación.

La complejidad y la vasta cantidad de información existente sobre el *fuzzing* es el primer problema que se presenta. Esto suele derivar en que muchos de estos avances son proyectos separados que sólo son implementados como pruebas de concepto de forma aislada, sin ser integrados con otras mejoras, u optimizaciones de forma general, que acaban siendo descontinuados o parcialmente integrados en herramientas *open source* [19].

[1]***crash*** [2] Fallo repentino y completo de un sistema o componente informático.

[2]**Vulnerabilidad** [12] La existencia de una debilidad, diseño o error de implementación que puede llevar a un evento inesperado e indeseable que comprometa la seguridad del sistema informático, red, aplicación o protocolo involucrados.

El segundo problema se debe a la monopolización de las herramientas e investigaciones hacia objetivos comerciales o de uso interno por parte de la industria u organizaciones. Esto es especialmente relevante en aquellos proyectos orientados a utilizar técnicas de *fuzzing* a gran escala o como parte de un modelo de desarrollo seguro. Tales proyectos presentan serias limitaciones para ser adaptados a distintos entornos diferentes de aquellos para los que fueron creados; ya sea debido a su origen privativo [20], [21] o a su enfoque comercial [22], [23].

Esto, en caso de productos comerciales, presenta serias limitaciones de cara a la privacidad dado que implica el acceso al código fuente [3] y el uso de *Software as a Service* (SaaS) en entornos de nube en alguna de sus modalidades. Respecto a aquellos proyectos de origen privativo, las limitaciones suelen venir por la dificultad de aprendizaje de la solución y por su adaptación a un entorno diferente, con esfuerzos reiterados en cada nueva evolución de la solución original.

La tendencia generalizada a utilizar infraestructuras en la nube promovidas en gran parte por las mismas organizaciones que derivan el desarrollo de las herramientas, en menor o mayor medida, una pérdida de autonomía y confidencialidad de los datos. Sirva de ejemplo un proveedor de nube específico [24] o la toma de decisiones de forma unilateral por el prestador de servicio [25].

Un tercer problema es debido a la dificultad y especialización necesarias para poner en marcha el *fuzzing* de forma que este aporte valor sin un esfuerzo recurrente y pueda aplicar los resultados de forma efectiva o reutilizar los esfuerzos previos realizados para casos similares.

Realizar pruebas de *fuzzing* en software es una tarea compleja puesto que requiere preparar [26] el código fuente para facilitar la automatización de entradas. Los protocolos de red o los componentes del núcleo de los sistemas operativos son dos ejemplos de software que necesitan una preparación previa para lograr obtener resultados efectivos.

En los casos donde la entrada de datos presente una complejidad elevada (como por ejemplo reproductores multimedia, intérpretes de lenguajes o protocolos de comunicaciones) implica un proceso previo de generación de entradas [26]. Para ello es necesario preparar las entradas iniciales o semillas utilizando ficheros con estructuras complejas o programables, como pueden ser los documentos XML, JavaScript, etc.

De forma generalizada a todo proceso de *fuzzing* es necesaria una especialización para el aprovechamiento de los resultados para por ejemplo identificar puntos de mejora de seguridad[4]o en la calidad del software.

Estos problemas pueden ser abordados sobre una estructura ordenada y orientada a las fases comunes del *fuzzing*, con un diseño modular de componentes que permita la reutilización e integración de distintas soluciones. El marco de trabajo HOUSE que se presenta en este trabajo propone una aproximación que permite mantener la confidencialidad del código fuente, garantizando la escalabilidad y funcionamiento independiente de cada uno de sus módulos para evitar esfuerzos recurrentes. Además, facilita la investigación de cada uno de los componentes que forman parte del proceso de *fuzzing* y la adopción de estas técnicas en proyectos de desarrollo. Esto reduce la barrera inicial y además proporciona una estructura modular consistente que permite simplificar las tareas de integración, preparación-ejecución de pruebas y análisis de resultados. Debido a que brinda una separación funcional, con automatizaciones para su puesta en funcionamiento e integrado en una organización superior preparada para la paralelización y reutilización.

Los principales objetivos de esta investigación son:

1. Conocer el estado actual del estado de la cuestión del fuzzing enfocándolo a su uso en entornos computación de alto rendimiento.
2. Dar una visión global de las actividades previas, los procesos y componentes del fuzzing respecto a las últimas mejoras y las adaptaciones a entornos de computación escalable.

El resto del artículo se estructura de la siguiente manera. La Sección II relaciona las distintas disciplinas que convergen o tienen un punto común con el *fuzzing* (como son el desarrollo seguro de software y la seguridad), así como conceptos clave relacionados o que forman parte de su evolución. La Sección III desarrolla el flujo de trabajo en sus distintas fases atendiendo a los distintos procesos implicados. La Sección IV expone la experiencia de uso del marco de trabajo HOUSE en un entorno de supercomputación real. Por último, la Sección V detalla las conclusiones del trabajo. La documentación de uso del Marco de trabajo HOUSE esta disponible el repositorio de GitHub [27].

II. Contexto y trabajos relacionados

Con la evolución en complejidad y extensión de los sistemas informáticos y de comunicaciones, el software que los gobierna cuenta cada vez con más superficie susceptible de fallos, lo que a su vez supone una mayor dificultad para su búsqueda y prueba.

II-A. Tipos de pruebas software

De una forma sencilla podemos reducir los tipos de pruebas realizadas al software en activas o dinámicas y pasivas o estáticas. Activas son aquellas pruebas que implican la ejecución total o parcial del software y las pasivas se centran en analizar el código fuente.

Dentro de las activas o dinámicas, atendiendo a un enfoque puramente de seguridad en el software [28], su identificación puede ser proactiva o reactiva.

Según nuestras evidencias, aparentemente no hay alternativas en la literatura al *fuzzing* desde un enfoque proactivo. El objetivo del *fuzzing* es identificar por medio de automatizaciones comportamientos no esperados en el software

[3]**Código fuente** [2]: Instrucciones y definiciones de datos del ordenador expresadas en una forma adecuada para su introducción en un ensamblador, compilador u otro traductor. También considerada la versión legible para el ser humano de la lista de instrucciones [programa] que hace que un ordenador realice una tarea.

[4]**Seguridad** [12] Todos los aspectos relacionados con la definición, adquisición y mantenimiento de la confidencialidad de los datos, la integridad, la disponibilidad y la responsabilidad.

habitualmente en forma de *crashes*. Algunos de estos fallos en el software pueden ser utilizados para comprometer la seguridad del sistema donde este es ejecutado o la información a la que tiene acceso. Debido a la automatización, es necesario acotar las pruebas a un método de entrada de datos específico en el software y limitarlas a un conjunto de funcionalidades que interactúen con las entradas de la manera más determinista posible. Ejemplos de *fuzzers* son herramientas como BFF [29], radamsa [30] o AFL [31].

Observando el comportamiento del software en un entorno de producción, es posible identificar de manera reactiva comportamientos no esperados. Estos por ejemplo pueden ser controles basados en contexto, como Security-Enhanced Linux *(SELinux)* [32] o Application Armor *(AppArmor)* [33], para distribuciones Linux o soluciones comerciales de seguridad basadas en la monitorización de eventos y protecciones [34], [35] siendo las más comunes combinaciones de Endpoint Detection and Response *(EDR)* y Endpoint Protection Platform *(EPP)*.

Existen numerosas herramientas públicas para realizar pruebas automatizadas sobre el software de manera pasiva utilizando su código fuente. Como por ejemplo CodeQL [36] centrada en facilitar el análisis semántico del código fuente como si se tratase de consultas a una base de datos, u otras herramientas más tradicionales como los analizadores de código Sonarqube, Yasca o Flawfinder [37].

Para realizar análisis más completos es necesario combinar distintos enfoques como el proactivo por medio de *fuzzing* y el pasivo con herramientas que permitan minimizar o priorizar qué funcionalidades son susceptibles a tener más bugs[5][6]. Este es el caso cuando por complejidad no sea posible realizar pruebas con cobertura completa del alcance o cuando se requieren adaptaciones a medida para permitir la integración de las herramientas de pruebas con el software objetivo. La misma problemática se repite en el caso de querer utilizar infraestructuras de cómputo intensivo como supercomputadores o clusters en la nube [17].

Respecto a los datos utilizados en el proceso, hay algunos componentes de generación de datos de entrada, limpieza de falsos positivos y análisis de resultados, la mayoría teóricos o con una implementación muy dirigida.

Este escenario tan heterogéneo repleto de piezas aisladas dispares o rígidos componentes, hace difícil introducir de forma mantenible en el tiempo y de manera eficiente pruebas de *fuzzing* en el desarrollo y ralentiza las investigaciones de mejoras de los componentes del proceso.

II-B. Tipos de fuzzing

Los tipos de *fuzzing* se clasifican en función de los detalles de la implementación disponible, estando claramente diferenciados dos tipos por la utilización del código fuente para la realización de las pruebas o el uso de software ya compilado en formato binario [28]. Estos dos tipos son el *white-box* fuzzing [39] y el *black-box* fuzzing [40] respectivamente.

Entre estos dos tipos se situaría un tercero, el *grey-box* fuzzing [26], que tiene características de ambos, como es el caso de las pruebas donde se utiliza instrumentación en el código fuente[41] y el análisis BVA (*Boundary Value Analysis*) [42] sobre los comportamientos observados en las ejecuciones del binario.

En la actualidad hay múltiples proyectos alrededor del *fuzzing*, desde el análisis de resultados [43], creación de conjuntos de datos de entrada de forma automatizada [44], aquellos específicos de tareas internas como optimizaciones [45] y creación de algoritmos de prueba [46]. Todos ellos necesitan, para poder ser completados de forma práctica, unas condiciones que en la mayoría de los casos exigen un esfuerzo de creación de un entorno completo de *fuzzing*. Similar caso sucede a desarrolladores que demanden poder realizar pruebas de calidad o seguridad en su software y quieran mantener la confidencialidad de su código fuente y evitar los esfuerzos de puesta en marcha de un entorno de *fuzzing* y realizar desde cero su integración o adaptación.

Las pruebas proactivas y activas con fines de calidad centradas en la seguridad [28] como el *fuzzing* están cada vez más presentes en los ciclos de vida de desarrollo de software SDLC (*Software Development Life Cycle)*, siendo parte fundamental de los procesos para la implantación de un modelo de desarrollo seguro SSDLC (*Secure Software Life Cycle)* [47].

III. Marco de trabajo HOUSE

HOUSE es un marco de trabajo con un enfoque funcional que permite organizar todas las actividades necesarias de cara a aplicar técnicas *fuzzing* sobre un proyecto software. Su nombre hace referencia a la clásica metáfora [48] de describir un proceso como el diseño y construcción de una casa y, en este caso, describe el proceso de desarrollo seguro del software pero desde el punto de vista del *fuzzing*.

En cuanto a sus características, HOUSE posee una arquitectura modular e interoperable cuyo funcionamiento se basa en aplicar el *fuzzing* a un proyecto software a través de un proceso compuesto por 4 fases, (ver Fig. 1):

1. **Común o de preparación del entorno de trabajo**. Es la única fase transversal a todo el proceso ya que se suele realizar una única vez, bien en la puesta en marcha de las herramientas necesarias para el *fuzzing* o bien durante el despliegue inicial del marco de trabajo.
2. ***Prefuzzing* o de preparación de las pruebas software**. En esta fase se elige el tipo de *fuzzing* a utilizar en función del grado de acceso y las necesidades de adaptación del código fuente. También admite los requisitos o preferencias del tipo de pruebas e incluso los resultados que se persigan con él. En función del tipo elegido, se preparan los binarios correspondientes: empleando binarios ya compilados o realizando una compilación

[5]**Bug** [2] Anomalía, defecto, error, excepción o fallo no controlado en un software.

[6]**Error** [38] La diferencia o discrepancia entre un valor o condición calculado, observado o medido y el valor o condición verdadero, establecido o teóricamente correcto.

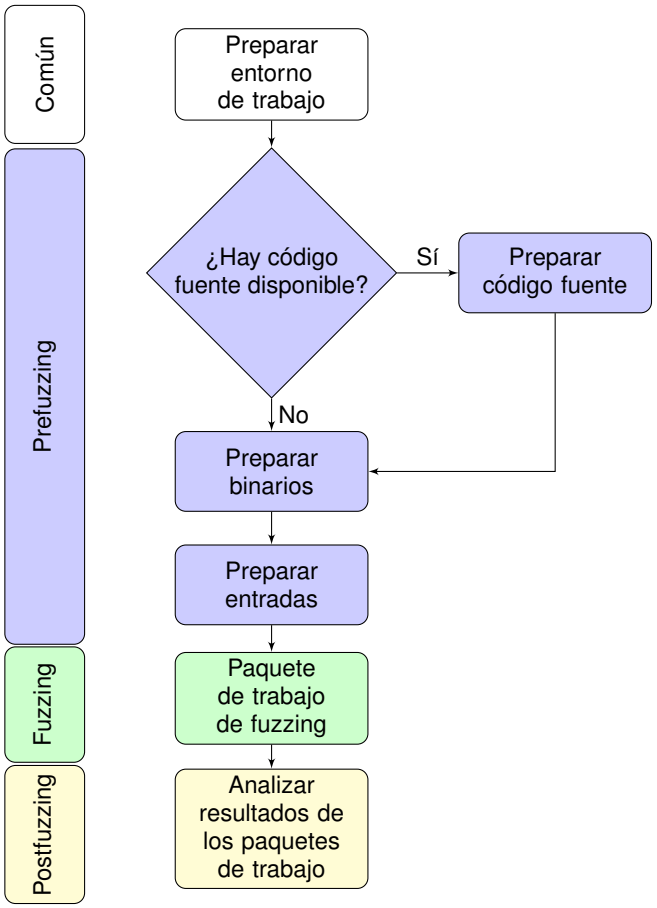


Figura 1. Diagrama de flujo para describir el funcionamiento de HOUSE según las diferentes fases de fuzzing.

específica si es necesario. En función del binario seleccionado, también se confecciona o valida un conjunto de entradas para esa campaña de *fuzzing* [26].

3. ***Fuzzing* o de ejecución de las propias pruebas**. Fase en la que se somete el software bajo prueba (PUT - *Program Under Test*) al *fuzzing* utilizando la configuración realizada previamente.
4. ***Postfuzzing*g o de análisis de resultados y tareas posteriores**. En esta última fase, bien si se desea o bien si por volumen de trabajo es necesario automatizar la tarea de análisis de resultados, es posible ejecutar un paquete de trabajo creado expresamente para poner en valor los resultados del *fuzzing*.

A su vez, dicho proceso trabaja sobre un conjunto de datos organizados en seis módulos, que pueden ser orquestados de manera escalable en un entorno de HPC por medio de cargas de trabajo agrupadas por paquetes:

1. ***TooL Box* (0.TLB)**. Colección de herramientas y utilidades auxiliares a cualquiera de las fases del *fuzzing*.
2. ***Source Code Integration* (1.SCI)**. Colección de códigos fuente, incluyendo parches, integraciones y optimizaciones previas al *fuzzing*.
3. ***Binaries Ready to Fuzz* (2.BRF)**. Colección de binarios preparada para el *fuzzing*.
4. ***WorkLoad Package* (3.WLP)**. Colección de paquetes con cargas de trabajo[7], que pueden ser de cualquier fase del *fuzzing*, como herramientas que analicen el código fuente para identificar áreas más propensas a fallos, tareas propias de *fuzzing* o de analítica de resultados.
5. ***Input Repository* (A.IR)**. Repositorio de entradas como semillas para generadores de entradas, diccionarios, archivos de muestra, etc.
6. ***Output Repository* (B.OR)**. Repositorio de salidas con los contextos de ejecuciones de los *fuzzers*, resultados y cualquier información útil generada por el flujo de trabajo como logs.

Cabe señalar que cada módulo puede estar formado por componentes, unidades funcionales o una combinación de ambos. La principal diferencia entre dichos elementos radica en que las unidades funcionales realizan tareas de automatización o ejecución y los componentes son solo información. Por otro lado, los módulos identificados por letras (como **A.IR** y **B.OR**) están formados únicamente por datos, mientras que los identificados por números (**0.TLB**, **1.SCI**, **2.BRF** y **3.WLP**) tienen al menos una actividad que forma parte del flujo de *fuzzing*. En relación a estos últimos, el propio número indica su grado de dependencia dentro del proceso del *fuzzing*; siendo 0 la más baja y 3 la más alta.

Teniendo en cuenta todo lo anterior, esta estructura de fases junto a su modularización hacen que HOUSE promueva la reutilización de las diferentes partes del proceso de *fuzzing*; permitiendo así su escalabilidad o paralelización. Adicionalmente, dicha estructura también permite que el marco de trabajo sea agnóstico a las herramientas *fuzzing* empleadas. Por ejemplo, es posible cambiar el *fuzzer* utilizado sin tener que rehacer las correspondientes adaptaciones del código fuente, sin tener que preparar nuevos binarios o, incluso, sin tener que parar los trabajos en curso. De igual modo, también permite la reutilización de los conjuntos de datos de entrada creados para un proyecto concreto o el hecho de combinarlos con herramientas que los generen desde archivos de muestra.

III-A. Flujo de trabajo y procesos asociados

La Fig. 2 representa el flujo de trabajo de HOUSE utilizando el lenguaje de modelado *Business Process Model and Notation* (BPMN) en su versión 2.0.2 [49]. La utilización de BPMN responde a la necesidad de modelar de extremo a extremo todo el flujo de *fuzzing* al detalle suficiente. De haber utilizado el lenguaje UML *(Unified Modeling Language)*, el resultado hubiese sido menos preciso y más complejo al tener menos objetos representables y estar centrado en los flujos de datos.

Dicho lo anterior, en las columnas del diagrama BPMN de la Fig. 2 se sitúan los módulos de HOUSE y en las filas sus fases. Los siguientes apartados detallan individualmente esas fases dentro de cada módulo del marco de trabajo.

[7]**Carga de trabajo** [2]: Combinación de tareas que se ejecutan en un sistema informático determinado. Sus principales características es incluir los requisitos de entrada y salida, cantidad, tipo de calculo y computo requerido.

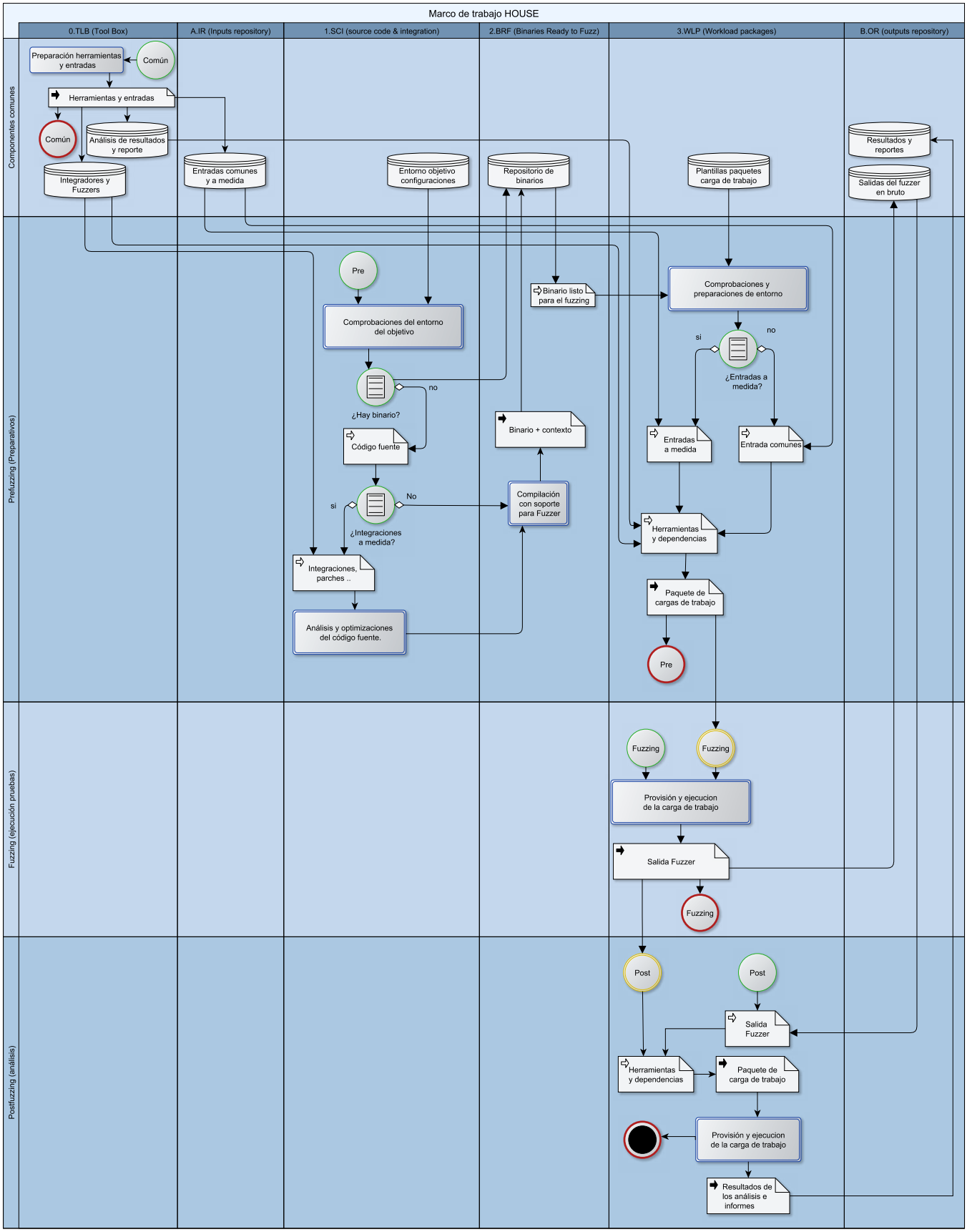


Figura 2. Diagrama BPMN del marco de trabajo HOUSE.

III-A1. Común: Esta fase aporta el esqueleto necesario para modelar el proceso de *fuzzing* dentro del marco de trabajo HOUSE. También define los elementos esenciales para las propias tareas de *fuzzing* o las adyacentes (como por ejemplo los analizadores de código, los generadores de entradas o el reporte de resultados). Dicho de otra manera, esta fase es responsable de habilitar—entre otros elementos—cualquier herramienta, librería o conjunto de datos de entradas que se vayan a utilizar en el entorno de ejecución de las pruebas. Todo esto se realiza tanto en el módulo **0.TLB** (para el caso de herramientas tales como *fuzzers*, analizadores u optimizadores de código fuente, generadores de conjuntos de datos de entrada, etc) como en el módulo **A.IR** (para los diccionarios, muestras de ficheros o conjuntos de datos utilizados).

III-A2. Prefuzzing: Todas aquellas tareas previas al *fuzzing* tales como preparativos, configuraciones, provisiones de muestras o conjuntos de datos forman parte de esta fase. Según el tipo de *fuzzing* a realizar, las tareas pueden comenzar en el módulo **1.SCI** o en el **2.BRF**.

Por un lado, el primer caso se da cuando el tipo de *fuzzing* seleccionado es *white-box* o *grey-box*; en función de la dependencia de las pruebas del código fuente. En este caso, el módulo es el encargado de realizar las diferentes actividades necesarias para preparar el código fuente de cara a su compilación y atendiendo a las necesidades de alcance, estrategia, requisitos y herramientas elegidas para ese flujo. Por otro lado, el segundo caso se da cuando el tipo de *fuzzing* seleccionado es *black-box* y, en cuyo caso, el módulo **1.SCI** no está implicado. Usándose directamente el módulo **2.BRF**.

Así mismo, el módulo **2.BRF** es común a cualquier flujo de *fuzzing* independientemente del tipo empleado. La finalidad de este módulo es realizar el almacenamiento del binario objetivo del *fuzzing* y el código utilizado para su generación si está disponible; llevando a cabo la compilación del código fuente previamente adecuado dentro del módulo **1.SCI**. Más aún, en el módulo **2.BRF** se realizan, si procede, las comprobaciones necesarias para verificar que los binarios son funcionales y también la alineación con la estrategia y requisitos planteados.

Cabe señalar que un punto común entre los módulos **1.SCI** y **2.BRF** es el código fuente utilizado en alguno de los tipos de *fuzzing*. No obstante, existe una diferencia fundamental entre ambos y es el contexto y la finalidad de este último. En el caso de **1.SCI**, el objetivo es trabajar con el código fuente original del software y realizar las transformaciones necesarias para el *fuzzing*. Por su parte, en el caso de **2.BRF** se utiliza un código modificado para un contexto específico de *fuzzing*. Al mismo tiempo, a nivel de almacenamiento también hay diferencias. Mientras que en **1.SCI** se asume que el código fuente es una única dependencia asociada a una versión de software, **2.BRF** mantiene un repositorio con el código modificado y los binarios asociados a este.

También se debe agregar que en esta fase se confecciona el paquete de trabajo para el *fuzzing* o para cualquier tarea de análisis o generación de dependencias para este, en el módulo **3.WLP**; en función del tipo de herramienta, técnica empleada y entorno de ejecución de las pruebas. Si es necesario, también se alinean los conjuntos de datos del repositorio **A.IR**.

Otros posibles usos de los paquetes de trabajo con el fin de ejecutar las tareas de forma escalable o paralelizable son: Los de análisis de código estático para preparar la estrategia de *fuzzing*, el filtrado de resultados y el uso de generadores de entradas.

III-A3. fuzzing*:* Esta fase, común a todos los tipos de *fuzzing*, comprende únicamente el módulo **3.WLP**. En ella se realizan las tareas de provisión del paquete de trabajo, el seguimiento de su ejecución y el volcado de los resultados en el repositorio de datos de salida en crudo **B.OR**. Un ejemplo de tarea realizada en esta fase dentro del componente **3.WLP** sería la gestión de colas de trabajo asociadas a un paquete.

III-A4. Postfuzzing: Esta última fase se realiza en el módulo **3.WLP**. Se centra en el procesamiento y análisis de los resultados de los paquetes de trabajo previos de *fuzzing*; ultimando para ello el repositorio **B.OR**. Ejemplos de tareas pertenecientes a esta fase podrían ser el uso de herramientas de filtrado de falsos positivos, la generación de reportes, etc.

IV. Prueba de Concepto

El marco de trabajo HOUSE es *open source* y está disponible públicamente en GitHub [27]. De cara a validar su propuesta, se está llevando a cabo una prueba de concepto en el entorno de ejecución del HPC Caléndula [50]. Este entorno utiliza Slurm [51] como sistema de gestión de cargas y las arquitecturas Intel, Cascade Lake y Haswell, sobre el sistema operativo CentOS 7.7. Calendula cumple con las políticas y requisitos del Esquema nacional de seguridad [52].

Cabe destacar que para la prueba de concepto no se ha utilizado ningún sistema de contenedores en ninguno de los módulos, ni se aplican excepción de políticas de seguridad o uso de permisos especiales.

La prueba de concepto consiste en desarrollar diversos scripts para Slurm e implementar casos de uso básicos con el *fuzzer* AFL++ [53], un proyecto *open source* que integra las ultimas técnicas de *fuzzing* y continúa el proyecto discontinuado de AFL. El tipo de *fuzzing* elegido para esta prueba fue el *white-box*, utilizando la cobertura de código sobre los binarios `date` y `expr` de la colección de utilidades coreutils, perteneciente a los sistemas GNU [54]. También se incluyó un caso básico de Buffer Overflow. En relación a la entrada de datos, en ambos casos se utilizó la entrada estándar.

Desde el inicio de las actividades se han resuelto distintos problemas derivados de la naturaleza del HPC Calendula, un entorno multipropósito y multiusuario con requisitos de seguridad y privacidad elevados que condicionan el desarrollo.

Los principales problemas identificados son el uso de nodos compartidos, lo que imposibilita relajar las protecciones y mecanismo de monitorización existentes. Destacan las limitaciones de uso de herramientas basadas en ASAN [55] o la monitorización de *crashes* usando ABRT [56], ya sea para evitar bloquear recursos comunes de los nodos o por su uso como parte de las protecciones de los sistemas. Esto se soluciona utilizando otras estrategias de detección basadas totalmente en instrumentación.

Inicialmente se trató de evitar complejidad usando la instrumentación mas básica del AFL++, afl-gcc similar a la del experimento, para reducir el número de dependencias y la complejidad para la prueba de concepto. Por los problemas anteriormente mencionados, se está trabajando en la adecuación de instrumentación apoyada en LLVM.

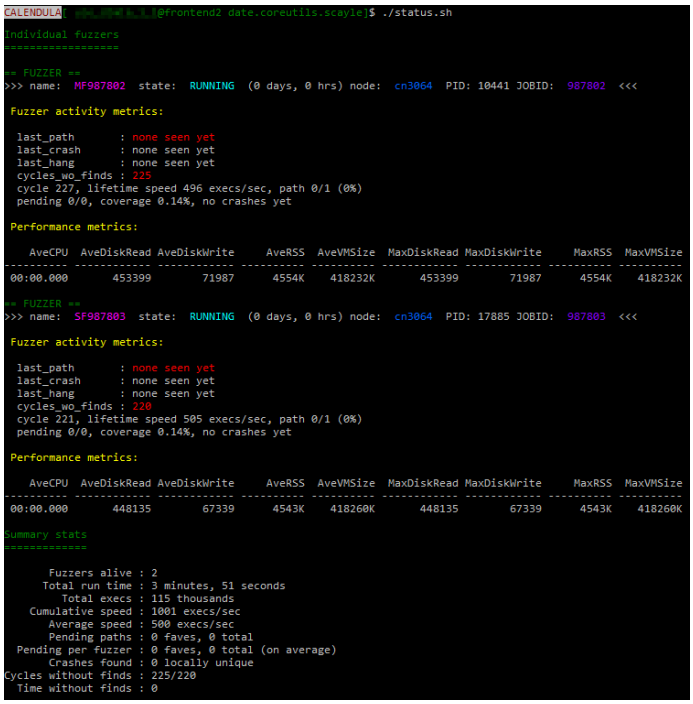


Figura 3. Ilustración status de los trabajos de fuzzing.

En la Fig. 3 se puede observar la ejecución de un paquete de trabajo de *fuzzing* utilizando dos trabajos y las métricas de rendimiento y actividad. Así como el sumatorio total de la actividad de todos los *fuzzers*.

V. Conclusiones

El *fuzzing* está cobrando cada vez más protagonismo como actividad para garantizar la calidad y la seguridad del software. Esta tendencia es utilizada por empresas y organizaciones para poner en valor sus desarrollos o servicios de forma comercial, a la vez que anima al desarrollo de mejoras y herramientas. Esta situación provoca la dependencia tecnológica y un riesgo a la confidencialidad para su adopción en un ciclo de desarrollo seguro y dificulta la puesta en marcha. Esta dependencia también frena la adopción de mejoras, el uso de alternativas totalmente libres y agnósticas de la tecnología.

HOUSE propone un modelado del flujo de trabajo del *fuzzing* caracterizado por tener un enfoque abierto y agnóstico a las herramientas. Con esta finalidad, el marco de trabajo propone una visión ordenada de todas las fases previas y posteriores a las pruebas. De esta manera, HOUSE permite resolver o simplificar las tareas más tediosas y complejas relacionadas con la aplicación de técnicas de *fuzzing*.

Este flujo se basa en los pasos y elementos establecidos en la Fig. 2.

Así pues, esta investigación sienta las bases teóricas comunes y las prácticas mínimas de un marco de trabajo que permite reducir el esfuerzo inicial a la hora de poner en marcha un entorno de *fuzzing* escalable que garantice la confidencialidad de la información. De igual modo, HOUSE facilita, tanto a investigadores como a desarrolladores, el hecho de poder trabajar en un componente o funcionalidad específica o el hecho de incluir aquellas características de seguridad proactiva en el software dentro de su ciclo de desarrollo. Estos aspectos se traducen en una reducción de la carga de trabajo en tareas auxiliares o distintas a las de los objetivos planificados.

Esta primera versión del marco de trabajo brinda la estructura de directorios, repositorios de datos y recursos para utilizar el AFL++ [53] con instrumentación básica, incluyendo las automatizaciones para la creación de cargas de trabajo, usando un sistema de colas Slurm [51].

Agradecimientos

Este trabajo ha sido parcialmente financiado por INCIBE mediante la Adenda 4 al convenio marco con la Universidad de León. Los autores agradecen al centro de supercomputación SCAYLE la posibilidad de utilizar sus infraestructuras para validar la propuesta presentada en este trabajo de investigación.

Referencias

[1] M. Sutton, A. Greene, and P. Amini, *Fuzzing: Brute Force Vulnerabilty Discovery*. Upper Saddle River, NJ: Addison-Wesley, 2007.

[2] IEEE Communications Society, "IEEE Standard Glossary of Software Engineering Terminology," *Office*, vol. 121990, no. 1, p. 1, 1990. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=159342

[3] B. P. Miller, L. Fredriksen, and B. So, "An empirical study of the reliability of UNIX utilities," *Communications of the ACM*, vol. 33, no. 12, pp. 32–44, Dec. 1990. [Online]. Available: https://dl.acm.org/doi/10.1145/96267.96279

[4] B. P. Miller, D. Koski, C. P. Lee, V. Maganty, R. Murthy, A. Natarajan, and J. Steidl, "Fuzz Revisited: A Re-examination of the Reliability of UNIX Utilities and Services," p. 23, 1995.

[5] B. Dolan-Gavitt, A. Srivastava, P. Traynor, and J. Giffin, "Robust signatures for kernel data structures," in *Proceedings of the 16th ACM Conference on Computer and Communications Security - CCS '09*. Chicago, Illinois, USA: ACM Press, 2009, p. 566. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1653662.1653730

[6] T. Xie, N. Tillmann, J. de Halleux, and W. Schulte, "Fitness-guided path exploration in dynamic symbolic execution," in *2009 IEEE/IFIP International Conference on Dependable Systems & Networks*. Lisbon, Portugal: IEEE, Jun. 2009, pp. 359–368. [Online]. Available: http://ieeexplore.ieee.org/document/5270315/

[7] J. Wei, J. Chen, Y. Feng, K. Ferles, and I. Dillig, "Singularity: Pattern fuzzing for worst case complexity," in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. Lake Buena Vista FL USA: ACM, Oct. 2018, pp. 213–223. [Online]. Available: https://dl.acm.org/doi/10.1145/3236024.3236039

[8] S. Veggalam, S. Rawat, I. Haller, and H. Bos, "IFuzzer: An Evolutionary Interpreter Fuzzer Using Genetic Programming," in *Computer Security – ESORICS 2016*, I. Askoxylakis, S. Ioannidis, S. Katsikas, and C. Meadows, Eds. Cham: Springer International Publishing, 2016, vol. 9878, pp. 581–601. [Online]. Available: http://link.springer.com/10.1007/978-3-319-45744-4_29

[9] M. Y. Wong and D. Lie, "IntelliDroid: A Targeted Input Generator for the Dynamic Analysis of Android Malware," in *Proceedings 2016 Network and Distributed System Security Symposium*. San Diego, CA: Internet Society, 2016. [Online]. Available: https://www.ndss-symposium.org/wp-content/uploads/2017/09/intellidroid-targeted-input-generator-dynamic-analysis-android-malware.pdf

[10] X. Du, B. Chen, Y. Li, J. Guo, Y. Zhou, Y. Liu, and Y. Jiang, "LEOPARD: Identifying Vulnerable Code for Vulnerability Assessment through Program Metrics," 2019. [Online]. Available: http://arxiv.org/abs/1901.11479

[11] M. Z. Nasrabadi, S. Parsa, and A. Kalaee, "Neural Fuzzing: A Neural Approach to Generate Test Data for File Format Fuzzing," 2018. [Online]. Available: http://arxiv.org/abs/1812.09961

[12] ENISA, "Risk Management: Implementation principles and Inventories for Risk Management/Risk Assessment methods and tools "Survey of existing Risk Management and Risk Assessment Methods") Conducted by the Technical Department of ENISA Section Risk Management," 2006.

[13] Y. Li, S. Ji, C. Lv, Y. Chen, J. Chen, Q. Gu, and C. Wu, "V-Fuzz: Vulnerability-Oriented Evolutionary Fuzzing," pp. 1–16, 2019. [Online]. Available: http://arxiv.org/abs/1901.01142

[14] E. Güler, C. Aschermann, A. Abbasi, and T. Holz, "ANTIFUZZ: Impeding Fuzzing Audits of Binary Executables," in *28th USENIX Security Symposium*. Usenix Association, Aug. 2019, p. 18.

[15] C. R. Cioce, D. G. Loffredo, and N. J. Salim, "Program Fuzzing on High Performance Computing Resources." Tech. Rep. SAND2019-0674, 1492735, Jan. 2019. [Online]. Available: http://www.osti.gov/servlets/purl/1492735/

[16] Amazon, "AWS Plugin for Slurm - repository." [Online]. Available: https://github.com/aws-samples/aws-plugin-for-slurm

[17] Microsoft, "Azure CycleCloud," Feb. 2021. [Online]. Available: https://docs.microsoft.com/en-us/azure/cyclecloud/slurm?view=cyclecloud-8

[18] SchedMD, "Cloud Scheduling Guide." [Online]. Available: https://slurm.schedmd.com/elastic_computing.html

[19] A. Fioraldi, D. Maier, H. Eißfeldt, and M. Heuse, "AFL++: Combining Incremental Steps of Fuzzing Research," *USENIX*, p. 12, 2020.

[20] C. Cioce, N. Salim, J. Rigdon, and D. Loffredo, "Multi-Node Program Fuzzing on High Performance Computing Resources." Tech. Rep. SAND2020-8215, 1650237, 690013, Aug. 2020. [Online]. Available: https://www.osti.gov/servlets/purl/1650237/

[21] Abhishek Arya, Oliver Chang, Max Moroz, Martin Barbella, and Jonathan Metzman, "Google Online Security Blog: Open sourcing ClusterFuzz," Feb. 2019. [Online]. Available: https://security.googleblog.com/2019/02/open-sourcing-clusterfuzz.html

[22] Microsoft Research, "Microsoft Security Risk Detection ("Project Springfield") - Microsoft Research," Jan. 2015. [Online]. Available: https://www.microsoft.com/en-us/research/project/project-springfield/

[23] Microsoft, "GitHub - microsoft/onefuzz: A self-hosted Fuzzing-As-A-Service platform." [Online]. Available: https://github.com/microsoft/onefuzz

[24] Mike Aizatsky, Kostya Serebryany, Oliver Chang, Abhishek Arya, and Meredith Whittaker, "Google Testing Blog: Announcing OSS-Fuzz: Continuous Fuzzing for Open Source Software," Dec. 2016. [Online]. Available: https://testing.googleblog.com/2016/12/announcing-oss-fuzz-continuous-fuzzing.html

[25] Microsoft, "Microsoft Security Risk Detection." [Online]. Available: https://www.microsoft.com/en-us/security-risk-detection/

[26] V. J. M. Manes, H. Han, C. Han, S. K. Cha, M. Egele, E. J. Schwartz, and M. Woo, "The Art, Science, and Engineering of Fuzzing: A Survey," *arXiv:1812.00140 [cs]*, Apr. 2019. [Online]. Available: http://arxiv.org/abs/1812.00140

[27] Francisco Borja Garnelo Del Rio, "HOUSE framework repository," 2021. [Online]. Available: https://github.com/b0rh/HOUSE

[28] A. Takanen, J. DeMott, C. Miller, and A. Kettunen, Eds., *Fuzzing for Software Security Testing and Quality Assurance*, 2nd ed., ser. Artech House Information Security and Privacy Series. Norwood, MA: Artech House, 2018.

[29] "CERT BFF," Oct. 2016. [Online]. Available: https://resources.sei.cmu.edu/library/asset-view.cfm?assetID=507974

[30] Aki Helin, "Radamsa source code repository." [Online]. Available: https://gitlab.com/akihe/radamsa

[31] Michał Zalewski, "AFL - american fuzzy lop." [Online]. Available: https://lcamtuf.coredump.cx/afl/

[32] NSA, "Security-Enhanced Linux," 2008. [Online]. Available: https://web.archive.org/web/20200915000700/https://www.nsa.gov/What-We-Do/Research/SELinux/

[33] Chris Brown, "Protect your applications with AppArmor - Linux.com," Aug. 2006. [Online]. Available: https://www.linux.com/news/protect-your-applications-apparmor/

[34] N. N. A. Sjarif, S. Chuprat, M. N. Mahrin, N. A. Ahmad, A. Ariffin, F. M. Senan, N. A. Zamani, and A. Saupi, "Endpoint Detection and Response: Why Use Machine Learning?" in *2019 International Conference on Information and Communication Technology Convergence (ICTC)*. Jeju Island, Korea (South): IEEE, Oct. 2019, pp. 283–288. [Online]. Available: https://ieeexplore.ieee.org/document/8939836/

[35] Gilad Maayan, "Comparing endpoint security: EPP vs. EDR vs. XDR - Infosec Resources," Dec. 2020. [Online]. Available: https://resources.infosecinstitute.com/topic/comparing-endpoint-security-epp-vs-edr-vs-xdr/

[36] GitHub Security Lab, "CodeQL." [Online]. Available: https://securitylab.github.com/tools/codeql/

[37] NIST, "Source Code Security Analyzers." [Online]. Available: https://samate.nist.gov/index.php/Source_Code_Security_Analyzers.html

[38] "ISO/IEC 2382:2015(en) Information technology — Vocabulary," 2015. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v1:en

[39] P. Godefroid, M. Y. Levin, and D. Molnar, "Automated Whitebox Fuzz Testing," *ResearchGate*, p. 17, Jan. 2008.

[40] G. J. Myers, C. Sandler, and T. Badgett, *The Art of Software Testing*, 3rd ed. Hoboken, N.J: John Wiley & Sons, 2012.

[41] V. Ganesh, T. Leek, and M. Rinard, "Taint-based directed whitebox fuzzing," in *2009 IEEE 31st International Conference on Software Engineering*. Vancouver, BC, Canada: IEEE, 2009, pp. 474–484. [Online]. Available: http://ieeexplore.ieee.org/document/5070546/

[42] Muthu Ramachandran, "Testing software components using boundary value analysis," in *Proceedings of the 20th IEEE Instrumentation Technology Conference (Cat No 03CH37412) EURMIC-03*. Belek-Antalya, Turkey: IEEE, 2003, pp. 94–98. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1231572

[43] M. Roper, "Using Machine Learning to Classify Test Outcomes," in *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*. Newark, CA, USA: IEEE, Apr. 2019, pp. 99–100. [Online]. Available: https://ieeexplore.ieee.org/document/8718223/

[44] M. Z. Nasrabadi, S. Parsa, and A. Kalaee, "Format-aware Learn&Fuzz: Deep Test Data Generation for Efficient Fuzzing," *Neural Computing and Applications*, vol. 33, no. 5, pp. 1497–1513, Mar. 2021. [Online]. Available: http://arxiv.org/abs/1812.09961

[45] L. Cheng, Y. Zhang, Y. Zhang, C. Wu, Z. Li, Y. Fu, and H. Li, "Optimizing seed inputs in fuzzing with machine learning," pp. 1–2, 2019. [Online]. Available: http://arxiv.org/abs/1902.02538

[46] J. Wang, B. Chen, L. Wei, and Y. Liu, "Superion: Grammar-Aware Greybox Fuzzing," 2018. [Online]. Available: http://arxiv.org/abs/1812.01197

[47] Nooper Davis, "Secure Software Development Life Cycle Processes," Jul. 2013.

[48] Gerbrand van Dieijen, "Metaphors in software development," Jul. 2010. [Online]. Available: https://xebia.com/blog/metaphors-in-software-development/

[49] OMG, "Business Process Model and Notation (BPMN) Version 2.0.2," Dec. 2013. [Online]. Available: https://www.omg.org/spec/BPMN/2.0.2/

[50] "SCAYLE, Supercomputación Castilla y León." [Online]. Available: https://www.scayle.es/

[51] "Slurm Workload Manager." [Online]. Available: https://slurm.schedmd.com/

[52] "BOE-A-2010-1330 Real Decreto 3/2010, de 8 de enero, por el que se regula el Esquema Nacional de Seguridad en el ámbito de la Administración Electrónica." [Online]. Available: https://www.boe.es/buscar/act.php?id=BOE-A-2010-1330&b=23&tn=1&p=20151104

[53] "AFLplusplus is the daughter of the American Fuzzy Lop fuzzer by Michal "lcamtuf" Zalewski." [Online]. Available: https://aflplus.plus/

[54] GNU, "Coreutils." [Online]. Available: https://www.gnu.org/software/coreutils/coreutils.html

[55] K. Serebryany, D. Bruening, A. Potapenko, and D. Vyukov, "Address-Sanitizer: A Fast Address Sanity Checker," *USENIX*, p. 10, 2012.

[56] "Chapter 25. Automatic Bug Reporting Tool (ABRT) Red Hat Enterprise Linux 7 — Red Hat Customer Portal." [Online]. Available: https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7/html/system_administrators_guide/ch-abrt

http://doi.org/10.18239/jornadas_2021.34.40

Reducción de dimensionalidad sin pérdida en representaciones semánticas de texto

Iñaki Vélez de Mendizábal, Enaitz Ezpeleta, Urko Zurutuza
Mondragon Unibertsitatea
Loramendi 4, 20500, Mondragón
{ivelez, eezpeleta, uzurutuza} @ mondragon.edu
https://orcid.org/{0000-0001-5460-9549, 0000-0003-4121-8869, 0000-0003-3720-6048}

Resumen- **El spam supone actualmente más del 50% del tráfico de correo electrónico. Es la vía de entrada para muchos de los ataques de secuestro de información (ransomware) que sufren las empresas. En este trabajo se propone la utilización de información semántica en los filtros antispam, sustituyendo y agrupando palabras como 'Viagra', 'Cialis' o 'Tadalafil' por su hiperónimo 'anti_impotence_drug' y utilizando synsets (conjuntos de sinónimos) para su representación. Se ha diseñado y probado un sistema de generalización de conceptos/palabras sin pérdida de información, que combina la información semántica y los algoritmos genéticos multi-objetivo. Los resultados obtenidos demuestran que es posible mejorar la detección de los mensajes legítimos, así como aumentar la velocidad de clasificación.**

***Index Terms*- filtrado spam, ciberseguridad, representación basada en synsets, reducción de dimensionalidad, información semántica, algoritmos genéticos multi-objetivo.**

Tipo de contribución: *Investigación ya publicada. Artículo original "SDRS: A new lossless dimensionality reduction for text corpora" publicado en Information Processing & Management.*

I. Introducción

Uno de los mayores problemas relacionados con el spam que sufren las empresas es la propagación del ransomware. Aunque no son muchos los ataques reconocidos por las empresas, algunos se han hecho públicos debido a su repercusión, como los sufridos por las entidades Prosegur, Everis, Cadena Ser o el ayuntamiento de Jerez. Ataques todos ellos, provocados por Ryuk [1], los cuales comienzan por la recepción de un correo electrónico de tipo phishing. El ataque de ransomware Ryuk, comienza por infectar los equipos con el bot Emotet. Esto se consigue mediante el envío de mensajes de correo spam que contienen un documento con una macro maliciosa, que el bot descarga si la víctima permite la ejecución de la macro.

En este trabajo se plantea una nueva forma de abordar la problemática de los mensajes de tipo spam, que consiste en la utilización de la información semántica contenida en ellos. Existen estudios previos [2] que utilizan synsets (conjuntos de sinónimos o *synonym set*) de Wordnet[1] para representar la información de los textos y que han conseguido una mayor eficacia en la clasificación del spam.

Un problema que se produce en el tratamiento de textos mediante técnicas de Natural Language Processing (NLP) es la dimensión del vector de características, utilizado para representar el texto. Cuanto más grande es este vector, más aumentan las necesidades de memoria y cálculo de los equipos de procesamiento. El objetivo principal de este trabajo es comprobar el rendimiento de un nuevo sistema de reducción de dimensionalidad basado en el uso de información semántica para representaciones de textos en forma de synsets.

II. Estado del arte

En los últimos años se ha desarrollado un nuevo método de representación para los textos: el concepto. El concepto puede ser definido como la representación del significado específico de la palabra o conjunto de palabras y sus sinónimos dentro de un determinado contexto. Por ejemplo, los textos "automóvil barato a la venta", "coche barato en venta", "oferta de auto económico" tienen el mismo significado, pero están compuestos por diferentes palabras.

La representación de conceptos por si sola, no consigue resolver el problema del elevado número de características que es necesario evaluar para entrenar los clasificadores basados en técnicas de Machine Learning (ML). En un trabajo previo en el que se utiliza información semántica de Bahgat et al. [3] se aprovechan las relaciones de sinonimia entre los términos para reducir levemente la dimensionalidad. Sin embargo, recientemente un estudio [2] propuso el uso de las relaciones de hiperonimia entre synsets para reducir la dimensionalidad.

III. Metodologia y resultados

Una de las principales novedades en la presente propuesta es la utilización de synsets de BabelNet[2]. Adicionalmente las palabras también se desambiguan para relacionarlas con el synset que mejor se ajusta a su significado. Usando este esquema, los mensajes originales se han transformado en un dataset (D) donde los mensajes se han representado en un vector de synsets (Fig. 1).

Para seleccionar los synsets a generalizar y el nivel de generalización de cada uno de ellos, se ha planteado un problema de optimización multi-objetivo para cuya resolución se ha utilizado un algoritmo MOEA (Multi-Objective

[1] https://wordnet.princeton.edu/

[2] Disponible en www.babelnet.org

Evolutionary Algorithm). En concreto se ha marcado como objetivo la minimización de Falsos Positivos (FP, mensajes legítimos clasificados como spam) y Falsos Negativos (FN, mensajes spam clasificados como legítimos). Como el objetivo principal de este trabajo es la reducción de dimensionalidad, éste se ha añadido también como objetivo.

msg	bn:00069736n (scooter)	bn:00004618n (antivirus)	bn:00041948n (meeting)	bn:00067997n (roadster)	bn:00007309n (car)	class
1	1	0	0	0	1	spam
2	0	0	0	1	1	spam
3	0	0	2	0	0	ham
4	0	0	1	0	1	ham
5	0	1	0	0	0	ham

Fig. 1 Ejemplo de dataset (D) para selección de características

El cromosoma empleado en este estudio C = {c_1, c_2, c_3,…,c_n} representa el número de generalizaciones a aplicar al synset en cada posición. Por ejemplo, el resultado de la aplicación de un cromosoma como C = {1,1,0,0,2} puede verse en la Fig. 2. El proceso ha consistido en la generalización de dos niveles del synset 'bn:00007309n' (car), donde la primera generalización nos lleva hasta 'bn:00007385n' (motor vehicle), llegando en la segunda generalización hasta el synset 'bn:00079675n' (vehicle). El synset 'bn:00069736n' (scooter) es generalizado un nivel hasta 'bn:00080980n' (wheeled vehicle), pero como este synset es un hipónimo del synset previamente generalizado 'bn:00079675n' (vehicle), ambos son fusionados en el synset más general, tal y como se hace también con el synset bn:00067997n (roadster).

msg	bn:00079675n (vehicle)	bn:00021492n (program)	bn:00041948n (meeting)	class
1	2	0	0	spam
2	2	0	0	spam
3	0	0	2	ham
4	1	0	1	ham
5	0	1	0	ham

Fig. 2 Transformación del dataset (D) por el cromosoma C = {1,1,0,0,2}

Para poder evaluar los resultados de la propuesta, se ha utilizado como dataset el "*YouTube Spam Collection*"[1] y se ha diseñado el protocolo experimental que se muestra en Fig. 3.

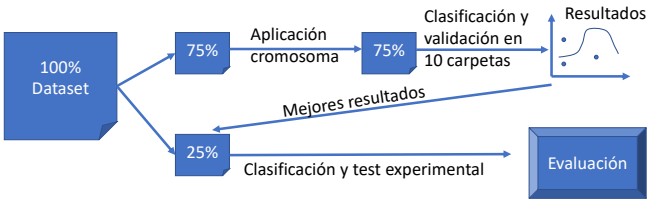


Fig. 3 Flujo protocolo experimental

En la Fig. 4 puede verse la comparativa entre realizar una reducción del vector de características utilizando tokens e Information Gain (IG), o synsets tal como se propone en este trabajo de investigación. Puede verse como utilizando una representación basada en synsets, el número de mensajes clasificados erróneamente como negativos (Falsos Negativos o FN) es mayor, pero el número de mensajes clasificados erróneamente como positivos (Falsos Positivos FP) se reduce considerablemente.

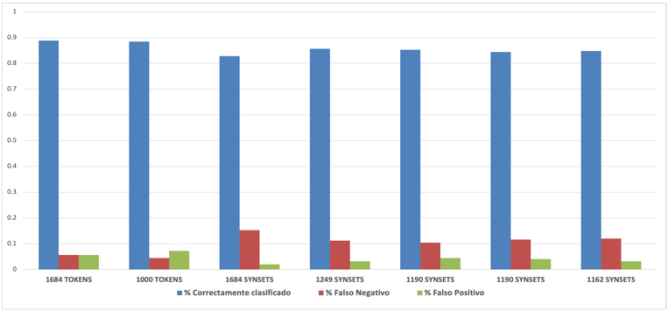


Fig. 4 Comparativa reducción tokens+IG o synsets+propuesta

IV. Conclusiones

Este trabajo realiza una comparativa experimental del rendimiento que pueden obtener los filtros antispam al utilizar representaciones de los textos basadas en synsets y en tokens. Los resultados muestran que los primeros funcionan mucho mejor en los contextos en los que el coste de los FP y FN es asimétrico, que es la situación más común para los usuarios.

Desde un punto de vista práctico, el decremento del número de palabras redundantes, las cuales se agrupan bajo un único synset, reduce los requerimientos de calculo para la fase de entrenamiento del clasificador. Desde un punto de vista teórico, al combinar estas palabras altamente dependientes en un synset se aumenta la independencia de las características, lo que aumenta el rendimiento de algunos clasificadores, como por ejemplo el Naïve Bayes.

Los resultados experimentales permiten concluir que la representación de synsets es mejor para la detección de mensajes legítimos que para detectar los mensajes spam.

Agradecimientos

Este trabajo ha sido parcialmente financiado por MINECO/AEI/FEDER a través del proyecto TIN2017-84658-C2-2-R, por el Departamento de Desarrollo Económico e Infraestructuras a través del contrato KK-2020/00054 y por el proyecto TRUSTIND del departamento de Desarrollo Económico e Infraestructuras del Gobierno Vasco. Ha sido desarrollado por el grupo de Sistemas Inteligentes para Sistemas Industriales, financiado por el Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco.

Referencias

[1] N. A. Hassan, 'Ransomware Families', in *Ransomware Revealed*, Berkeley, CA: Apress, 2019, pp. 47–68.

[2] J. R. Méndez, T. R. Cotos-Yañez, and D. Ruano-Ordás, 'A new semantic-based feature selection method for spam filtering', *Appl. Soft Comput. J.*, vol. 76, pp. 89–104, Mar. 2019.

[3] E. M. Bahgat, S. Rady, W. Gad, and I. F. Moawad, 'Efficient email classification approach based on semantic methods', *Ain Shams Eng. J.*, vol. 9, no. 4, pp. 3259–3269, Dec. 2018.

[1] https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection

Sesión de Investigación B1: Inteligencia Artificial en ciberseguridad
http://doi.org/10.18239/jornadas_2021.34.41

Entrenamiento optimizado de redes neuronales para reconocimiento biométrico

Gonzalo García Miranda
Universidad de León
Campus de Vegazana s/n, 24071, León
ggarcm04@estudiantes.unileon.es

Alberto Calvo García
Universidad de León
Campus de Vegazana s/n, 24071, León
acalvg05@estudiantes.unileon.es

Claudia Álvarez Aparicio
ORCID 0000-0002-7465-8054
Grupo de Robótica
Universidad de León
Campus de Vegazana s/n, 24071, León
calvaa@unileon.es

Ángel Manuel Guerrero Higueras
ORCID 0000-0001-8277-0700
Grupo de Robótica
Universidad de León
Campus de Vegazana s/n, 24071, León
am.guerrero@unileon.es

Francisco Javier Rodríguez Lera
ORCID 0000-0002-8400-7079
Grupo de Robótica
Universidad de León
Campus de Vegazana s/n, 24071, León
calvaa@unileon.es

Camino Fernández Llamas
ORCID 0000-0002-8705-4786
Grupo de Robótica
Universidad de León
Campus de Vegazana s/n, 24071, León
camino.fernandez@unileon.es

***Resumen*—Los sistemas actuales de autenticación biométrica para un grupo de individuos suelen requerir de un reentrenamiento completo cuando se registran nuevos usuarios, causando la indisponibilidad del sistema durante todo ese tiempo. En el intento de solventar dicho problema se propone un sistema de autenticación que permite autenticar a un usuario registrado con su huella dactilar sin afectar a la disponibilidad. Este sistema es gestionado por un algoritmo que estima y gestiona un cierto número necesario de redes neuronales que funcionan de forma paralela de modo que el proceso de alta de nuevos usuarios, además de ser rápido, no afecta en ningún caso al funcionamiento del sistema ni al resto de usuarios.**

***Palabras Clave*—Autenticación, biometría, huella dactilar, redes neuronales.**

Tipo de contribución: *Investigación original.*

I. Introducción

En la actualidad son frecuentes los dispositivos que utilizan un sistema de autenticación biométrica para identificar a sus usuarios y proteger así la información de éstos. La popularización y normalización del reconocimiento biométrico ha contribuido a la aceptación general de este tipo de autenticación por parte de los usuarios, siendo cada vez más común en organizaciones y empresas.

Los sistemas biométricos aplicados en la autenticación de un grupo de usuarios habitualmente usan técnicas de aprendizaje automático y/o redes neuronales para identificar a los sujetos. En estos casos, los procesos de alta de nuevos usuarios suelen ser bastante costosos, tanto en tiempo como en operatividad, puesto que requieren de un nuevo entrenamiento para incluir a los últimos individuos registrados, haciendo que el sistema quede indisponible temporalmente para los demás o, en su defecto, para los nuevos (ya que todavía no podrían hacer uso de él). Así pues, cuando la cantidad de usuarios aumenta el problema se va agravando con tiempos de indisponibilidad cada vez más elevados, pudiendo incluso consumir demasiados recursos. La solución más común consiste en posponer el alta de los nuevos usuarios a determinados momentos en los que esta indisponibilidad del sistema no afecte al funcionamiento general de la red. De esta manera, aunque es cierto que el funcionamiento del sistema para el resto de usuarios no se ve comprometido en ningún caso, se alarga el proceso de alta más de lo estrictamente necesario en tiempos de computación.

La idea aquí presentada consiste en compartimentar el sistema de autenticación biométrica de los usuarios en una serie de subsistemas, dedicados cada uno a un subconjunto tomado de entre el total de los usuarios registrados en el sistema. De este modo se pretende garantizar la disponibilidad del sistema para todos los usuarios, reduciendo directamente los tiempos de entrenamiento mediante un reentrenamiento parcial que incluya al nuevo individuo casi inmediatamente. En consecuencia, se minimiza el tiempo de espera en el que el nuevo usuario registrado puede hacer un uso normal y continuado del sistema, sin afectar a la disponibilidad para el resto de usuarios. Además, análogamente, este procedimiento permite dar de baja a un usuario prácticamente de forma instantánea.

Este trabajo tiene como objetivo principal el diseño de un algoritmo, denominado Algoritmo Gestor de Redes Neuronales (AG-RN), que permite gestionar las redes neuronales que conforman el sistema de autenticación biométrico para un grupo de individuos, de manera que se puedan dar de alta nuevos (o de baja antiguos) usuarios de forma casi inmediata garantizando la disponibilidad constante del sistema. En concreto, se ha desarrollado una prueba de concepto (PoC, por sus siglas en inglés) para reconocimiento de huella dactilar con el fin de demostrar el correcto funcionamiento, la efectividad general y la reducción de los tiempos de entrenamiento ante el aumento del número de individuos registrados.

La estructura de este artículo es la siguiente. En la sección II se expone una breve revisión metódica de otras posibles soluciones a situaciones similares para poner en contexto el problema en cuestión y la solución planteada. En la sección III se explica el procedimiento desarrollado en relación al conjunto de datos utilizado y al diseño de la arquitectura del

sistema a implementar. En la sección IV se presenta el análisis de la solución aplicada, detallando los criterios en cuanto a tiempos de entrenamiento y la precisión del sistema requeridos para luego diseñar el algoritmo que finalmente se implementa en la PoC. En la sección V se describen las pruebas de verificación para demostrar que la PoC es correcta y, por tanto, satisfactoria. En la sección VI se infiere la importancia de la solución planteada junto con sus posibles inconvenientes y líneas de investigación futuras.

II. Estado del arte

Fundamentalmente los procedimientos de autenticación se realizan mediante información secreta (como contraseñas), objetos especiales (como tarjetas personales), o bien atributos biométricos (como huellas dactilares). Cada una de estas formas de autenticación conlleva diferentes grados de seguridad, puesto que la información secreta puede ser olvidada, robada o compartida; y los objetos especiales pueden ser sustraídos, deteriorados o duplicados. Sin embargo, los atributos biométricos de una persona son únicos, difícilmente usurpables o replicables y, salvo accidente grave, inmutables. No obstante, a medida que aumenta la seguridad que proporcionan estos métodos también se incrementa la complicación en su implementación. En particular, la captación de los atributos biométricos requiere de un dispositivo complejo capaz de analizar el cuerpo, realizando así una lectura de datos que posteriormente son tratados y codificados en el formato deseado, la llamada plantilla biométrica [1].

La principal desventaja de los sistemas de autenticación biométrica son las interferencias en la lectura de datos, pudiendo arrojar una mayor tasa de falsos negativos que el resto de tipos de autenticación. Sin embargo, se pueden adaptar las plantillas de identificación con los sucesivos usos lícitos del sistema como solución [2]. Por otro lado, aunque en general los datos biométricos son inalterables, hay trabajos que intentan conseguir la identificación mediante diversas modificaciones [3], [4]. Además, esta inmutabilidad origina una contingencia inherente conocida como riesgo de irrevocabilidad, que se refiere a la incapacidad de actualizar, reeditar o destruir el atributo biométrico una vez éste haya sido comprometido. Para mitigar este riesgo una opción común es la autenticación multifactor [5], generando plantillas biométricas dependientes de otros factores como, por ejemplo, una contraseña y que, por tanto, sí permiten la actualización, reedición o destrucción de éstas [6]. En particular, también es importante destacar que los rasgos biométricos se consideran datos de carácter personal a todos los efectos legales. En consecuencia su uso, almacenamiento o tratamiento está sometido al cumplimiento de las distintas exigencias de carácter jurídico, técnico, físico y organizativo como, por ejemplo, prevee el Real Decreto de la Ley Orgánica de Protección de Datos de Carácter Personal (RDLOPD) [7].

La clasificación de los sistemas de identificación biométrica está basada, sustancialmente, en: atributos físicos (como huella dactilar o palmar, reconocimiento facial, escáner de iris o retina, etc.) o comportamiento (como el reconocimiento de voz, firma o tecleo). Principalmente, el más utilizado es el atributo físico de huella dactilar, quizás por su carácter orgánico único, ya que supone un desarrollo diferente para cada individuo incluso entre los propios dedos.

Tal y como se describe en los manuales originarios del tratamiento de huellas dactilares, los primeros trabajos se centraron en la detección y comparación de las características propias de cada huella y, más particularmente, en los diferentes patrones que forman las pequeñas rugosidades de las palmas de las manos y los dedos [8]. De este modo, mediante el reconocimiento de patrones en la huella, la identificación es inequívoca asignando a cada entrada una puntuación basada en las características genuinas de la huella que pertenecen a un solo individuo [9]. Un método común de reconocimiento de patrones es la identificación de las propias *minutiaes*, que son las líneas que según su grosor, separación y formas configuran estos patrones. Su distancia y sus intersecciones o falta de ellas permite definir de un modo más fino estos puntos de comparación entre huellas, lo cual suele inducir resultados bastante satisfactorios [10]. Ahora bien, este método difiere un poco del reconocimiendo de patrones en sí mismo siendo más próximo a lo que hace una red neuronal convolucional (CNN, del inglés *Convolutional Neural Network*). El motivo principal se basa en la necesidad de preprocesar la imagen de la huella antes de su tratamiento, para encontrar así esos puntos asignándoles un cierto valor que luego es comparado con los obtenidos de la huella a autenticar. Existen muchos métodos diferentes, continuándose avanzando actualmente en los métodos de detección, mejora de imágenes y codificación de las mismas [11], [12], [13]. Además, la manera de obtener huellas dactilares puede ser muy variada influyendo en la calidad de la imagen recopilada, pudiendo afectar gravemente a los procedimientos de identificación y codificación de *minutiaes* (ya que los errores de lectura pueden ser interpretados como rasgos propios). Aunque ambos métodos son ampliamente utilizados y obtienen buenos resultados, algunos trabajos afirman que el reconocimiento de patrones puede ser más correcto al abstraerse de ese proceso de extracción de *minutiaes*, puesto que requiere de una codificación adicional [14].

Las CNNs se pueden definir como la replicación artificial de la estructura cerebral que posibilita el aprendizaje mediante un conjunto estratificado de capas de nodos que filtran sucesivamente información para el análisis de datos visuales, destacando su aplicación en visión por ordenador como en la clasificación de imágenes y el reconocimiento de objetos. Así pues, este tipo de red neuronal se puede emplear en la identificación de huellas dactilares mediante las *minutiaes*. Este método utiliza toda la imagen de la huella aplicando sucesivas capas de convolución que, a modo de filtros, permiten distinguir sucesivamente patrones más específicos dentro de la imagen. Su uso requiere de un procesamiento previo de la imagen para normalizarla, limpiar el posible ruido y estilizar las líneas que forman estas *minutiaes* o definir una serie de patrones a identificar dentro de la huella, sin tener así que declararlo después de forma explícita [15]. De esta manera se pueden identificar y extraer los puntos de la imagen para luego calcular sus distancias relativas y orientaciones, siendo finalmente los parámetros de entrada de la red neuronal [12], [16], [17]. Así pues, se pueden crear *autoencoders* mediante la sucesiva aplicación de capas de convolución en conjunción con capas de "Max Pooling". Estos *autoencoders* crean unos descriptores de las imágenes mediante un aprendizaje no supervisado que pueden utilizarse para, entre otras cosas,

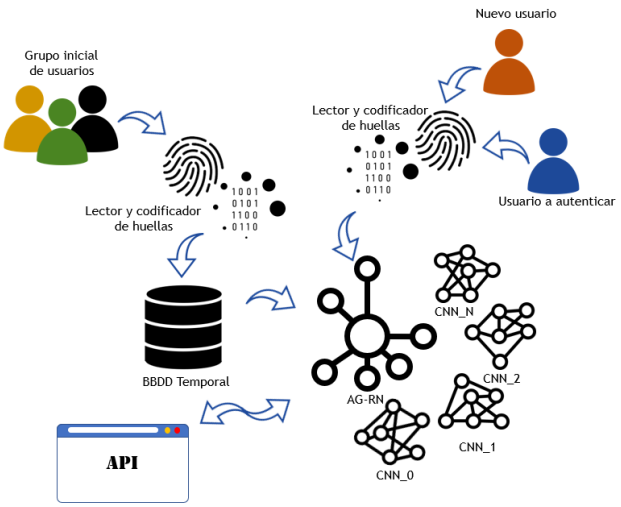


Figura 1. Esquema ideal del sistema.

tomar automáticamente las características más importantes de una imagen y, a partir de ellas, relacionar dos imágenes tomadas desde ángulos distintos o con alguna modificación distinta entre ambas [18]. Todas estas características que aportan las CNN y su uso para crear *autoencoders* permiten un reconocimiento de patrones sin la necesidad de explicitarlos, lo cual puede hacerse extensivo al reconocimiento de patrones dentro de las propias huellas dactilares. Es decir, tal y como demuestran Chowdhury *et al.* [15], es posible entrenar una CNN para que realice reconocimiento de huellas dactilares mediante los patrones que las *minutiaes* forman en ellas, sin que la propia red sea "consciente" de que eso es lo que está haciendo.

III. Metodología

Como el objetivo es implementar un sistema de autenticación mediante huella digital que permite identificar individuos realizando el registro o borrado de usuarios en el momento (es decir, sin requerir de un reentrenamiento total), es preciso, por un lado, recolectar un conjunto de datos de huellas dactilares lo suficientemente grande y, por otro lado, diseñar una arquitectura conveniente de las CNNs para que el algoritmo AG-RN pueda gestionar el número variable de redes con una interfaz que permita la utilización de este sistema por parte de terceros. En la Fig. 1 se muestra el esquema ideal del sistema.

El conjunto de datos utilizado proviene del recolectado en una base de datos preexistente, debido a que no se ha dispuesto de un procedimiento para ello. Además, las CNNs necesitan tener un tamaño manejable (en términos computacionales) y cada una debe ser entrenada con un subconjunto diferente y reducido de los datos totales de los individuos, pero manteniendo un nivel suficiente de redundancia de los datos de cada individuo a lo largo de varias de estas redes. En particular, el desarrollo general del sistema se ha realizado casi exclusivamente en el lenguaje Python por su uso común y flexible. Adicionalmente la interfaz del sistema es muy básica, a modo ilustrativo, consistente en el uso del intérprete Python de Jupyter para probar y evaluar de forma sencilla el algoritmo.

III-A. Conjunto de datos

La base de datos de huellas dactilares seleccionada para el sistema es Sokoto Coventry Fingerprint (SOCOFing) [19], que provee imágenes de huellas dactilares de cada uno de los 10 dedos de 600 personas. En especial, para mantener un conjunto grande de muestras, se ha decidido interpretar cada una de estas huellas como un individuo diferente, ya que a efectos de nuestro estudio en términos de autenticación se trata de diferentes plantillas biométricas. Además, SOCOFing provee de unas imágenes alteradas artificialmente de cada una de las huellas originales, simulando en tres grados distintos (fácil, medio y difícil) los tres tipos diferentes de alteraciones: obliteración, rotación central y corte en Z [3], [4]. No obstante, no es el objetivo de este trabajo centrarse en la identificación de sujetos cuyas huellas hayan sido alteradas, por lo que su uso principal aquí es para simular malas condiciones en el proceso de lectura que generan imágenes modificadas.

Las imágenes se proporcionan en formato BMP y sus dimensiones son 96x103 píxeles. Sin embargo, el marco de las imágenes ha sido recortado para dejar únicamente la imagen de la huella. Por tanto, la dimensión final es 90x90 píxeles leídas en escala de grises y almacenadas en ficheros NPZ, diferenciando cada uno de los conjuntos de datos proporcionados por SOCOFing (original, fácil, medio y difícil). Además, por cuestiones de eficiencia, se ha traducido el nombre de las imágenes a formato numérico donde el patrón general "ID_SEXO_MANO_DEDO_finger" para el siguiente ejemplo "11_M_Right_thumb_finger" se convierte en la etiqueta "[11, 0, 1, 0]". Estas etiquetas han sido almacenadas en archivos NPY, especiales para almacenar series de números NumPy.

Para comprobar diferentes funcionamientos de la red a lo largo de todo el proceso de desarrollo, se han creado tres subconjuntos de datos de diferentes tamaños del conjunto total de huellas que provee SOCOFing:

- El conjunto total de los datos (600 personas): Utilizado para el desarrollo de las CNNs que usa como base el AG-RN y también para el análisis de los tiempos de entrenamiento.
- Un conjunto con casi todos los datos (590 personas): Utilizado como conjunto de datos total sobre el que se crea el gestor y sobre el que se realizan los entrenamientos iniciales de las redes.
- Un conjunto con el resto de los datos (10 personas): Utilizado para añadir nuevos usuarios al gestor.

Además, se ha decidido utilizar la técnica de incremento de datos para aumentar el número de imágenes de huellas de cada individuo y garantizar así un buen entrenamiento. Esta técnica consiste en obtener, a partir de los datos de los que se dispone y mediante modificaciones artificiales de los mismos, un conjunto de datos mayor que permite un entrenamiento más extenso. Para ello se ha diseñado un proceso muy simple que ayuda a obtener una imagen algo distorsionada a partir de otra, lo cual permite simular múltiples lecturas de una misma huella por parte de un sensor con las pequeñas imperfecciones y/o modificaciones que la propia lectura pueda introducir. Se ha hecho uso de la librería imgaug [20], en concreto de

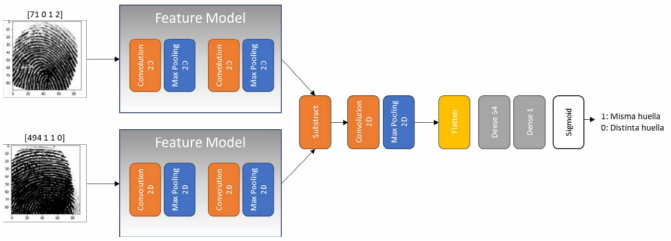


Figura 2. Arquitectura de la CNN.

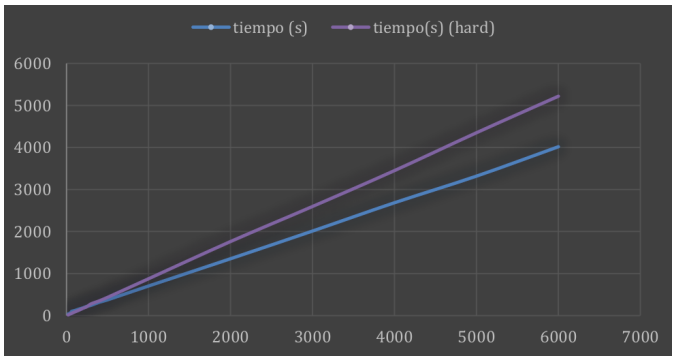


Figura 3. Tiempos de entrenamiento por volumen de usuarios.

augmenters, para asegurar que la imagen obtenida no se va a repetir mediante las siguientes técnicas:

- Escalar: Modifica el tamaño entre 0.9 y 1.1 veces el original.
- Traslación: Modifica la posición para que los mismos puntos no se encuentren en los mismos píxeles.
- Rotación: Se gira entre $-30°$ y $30°$.
- Desenfoque Gaussiano: Combina en cada pixel aquellos que lo rodean.

III-B. Arquitectura de la CNN

Una vez leído y codificado el conjunto de datos, se construye la red neuronal basada en reconocimiento de patrones de las *minutiaes* mediante convolución sin tener así que especificarlo explícitamente. Esta construcción se inspira en la propuesta por Chowdhury *et al.* [15].

La CNN está diseñada para recibir 2 imágenes de 90x90 y devolver la probabilidad de que las imágenes coincidan con la misma huella. Sobre cada una de las entradas se extraen sus características por medio de una red de extracción de características creada mediante la aplicación sucesiva de dos capas combinadas de "Convolution 2D" y "Max Pooling 2D". De este modo se generan los *autoencoders* necesarios para crear los descriptores sobre las distintas imágenes de las huellas dactilares. En particular, se aplica dos veces sobre las imágenes de entrada para distinguir, primero, el fondo de la huella y, posteriormente, hacer una distinción más fina dentro de la huella que permita captar los patrones que forman las *minutiaes*. Así pues, estas redes de extracción de características son las que crean la entrada que se aplica a la red que finalmente da la predicción. Esta segunda red recibe ambas entradas codificadas, dado que necesita comparar los patrones de características de ambas imágenes mediante la capa Substract. A continuación, se aplica de nuevo una capa de "Convolution 2D" y "Max Pooling 2D" para realizar una extracción de características de la comparación. Este nuevo descriptor proporciona una salida en 2D, que se aplana mediante una capa Flatten para pasarla a una nueva capa densa de 64 neuronas. La salida de esta última capa es una única neurona que da el resultado final, la predicción entre 0 y 1 que estima la probabilidad de acierto. En la Fig. 2 se muestra esta arquitectura de la CNN.

Antes de utilizar el modelo es necesario compilarlo, para lo que se necesitan elegir ciertos parámetros que se utilizan durante el entrenamiento de la red. En este caso como función de pérdidas se ha escogido la binary_crossentropy, dado que se usa en los modelos de clasificación binarios como el que se utiliza aquí, y como optimizador se ha seleccionado Adam [21].

IV. Implementación

Una vez creada la red y el modelo se prueba su funcionamiento. Primero, como es común en el entrenamiento de redes neuronales, se divide el conjunto de datos en datos de entrenamiento y datos de prueba. En este caso se ha hecho una división 90 %-10 % respectivamente. Ahora bien, los datos almacenados son huellas con su correspondiente etiqueta. Dado que el objetivo de nuestra red es el de comparar dos huellas y mostrar el grado de certidumbre en un rango de 0 a 1, se debe hacer un cambio para poder dar estos datos de forma correcta. Se trata de, para cada una de las huellas que están en el conjunto de datos tomar dos huellas aumentadas, una creada a partir de la misma huella y otra tomada de forma aleatoria entre el resto de datos y asignar a cada par de huellas un 1 o un 0 dependiendo de si es la misma huella o no, lo cual se hace mediante una clase adicional.

IV-A. Estudio de tiempos de entrenamiento y precisión

Para comprobar en qué medida un aumento en el tamaño del conjunto de datos está influyendo tanto en los tiempos de entrenamiento como en la precisión de las redes que se entrenan, se ha diseñado un experimento haciendo lo siguiente:

1. Se crean conjuntos de datos de distintos tamaños que oscilan desde los 20 a los 600 individuos, tomados de manera aleatoria de entre la totalidad de los individuos que conforman los 600 individuos de SOCOFing.
2. Para cada uno de esos conjuntos de datos se entrena el modelo utilizando las mismas condiciones salvo el propio conjunto y se guardan los tiempos de entrenamiento.
3. Se entrenan los modelos incluyendo tanto las huellas con alteraciones difíciles como sin ellas, dado que esto afecta a los tiempos de entrenamiento al haber más lotes en cada período y, sobre todo, al ratio de falso rechazo (FRR, del inglés *False Rejection Rate*), puesto que algunas alteraciones son bastante extremas.
4. Se prueba el modelo con 100 resultados positivos tomados de entre los valores de testeo del modelo y con otros 100 resultados negativos, lo cual permite obtener el FRR y el ratio de autenticación errónea (FAR, del inglés *False Acceptance Rate*) de cada una de las redes.

La evolución de los tiempos de entrenamiento del experimento se muestra en la Fig. 3.

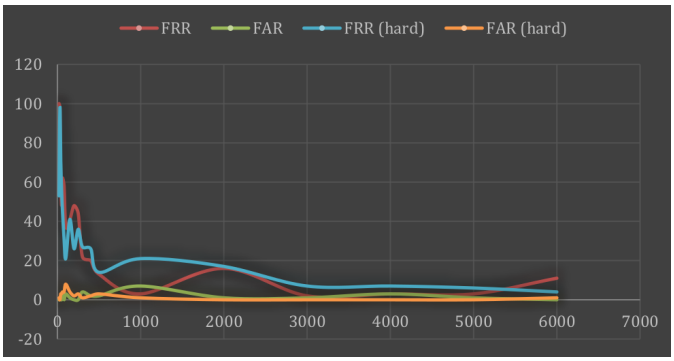


Figura 4. FAR y FRR por volumen de usuarios.

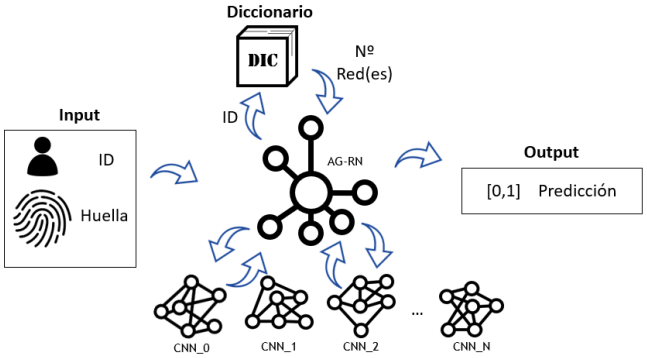


Figura 5. Diagrama del funcionamiento de la predicción del AG-RN.

Aunque este estudio de tiempos se ha realizado con un portátil estándar actual, estos resultados sirven para ilustrar la tendencia lineal general según aumenta el número de individuos. Asimismo, para las mismas redes que fueron entrenadas en el análisis de los tiempos de entrenamiento, se analiza la precisión de las predicciones cuando se produce este aumento en el número de individuos. Los resultados del FRR y del FAR se muestran en la Fig. 4.

Así pues, se observa que, aunque hay una cierta inestabilidad en redes con muy pocos individuos, la tendencia es claramente decreciente según se aumenta el número de usuarios incluidos en el entrenamiento. Las redes con muy pocos individuos presentan tasas de FRR muy elevadas, lo cual afectaría en gran medida a su usabilidad, mientras que para redes más grandes se mantiene en valores muy bajos, lo cual implica un funcionamiento muy certero de la red. En cuanto a la tasa de FAR se aprecia que, en general, se mantiene muy baja en todo momento salvo algún incremento puntual que se puede relacionar con una disparidad estadística. Por otro lado, tal y como se esperaba, el FAR es mayor en los conjuntos de datos que emplean las huellas con alteraciones difíciles, aunque con mayores conjuntos, las tasas tienden a igualarse. En definitiva, a partir de un número mínimo de usuarios la precisión de la red se mantiene en valores satisfactorios mientras que el tiempo de entrenamiento aumenta progresivamente según lo haga el número de individuos.

IV-B. Diseño del AG-RN

El primer paso para desarrollar el algoritmo es encontrar el Punto de Distribución (PD), entendido como el número óptimo de usuarios con el que la red tiene una respuesta correcta manteniendo un tiempo de entrenamiento aceptable. Entonces, el PD es variable para cada implementación del algoritmo en función de qué parámetro se considere más importante (es decir, un FAR bajo implica un PD alto mientras que un entrenamiento rápido implica un PD más bajo). Por tanto, hay que garantizar un balance para que la precisión de la red se mantenga como mínimo en un nivel aceptable antes de que el conjunto de datos sea demasiado grande. Por ejemplo, en el estudio realizado se ha decidio establecer un tiempo de entrenamiento máximo por red de 600 segundos y un FRR máximo del 30 %. Tras realizar un análisis sobre cómo afecta el tamaño de las redes a la precisión y el tiempo de entrenamiento, se ha obtenido un intervalo para el PD de [200, 700] y, seleccionando el valor más óptimo para el propósito de la prueba, el PD establecido es 250. Así pues, se subdivide el conjunto inicial en grupos de N usuarios, con N igual o muy cercano al PD seleccionado. No obstante, se precisa distribuir a cada usuario en al menos un subconjunto con el que se haya entrenado una red, de modo que así se puede autenticar al usuario de forma tan eficaz como con una red entrenada con el total de usuarios. En consecuencia, como requisito adicional, se hace necesario saber a qué subconjunto pertenece cada usuario mediante el uso, en este caso, de un diccionario.

El diccionario devuelve la(s) red(es) en la(s) que se encuentra el usuario. Esto representa la redundancia (Rd), ya que indica el número mínimo de redes en las que se debe incluir cada usuario, siendo también su valor ideal variable dependiendo de la implementación del AG-RN. La redundancia presenta varias funciones dentro del algoritmo, permitiendo la adicción inmediata de nuevos usuarios únicamente añadiendo y entrenando Rd nuevas redes. En este caso, sólo se debe controlar que tras sucesivas adicciones el número de redes no sea excesivo, puesto que esto afectaría al tiempo de respuesta del sistema. De la misma forma, también permite la eliminación de usuarios mediante la supresión de las redes en las que se encuentren y reentrenando nuevas redes para aquellos usuarios cuya redundancia haya caído por debajo de la establecida. Además, al tener no una sino Rd redes haciendo una predicción sobre los mismos datos de entrada, pero entrenadas con diferentes conjuntos de datos, las FFR y FAR son reducidas. Adicionalmente, una redundancia eficiente proporciona flexibilidad en el entrenamiento (ya que en los sucesivos entrenamientos de las redes por la adicción o eliminación de usuarios la mejora es sustancial, aunque es posible que en el inicial resulte mayor que entrenando una única red) y en el uso de memoria (debido a que se pueden cargar en primer lugar sólo aquellas redes en las que se encuentren los individuos que hagan un uso intensivo del sistema).

En la Fig. 5 se muestra un diagrama del funcionamiento del AG-RN.

Entonces, para distribuir de forma óptima los usuarios con una Rd determinada entre el total de redes de tamaño PD, se hace necesario un algoritmo que se ha denominado distribución sacudida. Este procedimiento consiste en, primero, distribuir los usuarios de forma secuencial en el número calculado de subconjuntos necesarios para garantizar la Rd

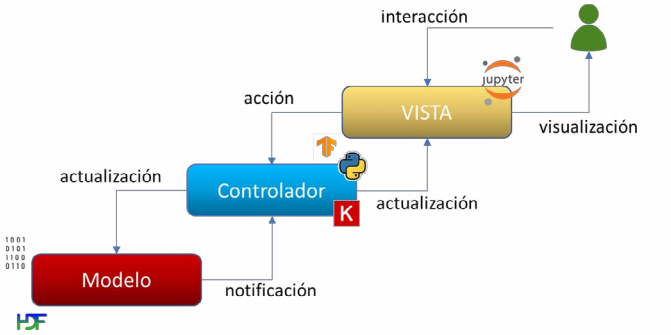


Figura 6. MVC aplicado en la PoC.

establecida (sin que éstos se repitan en el mismo subconjunto); y, segundo, realizar un intercambio semialeatorio entre los distintos subconjuntos (sin que éstos acaben estando repetidos en el mismo subconjunto) tantas veces como establece la Ec. 1, evitando así que muchos usuarios compartan todos sus subconjuntos lo que invalidaría la mayor parte de los beneficios de la redundancia.

$$\max\{\text{N.º Redes, PD}\} \cdot \text{PD}^2. \tag{1}$$

En el proceso de adición de nuevos usuarios, análogamente, la distribución de las nuevas redes consiste en añadir al propio usuario y después completar el subconjunto hasta obtener los PD usuarios, eligiendo entre el resto de forma aleatoria y evitando la repetición dentro del conjunto.

IV-C. Implementación de la PoC

La PoC implementada se encuentra en el siguiente repositorio público y accesible mediante licencia Creative Commons Zero: https://github.com/GonGMiranda/AG-RN.

Esta PoC exige los siguientes requisitos para poder validar el funcionamiento del AG-RN:

- Requisitos funcionales: Permitir crear un gestor que actúe sobre un conjunto inicial de datos con parámetros PD y Rd para realizar el proceso de distribución sacudida, el entrenamiento de redes, y la adición y eliminación de usuarios; guardar su estado y recuperarlo para las redes ya entrenadas; y crear un predictor que actúe sobre el gestor.
- Requisitos no funcionales: Permitir la predicción en paralelo de distintas instancias del gestor sin que se produzcan interrupciones entre ellas.

En cuanto a la interacción, se hace uso de un patrón de diseño clásico como es el Modelo-Vista-Controlador (MVC). Este diseño permite separar la interacción con el usuario, del control y modificación de los datos, y de su persistencia. Se utiliza como modelo el propio sistema de ficheros de Windows, almacenando el propio gestor en ficheros binarios y las redes entrenadas en ficheros HDF5. En referencia a la vista no se ha desarrollado una interfaz puesto que se hace uso de los cuadernos de Jupyter que permiten actuar de forma sencilla y visual para mostrar los resultados. El controlador está programado en Python, haciendo uso de la librería Keras, por su alto nivel de abstracción, y de TensorFlow, por ser la opción por defecto. En la Fig. 6 se puede observar el MVC aplicado en esta PoC.

Además, se crea una clase Predictor con un método sobre la que se ha aplicado el patrón Monostate [22], para predecir huellas dactilares. De esta manera los predictores se crean de forma paralela con el único propósito de hacer consultas sobre instancias distintas de la clase NNManager del gestor, sin posibilidad de modificar su estructura ni forzar la creación y borrado de usuarios o el entrenamiento de redes. Así pues, idealmente, el uso de un Predictor será único para cada huella que se quiera autenticar. Esto permitiría que cada uso contenga la red actualizada al momento de su utilización, aunque también puede realizarse la actualización periódicamente.

V. Evaluación

Con la intención de determinar que la implementación de la PoC ha sido realizada correctamente, se han diseñado una serie de casos de prueba como método de evaluación. La demostración de que los resultados de los casos de prueba han sido satisfactorios se encuentra en los cuadernos de Jupyter del repositorio de la implementación. Por tanto, a continuación sólo se mencionan tales casos de prueba, que en particular todos han sido verificados como satisfactorios lo que permite concluir que el software desarrollado del AG-RN en la PoC es correcto:

- Creación del conjunto de datos (fichero FPPreprocess.ipynb): La lectura de una imagen y su codificación es correcta, extrayéndose la etiqueta que identifica a su usuario. Además, la imagen codificada y su etiqueta se guardan y recuperan adecuadamente. Asimismo, se leen y codifican todas las imágenes y etiquetas tanto de una carpeta de huellas reales como de una de huellas alteradas, guardándose en los ficheros correspondientes.
- Desarrollo CNN (ficheros RNC.ipynb y Training stats.ipynb): El modelo de incremento de datos produce correctamente 9 imágenes distintas a partir de la original. El modelo de CNN se crea y entrena debidamente con un funcionamiento apropiado de la red tanto para una prueba positiva como para una negativa de una huella aleatoria. Además, se genera un conjunto de datos con huellas reales y alteradas de un número determinado de usuarios y se verifica que el FRR y el FAR son menores al 5 % y al 2 % respectivamente. Finalmente, se generan varios conjuntos de datos aleatorios y se entrenan distintas redes con ellos satisfactoriamente almacenando los resultados en un archivo Excel.
- Desarrollo AG-RN (fichero NNManager.ipynb): Se crea un nuevo NNManager configurado con los parámetros indicados y se carga adecuadamente, comprobándose que la distribución de los usuarios es correcta. Así pues, se verifica la autenticación de un usuario obteniéndose la predicción cuando las redes han sido parcialmente entrenadas y el rechazo en el otro caso. Además, se entrenan todas las redes del gestor y se guardan en sus respectivos ficheros HDF5 para verificar la autenticación negativa tanto para un usuario con una huella errónea como para otro cuya huella no pertenece al gestor correspondiente. Por último, se añade correctamente a un nuevo usuario, pudiéndose autenticar después de forma adecuada, y también se borra a un usuario satisfactoriamente.
- Predictor (fichero PredictorAG-RN.ipynb): Se verifica que el predictor se crea sobre la red entrenada y que

autentica al usuario convenientemente. Además, se comprueba que no se autentica a un usuario inválido, y que cuando el predictor no se crea se alerta mediante el error de que el gestor no existe.

VI. Conclusiones

La autenticación de usuarios es, actualmente, uno de los procesos más importantes en la seguridad de la información. Por tanto, cada vez es más común el uso de métodos más difíciles de vulnerar como el uso de factores biométricos. No obstante, a veces la inmediatez en su aplicación se encuentra lejos de lo deseable, como por ejemplo al autenticar a un grupo de usuarios en una organización mediante el uso de huellas dactilares.

En este trabajo se presenta AG-RN, un algoritmo que permite la adición o eliminación de usuarios a un sistema de reconocimiento biométrico basado en redes neuronales sin ser necesario un reentrenamiento de la totalidad de la red. La potencia de este algoritmo está basada en la capacidad de gestión de varias redes neuronales, ofreciendo un balance óptimo entre el tamaño de cada una de estas redes que están entrenadas con subconjuntos diferentes del total de individuos, y la redundancia de los datos de cada individuo distribuido entre varias de estas redes. Así pues, se ha comprobado que se puede crear una red neuronal basada en capas de convolución y "Max Pooling" capaz de identificar si dos huellas dactilares son o no del mismo dedo de un mismo individuo, incluso aunque la lectura de las mismas no sea perfecta o que la huella haya sufrido alteraciones que conlleven una modificación visual en el patrón general formado por las *minutiaes*. Asimismo, se ha mostrado que el rendimiento de esta red empeora al ser entrenada con conjuntos más reducidos de individuos y que el tiempo de entrenamiento aumenta de forma lineal al aumentar el conjunto de usuarios.

Además, para demostrar la viabilidad de este algoritmo, se ha desarrollado una PoC en Python que hace uso de un modelo basado en una CNN que permite la autenticación de usuarios mediante su huella dactilar. En particular, esta PoC utiliza una base de datos preexistente de un conjunto de imágenes de huellas dactilares, y no incluye una interfaz de usuario sino que las pruebas pertinentes se han realizado en cuadernos de Jupyter, ya que son una forma más sencilla pero también visual de mostrar los resultados.

El repositorio del sistema de autenticación presentado es público y tiene una licencia libre con la intención de servir como material de apoyo a otros investigadores y, también, que pueda ser completada en el futuro con otras nuevas funcionalidades y diversas mejoras como, por ejemplo, un proceso de distribución de redes más eficiente, la inclusión de otros métodos de autenticación o la extensión del algoritmo a opciones multifactor.

Finalmente, se considera que el algoritmo propuesto puede dar solución a un problema real, demostrando que un sistema de autenticación biométrica basado en una implantación del AG-RN permite registrar o eliminar a un usuario prácticamente de forma inmediata, sin afectar al funcionamiento general del sistema o a los usuarios creados anteriormente.

Referencias

[1] A. K. Jain, A. Ross y S. Prabhakar: "An Introduction to Biometric Recognition", en *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4-20, 2004. DOI: 10.1109/TCSVT.2003.818349.

[2] R. Noval, y F. Perales: "Adaptative Templates in Biometric Authenticacion", en *WSCG International Conferences in Central Europe on Computer Graphics, Visualization and Computer Vision*, 2014.

[3] J. Feng, A. K. Jain y A. Ross: "Detecting Altered Fingerprints", en *2010 20th International Conference on Pattern Recognition*, pp. 1622-1625, 2010. DOI: 10.1109/ICPR.2010.401.

[4] E. Tabassi, T. Chugh, D. Deb, y A. K. Jain: "Altered Fingerprints: Detection and Localization", en *repositorio arXiv*, Cornell University, 2018. ArXiv: 1805.00911v2 [cs.CV].

[5] J. R. Hamlet y L. G. Pierson: "Multi-Factor Authentication", *Sandia Corporation*, Patente US8868923B1, Estados Unidos, 2010.

[6] S. H. Khan, M. A. Akbar, F. Shahzad, M. Farooq y Z. Khan: "Secure Biometric Template Generation for Multi-Factor Authentication", en *Pattern Recognition*, vol. 48, n. 2, pp. 458-472, 2015. DOI: 10.1016/j.patcog.2014.08.024.

[7] España, Ministerio de Justicia: "Real Decreto 1720/2007, de 21 de diciembre, por el que se aprueba el Reglamento de desarrollo de la Ley Orgánica 15/1999, de 13 de diciembre, de protección de datos de carácter personal", en *Boletín Oficial del Estado*, 19 de enero de 2008, n. 17, pp. 4103-4136. Referencia: BOE-A-2008-979.

[8] F. Galton: "Finger Prints", Londres: *McMillan*, 1892.

[9] H. Abdulkareem: "Fingerprint Identification System using Neural Networks", en *College of Engineering Journal (NUCEJ)*, Nahrain University, vol. 15, pp. 234-244, 2012.

[10] F. Pernus, S. Kovacic y L. Gyergyek: "Minutiae-Based Fingerprint Recognition", en *Proceedings of the Fifth International Conference on Pattern Recognition*, pp. 1380-1382, 1980.

[11] D. Maio y D. Maltoni: "Direct Gray-Scale Minutiae Detection in Fingerprints", en *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 1, pp. 27-40, 1997. DOI: 10.1109/34.566808.

[12] N. Zaeri: "Minutiae-Based Fingerprint Extraction and Recognition", en *Biometrics*, 2011. DOI: 10.5772/17527.

[13] A. Farina, Z. M. Kovács-Vajna y A. Leone: "Fingerprint Minutiae Extraction from Skeletonized Binary Images", en *Pattern Recognition*, vol. 32, n. 5, pp. 877-889, 1999. DOI: 10.1016/S0031-3203(98)00107-1.

[14] S. Narwal y D. Kaur: "Comparison between Minutiae Based and Pattern Based Algorithm of Fingerprint Image", en *International Journal of Information Engineering and Electronic Business*, vol. 8, n. 2, pp. 23-29, 2016. DOI: 10.5815/ijieeb.2016.02.03.

[15] A. Chowdhury, S. Kirchgasser, A. Uhl y A. Ross: "Can a CNN Automatically Learn the Significance of Minutiae Points for Fingerprint Matching?", en *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 340-348, 2020. DOI: 10.1109/WACV45572.2020.9093301.

[16] T. R. Borah, K. K. Sarma y P. H. Talukdar: "Fingerprint Recognition using Artificial Neural Network", en *International Journal of Electronics Signals and Systems (IJESS)*, vol. 3, n. 1, pp. 98-101, 2013.

[17] F. A. A. Minhas, M. Arif y M. Hussain: "Fingerprint Identification and Verification System using Minutiae Matching", en *National Conference on Emerging Technologies*, pp. 141-147, 2004.

[18] L. Chen, F. Rottensteiner y C. Heipke: "Feature Descriptor by Convolution and Pooling Autoencoders", en *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XL-3/W2, pp. 31-38, 2015. DOI: 10.5194/isprsarchives-XL-3-W2-31-2015.

[19] Y. I. Shehu, A. Ruiz-García, V. Palade y A. James: "Sokoto Coventry Fingerprint Dataset (SOCOFing)", en *repositorio arXiv*, Cornell University, 2018. ArXiv: 1807.10609v1 [cs.CV].

[20] A. Jung: "Imgaug", en *repositorio GitHub*, 2020. URL: https://github.com/aleju/imgaug.

[21] D. P. Kingma y J. L. Ba: "Adam: A Method for Stochastic Optimization", en *3rd International Conference for Learning Representations (ICLR)*, 2015. ArXiv: 1412.6980 [cs.LG].

[22] Wiki: "Monostate Pattern" [recurso web]. URL: https://wiki.c2.com/?MonostatePattern.

Sesión de Investigación B2:
Gobierno y gestión de riesgos

Sesión de Investigación B2: Gobierno y gestión de riesgos
http://doi.org/10.18239/jornadas_2021.34.42

Verificación remota de controles de seguridad en contextos de seguridad adaptativa

Miguel Calvo
ORCID: 0000-0003-2418-1830
ETSII, Universidad Rey Juan Carlos
28933 Móstoles, Madrid
miguel.calvo@urjc.es

Marta Beltrán
ORCID: 0000-0002-1689-7479
ETSII, Universidad Rey Juan Carlos
28933 Móstoles, Madrid
marta.beltran@urjc.es

***Resumen*—La aparición de paradigmas muy escalables, distribuidos y heterogéneos como el web, el móvil, el *cloud* o el del Internet de las cosas ha hecho que la seguridad adaptativa sea cada vez más importante. Este tipo de seguridad permite que los controles de seguridad desplegados se comporten de manera inteligente y que se adapten de manera dinámica a su contexto, dependiendo del riesgo real que se corre y del que se tolera en ese momento. Esta adaptación puede afectar a su diseño, despliegue, configuración o modo de uso. En este trabajo se propone un mecanismo de verificación remota que permite asegurar que controles de seguridad distribuidos y, en muchas ocasiones, poco supervisados, se encuentran en un punto de partida seguro antes de realizar cualquier adaptación, ya que se trata de un proceso que puede llegar a ser costoso en tiempo y en recursos y que debe evitarse si es innecesario o poco efectivo. Este mecanismo, implementado por capas, garantiza flexibilidad, separación de dominios e inicio autorizado, y puede aplicarse a diferentes tipos de controles (se trata de un mecanismo genérico). La propuesta realizada se ha validado y evaluado en un caso de uso real.**

***Index Terms*—Controles dinámicos, Seguridad adaptativa, Verificación remota**

Tipo de contribución: *Investigación original*

I. Introducción

En escenarios web, móviles, *cloud*, *Internet of Things* (IoT) o de Industria 4.0; ningún diseño, despliegue, configuración o modo de uso de un control de seguridad es capaz de hacer frente al riesgo en todas las situaciones posibles. Por ello, la seguridad ya no puede ser estática; en muchos contextos no es posible definir una estrategia de seguridad inalterable, sino que hay que permitir a los administradores, gestores o responsables de seguridad adaptar los controles y contramedidas para alcanzar los niveles de seguridad adecuados en cada momento, en función de los cambios que se producen en los activos protegidos, la evolución de su contexto, y en última instancia, del riesgo percibido y del riesgo que la organización o los usuarios están dispuestos a tolerar.

Un enfoque que se ha propuesto en los últimos años para hacer frente a esta heterogeneidad, incertidumbre y dinamismo es el adaptativo, capaz de ajustar controles de seguridad aplicados al contexto y estado del activo, y al riesgo que estos factores implican en un momento dado [1], [2]. Se suele trabajar con modelos que modifiquen de forma autónoma el diseño, despliegue, configuración o modo de uso de los controles, tomando estas decisiones de adaptación con diferentes criterios y estrategias. Nos estamos acostumbrando a que un sistema de control de accesos adapte el número y tipo de autenticadores solicitados en función del riesgo que se corre al autorizarnos a realizar una determinada tarea. O a que un *Web Application Filter* modifique su comportamiento según del estado de Internet en un momento concreto.

Algo que se observa en todas estas soluciones dinámicas o adaptativas es que la toma de la decisión de adaptación y su implementación tienen un coste asociado: en tiempo y recursos. Por este motivo suelen implementarse en arquitecturas distribuidas, en varias capas que se reparten el trabajo a realizar y que pueden ejecutarse más próximas o más lejanas a los controles adaptados en función de los requisitos de latencia, ancho de banda, privacidad, etc.

Sin embargo, en la mayor parte de los casos no se realizan comprobaciones previas a la adaptación para verificar que los controles no se han visto comprometidos y que el punto de partida de la adaptación es el esperado (correcto o permitido). En este trabajo se propone un mecanismo de verificación remota que se puede incorporar en estas soluciones de seguridad adaptativa. El mecanismo propuesto también se implementa por capas para que sea fácil de añadir a las soluciones actuales. Permite verificar el correcto estado interno de los controles adaptados (en concreto, la integridad de su memoria), pero también de los componentes de la solución de seguridad adaptativa que se despliegan de manera remota. De esta forma se puede estar suficientemente seguro, por ejemplo, de que las decisiones de adaptación se han tomado con datos actuales y fiables. Se trata de un mecanismo genérico, que se puede emplear para cualquier tipo de control de seguridad, y que además garantiza la flexibilidad necesaria para el contexto para el que se propone y que sólo se puede iniciar de manera autorizada. En este trabajo se implementa un primer prototipo de la propuesta realizada para su validación y evaluación.

El resto del artículo se estructura de la siguiente manera. En la sección II se resume el estado del arte en seguridad adaptativa y verificación remota. A partir de las limitaciones encontradas en este estado del arte se explica la motivación para este trabajo. En la sección III se presenta la arquitectura habitual en las soluciones de seguridad adaptativa y se resumen las asunciones que se han tenido en cuenta para realizar esta investigación. La sección IV propone el mecanismo de verificación remota que supone la principal contribución de este trabajo, presentando su funcionamiento a alto nivel y definiendo todos los protocolos necesarios para su ejecución. La sección V presenta el primer prototipo que implementa este mecanismo para validarlo y evaluarlo. Finalmente, en la sección VI se discuten las principales conclusiones de este trabajo y algunas líneas interesantes para continuar investigando en el futuro.

II. Estado del arte y motivación

II-A. Controles dinámicos y seguridad adaptativa

Los controles de seguridad dinámicos se pueden reconfigurar y adaptar (incluso en tiempo real) en función del contexto, de acuerdo con las amenazas que se sufren o según el riesgo tolerado. Diferentes investigaciones han abordado esta idea, proponiendo controles aplicables según la situación (por ejemplo, [3] presenta una arquitectura de seguridad para SDN o *Software Defined Networking*) o según el contexto (como en [4], que muestra una solución para la configuración dinámica de recursos en *Cloud*). A pesar de ello, la seguridad dinámica continúa siendo un concepto ambiguo, sin una definición y un alcance claros. La seguridad adaptativa está estrechamente relacionada y, a menudo, es un concepto que aparece en las diferentes investigaciones que exploran los controles dinámicos de seguridad.

La seguridad adaptativa se centra en analizar eventos ([5], [6]) para proteger determinados recursos o sistemas de las amenazas antes de que se produzcan o cuando están sucediendo, permitiendo, de forma automática, aumentar o disminuir el nivel de protección. Esta seguridad se basa en una lógica de adaptación que suele ser PDCA (*Plan-Do-Check-Act*) [7] (utilizada en [8] para la gestión de redes SDN/NFV - *Network Function Virtualization*), OODA (*Observe-Orient-Decide-Act*) [9] o MAPE-K (*Monitor, Analyse, Plan, Execute – Knowledge*) [10], que es la más extendida de las tres. A su vez, la seguridad adaptativa puede clasificarse en seguridad sensible al contexto, cuyo enfoque se basa en el conocimiento del entorno (con diversos ámbitos de aplicación que van desde el control de accesos en [11], hasta la criptografía en [12]); y en seguridad incremental o inteligente, cuyo enfoque suele combinar diferentes técnicas y herramientas (*Big Data*, *Analytics*, etc.) para detectar desviaciones del comportamiento normal y actuar en consecuencia (también con aplicaciones muy variadas como la detección de anomalías en [13] o la autenticación y el cifrado en sistemas *Smart Grid* en [14]).

En relación con la seguridad adaptativa, y, aunque no es lo mismo, también se pueden encontrar investigaciones recientes en seguridad basada en el riesgo. La idea es identificar los diferentes riesgos que corre cada activo de una organización, intentando adaptar en cada caso los mecanismos de mitigación que aplican al presupuesto disponible y al riesgo tolerado. Las utilidades de la seguridad basada en el riesgo son tan variadas como las de seguridad adaptativa, y en muchos casos solapan. Destacan el control de accesos (como [15] en IoT) o la toma de decisiones ([16] para el enrutamiento en SDN).

II-B. Mecanismos de verificación remota

La verificación remota conlleva la existencia de al menos dos roles, verificador y probador, normalmente distribuidos geográficamente. El verificador, envía un desafío aleatorio (que puede contener un *nonce* o *time-stamp*, el identificador de la memoria verificada o las medidas requeridas, la rutina de verificación a ejecutar, etc.) al probador [17]. Tras recibirse el desafío, el probador tratará de ejecutar la rutina de verificación indicada, utilizando medidas propias para representar su estado interno, por ejemplo, considerando su configuración hardware, el contenido de la memoria o el ejecutable cargado y traducirá el estado para su posterior envío al verificador.

La mayoría de los mecanismos de verificación remota propuestos utilizan un *hash* o *checksum* para poder representar con una longitud fija el estado interno del probador. Se ha demostrado que esta técnica es eficiente en términos de procesamiento, almacenamiento, transmisión y comparación (el verificador únicamente necesita almacenar el *hash* o *checksum* de los estados válidos para llevar a cabo la verificación [18]). Entre las soluciones propuestas, existen las que generan el *hash* o *checksum* mediante software ([19]), comprobándose únicamente el contenido de la memoria; mediante hardware ([20], [21]), cuando el contenido de la memoria puede verse comprometido (siendo estas últimas más caras y menos flexibles, pero también más seguras); o híbridas/semánticas ([22]), apoyadas en la virtualización de hardware para superar las desventajas de la verificación basada en hardware.

Todas estas soluciones, se centran en la verificación de binarios (se resume el estado del probador mediante los binarios del software cargado en su memoria). Sin embargo, varias investigaciones han señalado algunas de sus limitaciones, como, por ejemplo, que pequeños cambios en el software obligan a modificar el estado de referencia o válido en el verificador (lo cual no siempre es sencillo); o que esta técnica puede revelar datos sobre la configuración del probador, el software instalado, etc. (y esto, a su vez, puede conllevar la filtración de información confidencial o a la exposición a nuevas amenazas de seguridad). Para intentar remediar esto, se ha explorado la verificación remota basada en propiedades ([23]) y la verificación remota basada en el comportamiento ([24]); aunque los mecanismos basados en la verificación de binarios siguen siendo los más utilizados por la heterogeneidad y el dinamismo de las propiedades y comportamientos del tiempo de ejecución.

II-C. Motivación

Teniendo en cuenta el estado del arte presentado, se puede plantear un caso de uso que muestra claramente la motivación de este trabajo. Los cortafuegos tradicionales han demostrado en el pasado ser una potente solución para proteger las redes corporativas contra el tráfico no deseado o las intrusiones, el código potencialmente dañino, etc. Las reglas de los cortafuegos son especialmente difíciles de mantener actualizadas. Estas reglas son complejas de adaptar o reconfigurar cuando se modifica la topología de la red, cuando se ofrecen nuevos servicios dentro de la red corporativa o cuando las amenazas específicas son más probables durante un periodo de tiempo determinado. La razón principal es que es necesario configurar hasta varios cientos o miles de reglas en los cortafuegos actuales. Una actualización puede requerir un análisis en profundidad de todas ellas para determinar qué reglas hay que cambiar y cómo hacerlo. Los responsables de seguridad no suelen tener tiempo para realizar este análisis, por lo que las reglas suelen permanecer sin cambios (además, sin relación con el riesgo afrontado o el tolerado). Esto implica que, o bien aplican políticas demasiado restrictivas (causando problemas de usabilidad, restringiendo la funcionalidad o aumentando los costes, por ejemplo), o bien no protegen la red adecuadamente (lo que acaba provocando incidentes de seguridad).

Si se opta por utilizar una solución de seguridad adaptativa, estos cortafuegos podrán modificar dinámicamente estas reglas de manera autónoma, siguiendo un modelo MAPE-K

u otra alternativa similar de las presentadas como estado del arte. Con cualquiera de ellas la adaptación de las reglas del cortafuegos implicará unos costes: recoger los requisitos de seguridad de la organización, monitorizar el estado de la red y de los cortafuegos, detectar los cambios que se producen en estos aspectos y que llevan a lanzar la adaptación, analizar las reglas actuales y decidir los cambios que deben hacerse para gestionar el riesgo adecuadamente en la nueva situación e implementar estos cambios. ¿Qué sentido tiene realizar todas estas tareas, que consumen tiempo y recursos, sin verificar antes que los cortafuegos que van a sufrir la adaptación están en el punto de partida esperado y que su estado interno es el correcto? ¿Para qué invertir tiempo y esfuerzo en adaptar un control de seguridad obsoleto o comprometido? ¿Y si lo que está comprometido es el módulo de la solución de adaptación que se encuentra desplegado de manera remota? En este caso se habría tomado la decisión de adaptación con datos parcialmente incompletos o incompletos directamente.

El mecanismo propuesto en este trabajo resuelve este problema, incorporando a las soluciones de seguridad adaptativa la capacidad de realizar verificación remota. Ninguna de las soluciones propuestas hasta el momento tiene esta capacidad. Además, todas las investigaciones previas que se centran en proponer mecanismos de verificación remota lo hacen en arquitecturas de dos capas (una para verificadores y otra para probadores), con patrones de verificación uno a uno; no permitiéndose otras topologías, patrones de interacción o esquemas de delegación. Y esto es necesario para resolver el problema que nos ocupa.

Por este motivo, el mecanismo propuesto en este trabajo debe cumplir con estas propiedades:

- Basado en binarios: debe verificar los binarios cargados en la memoria del componente o control que se verifica. En el futuro se añadirá la capacidad de verificar otros aspectos, pero en este primer trabajo se ha considerado que es el aspecto determinante, el que cómo mínimo se debe comprobar antes de comenzar con un proceso de adaptación.
- Por capas: debe permitir a soluciones de seguridad adaptativa que se ejecutan en plataformas sin límite de recursos, verificar remotamente a otros componentes de la propia solución que se ejecutan de manera distribuida y delegar en estos, cuando sea necesario, la verificación de los controles de seguridad, evitando la comunicación directa con estos controles. Es decir, debe permitir como mínimo arquitecturas de tres capas y verificaciones de uno (la solución de seguridad adaptativa) a muchos (componentes de esta solución que se encuentren distribuidos y los controles que pueden ser adaptados).
- Flexibilidad: debe soportar diferentes alternativas para comunicarse con las soluciones de seguridad adaptativa y para acceder a la memoria de los componentes o controles verificados.
- Generalidad: debe diseñarse e implementarse con capacidad para trabajar con diferentes componentes y controles, independientemente de las tecnologías, arquitecturas, fabricantes y vendedores.
- Separación de dominios: debe aislar el dominio de aplicación (el componente o control y su funcionalidad original) y el dominio de confianza (que implementa el propio mecanismo de verificación). Si el componente o control verificado se ven comprometidos, la solución de seguridad adaptativa puede seguir realizando mediciones precisas, no manipuladas y fiables.
- Inicio autorizado: sólo debe permitir que los verificadores autorizados realicen la verificación para evitar la fuga de información sensible a posibles adversarios a través de las mediciones de estado internas.

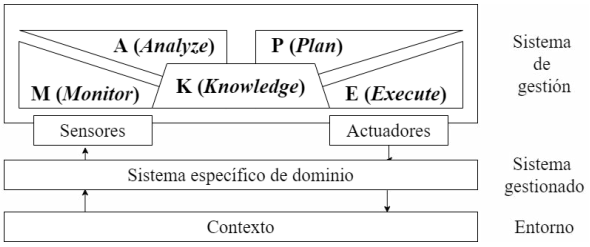


Figura 1. Soluciones de seguridad adaptativa.

III. Modelo para la seguridad adaptativa: Arquitectura y asunciones

Evaluando las soluciones de seguridad dinámica o adaptativa que se han propuesto hasta el momento (mencionadas en la sección II) se puede observar que casi todas siguen el modelo MAPE-K. Este modelo implica que algún componente de la solución se encarga de monitorizar los activos que deben protegerse, los controles que pueden utilizarse (y adaptarse) para ello y su contexto (función *Monitor*). Otros componentes analizan (función *Analyze*) esta información recogida en tiempo real para tomar decisiones acerca de las adaptaciones necesarias (función *Plan*). Y otros se encargan de materializar las adaptaciones decididas (función *Execute*) para conseguir el objetivo planteado, normalmente, gestionar el riesgo y mantenerlo por debajo de unos determinados niveles tolerados. Para todo esto suele utilizarse un conocimiento histórico y compartido entre los diferentes componentes (función *Knowledge*) que la propia solución puede ir generando o que se puede obtener de fuentes externas. Todo esto se resume en la figura 1.

Desde el punto de vista del despliegue y ejecución de todos estos componentes, en este trabajo se asume una arquitectura de tres capas para las soluciones de seguridad adaptativa (figura 2):

- Componentes de seguridad adaptativa centrales: Estos servicios (o recursos o aplicaciones) proporcionan capacidades de computación y comunicación, de almacenamiento, gestión y visualización de datos, etc. Desde estos servicios se ofrece, por ejemplo, un panel o consola de la solución de seguridad adaptativa para los administradores, gestores o responsables de seguridad. Son las entidades que necesitan verificar que otros componentes remotos de la solución y que los controles que se desea adaptar parten de un estado correcto o permitido antes de realizar ciertas tareas específicas. Se supone que estos servicios se ejecutan en servidores con muchos recursos (en términos de cómputo, memoria, almacenamiento, ancho de banda y energía), aislados o situados en grandes

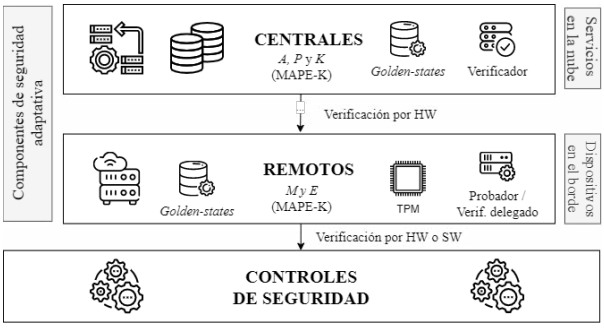


Figura 2. Despliegue de soluciones de seguridad adaptativa en tres capas.

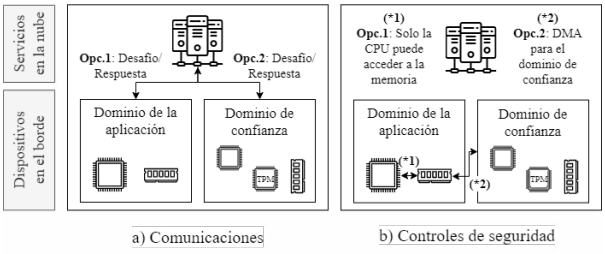


Figura 3. Alternativas propuestas para la comunicación y el acceso a la memoria.

centros de datos en la nube. Suelen ser los componentes que ejecutan las funciones *A*, *P* y *K* de MAPE-K.

- Componentes de seguridad adaptativa remotos: Estos componentes pueden ejecutarse en pasarelas, controladores independientes o integrados, terminales de detección o incluso *clusters* o microcentros de datos. En general, en dispositivos *edge* o de borde, cerca de los controles que se desea adaptar (y no en centros de datos lejanos). Los aspectos significativos que los convierten en dispositivos de borde son:
  1. Se sitúan cerca de los controles y más lejos de los servicios en la nube.
  2. Esta proximidad permite una comunicación eficiente y de baja latencia con estos controles. Por ello, son ideales para monitorizar los controles y ejecutar las adaptaciones. Es decir, para los elementos encargados de las funciones *M* y *E* de MAPE-K.
  3. Suelen contar con recursos más limitados para su ejecución que los servicios en la nube, por este motivo no son adecuados para realizar cómputo complejo o tareas que consuman mucha memoria.
  4. Actúan como intermediarios, *brokers* o *proxies* para los controles cuando necesitan interactuar con los servicios en la nube.
  5. Son capaces de incluir/integrar el mecanismo propuesto para verificación remota. Para ello necesitan incorporar un TPM (*Trusted Platform Module*).
- Controles de seguridad: Componentes hardware o software que pueden ser adaptados y a los que no se presupone ninguna característica ni capacidad concreta. Es decir, no se supone que dispongan de almacenamiento seguro para guardar cualquier tipo de secreto o credencial ni que puedan ejecutar funciones criptográficas robustas, por ejemplo.

IV. Mecanismo de verificación remota

IV-A. Explicación a alto nivel

La figura 2 muestra una descripción a alto nivel del mecanismo propuesto. Los componentes centrales pueden utilizar verificación remota basada en hardware cuando verifican el estado interno de los dispositivos de borde en los que se ejecutan el resto de los componentes de la solución (los componentes *M* y *E* habitualmente). Esto es así porque, como hemos establecido en la sección anterior, suponemos que hay un TPM disponible. Cuando los servicios en la nube necesitan verificar el estado interno de uno o más controles de seguridad, antes de tomar o de hacer efectiva una decisión de adaptación, por ejemplo, la verificación se delega en estos componentes que se ejecutan en los dispositivos de borde. En este caso se requiere un mecanismo de verificación basado en software porque no sería realista suponer que todos los controles de seguridad incorporan un TPM o se ejecutan en plataformas hardware que los incorporan.

Dentro de un dispositivo de borde, el dominio de aplicación consiste en un microcontrolador o procesador, diferentes tipos de memoria, periféricos, etc. En este dominio es donde reside la funcionalidad del dispositivo de borde, por tanto, los programas de aplicación y los datos de los componentes de la solución de seguridad adaptativa (que pueden estar monitorizando al control o a otros activos, calculando métricas de riesgo sencillas o reconfigurando el control, por ejemplo). Podría estar bajo el control de un adversario por completo, no se debe confiar en él por defecto. El dominio de confianza, por otro lado, es una raíz de confianza de hardware (un TPM), responsable de actuar como verificador para el servicio en la nube (confiando en la verificación basada en hardware) y como verificador delegado para los controles de seguridad (confiando en la verificación basada en software si estos controles no tienen una raíz de confianza hardware disponible).

Se pueden implementar diferentes arquitecturas para añadir el mecanismo de verificación a un dispositivo de borde en función de dos aspectos diferentes. El primero, las comunicaciones: el dominio de confianza puede utilizar su propia pila TCP/IP para comunicarse con el servicio en la nube y con los controles de seguridad. O bien, puede necesitar utilizar la pila TCP/IP ofrecida por el dominio de aplicación. El segundo, el acceso a la memoria: el dominio de confianza puede necesitar pasar por el controlador o procesador del dominio de aplicación para acceder a la memoria verificada (que es la memoria del dominio de aplicación). O bien, el acceso directo a la memoria (DMA) puede estar disponible para el dominio de confianza con el fin de realizar accesos directos a la memoria del dominio de aplicación. La figura 3 muestra estos diferentes enfoques, ambos ortogonales y soportados por nuestra propuesta para proporcionar la flexibilidad y generalidad deseadas.

IV-B. Definición de protocolos

IV-B1. Alta y actualización: La fase de alta (*enrolment*) o inicialización es necesaria, previa al despliegue de los diferentes componentes de la solución de seguridad adaptativa

en dispositivos de borde. La capacidad de realizar esta inicialización dinámicamente y de manera remota no es el objetivo de este trabajo, por lo que se asume una inicialización estática fuera de línea, lo que implica la creación y el almacenamiento de los *golden-states* (los estados internos o binarios que se consideran correctos o permitidos en cada caso) en los componentes centrales y en el dispositivo de borde, así como la compartición de un secreto *K* (basado en el E_K del TPM para este servicio en la nube específico) y la configuración de generadores de números aleatorios.

También se requiere una fase de actualización cada vez que se actualiza el componente de la solución de seguridad adaptativa que se ejecuta en el dispositivo de borde o el propio control de seguridad, lo que implica cambios en los binarios y, por tanto, en los *golden-states*.

IV-B2. Verificación basada en hardware: El protocolo propuesto se basa en la especificación TCG [25]. Como se muestra en la figura 4, el verificador (el servicio en la nube) envía el reto en el paso 1 con un *nonce* aleatorio (de un solo uso) para evitar ataques de repetición e identificar de forma exclusiva un procedimiento de atestación específico. El desafío también se cifra con el secreto compartido durante la fase de inscripción: $E_K(nonce)$.

Por un lado, esto garantiza la iniciación autorizada: si el verificador no conoce este secreto compartido, el dispositivo de borde no enviará una respuesta de verificación después del paso 2. Por otra parte, este cifrado mitiga los posibles impactos de las amenazas a la seguridad que plantea la opción 1 de la figura 3.a).

La interacción entre la aplicación y los dominios de confianza es necesaria dentro de esta arquitectura para enviar el desafío recibido desde el procesador del dominio de la aplicación al dominio de confianza, y no se puede asumir que esta interacción sea segura. Los dispositivos de borde suelen estar situados en entornos físicos poco protegidos, por lo que un adversario pasivo podría ser capaz de leer todos los mensajes intercambiados por el dominio de aplicación y el dominio de confianza. Pero como el dominio de aplicación podría estar completamente comprometido (*firmware*, controladores, sistema operativo, componentes de la solución de seguridad adaptativa, etc.), un adversario activo también podría ser capaz de modificar, reproducir o borrar todos estos mensajes. Si el desafío está correctamente cifrado (y la respuesta correspondiente también), un adversario no podrá leerlo ni modificarlo. El *replay* se impide mediante el uso del *nonce* antes mencionado. Y el borrado haría fracasar el procedimiento de verificación.

Si las comunicaciones se establecen directamente entre el servicio en la nube y el dominio de confianza (sin pasar por la pila TCP/IP del dominio de aplicación), este cifrado sólo contribuye a un enfoque de defensa en profundidad (por ejemplo, porque no se puede confiar en el canal de comunicación y hay que mitigar los ataques Man in the Middle, etc.).

El mecanismo propuesto se basa en auto mediciones periódicas, es decir, el dominio de confianza actualiza la información sobre el estado interno (verificando el contenido de la memoria y generando el correspondiente *hash* o suma de comprobación) periódicamente (con un período aleatorio dentro de un intervalo configurable) y de forma proactiva, sin

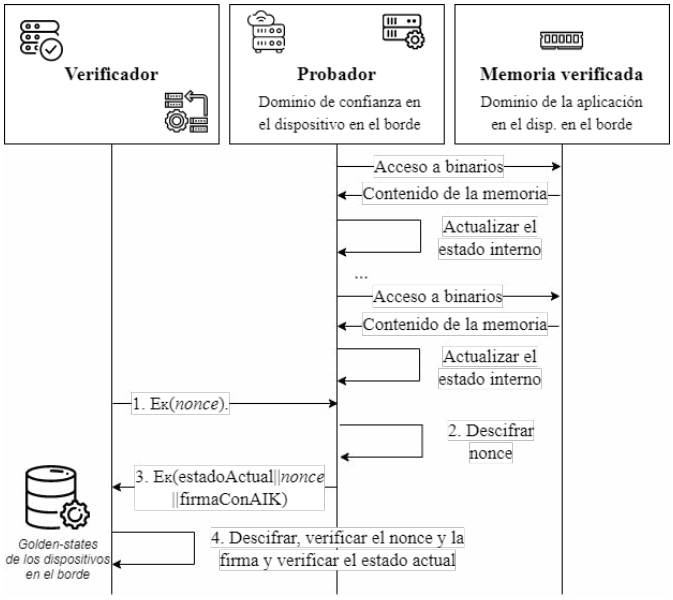


Figura 4. Verificación basada en hardware.

requerir que un verificador envíe un desafío. Por lo tanto, si el dominio de confianza es capaz de descifrar el reto, automáticamente cifra la concatenación del último estado producido con el *nonce* recibido y envía esta respuesta de vuelta al verificador firmada con un AIK (*Attestation Identity Key*) (paso 3):

$E_K(estadoActual||nonce||firmaConAIK)$

Al recibir esta respuesta, el verificador puede comprobar el *nonce* y el estado, y comparar el estado descifrado con el *golden-state* para saber si la verificación tiene éxito o si falla (paso 4). Si el verificador no recibe ninguna respuesta del dominio de confianza después de un periodo de tiempo determinado, el verificador asumirá que no está vivo o que ha sido comprometido, en ambos casos la verificación fallará.

En cuanto a las alternativas de acceso a la memoria, la opción 2 de la figura 3.b) permite realizar auto mediciones periódicas seguras, ya que el dominio de confianza es capaz de cargar direcciones de memoria sin interactuar con el procesador del dominio de aplicación, basándose en un hardware a prueba de manipulaciones. Pero la opción 1 implica un escenario más complejo, ya que un dominio de aplicación comprometido puede responder a las peticiones de acceso a memoria del dominio de confianza, redirigiendo estas peticiones a una sección de memoria (o dispositivo de memoria completo) que almacene los binarios originales para responder con un estado que coincida con el que se considera correcto (ataque de redirección de memoria).

Para evitar esta amenaza que surge cuando los accesos a la memoria no son directos y tienen que realizarse a través del procesador del dominio de la aplicación, se puede trabajar con propiedades de temporización: la latencia de un acceso a la memoria debe ser determinista dentro del dispositivo de borde. Si se mide la latencia de un acceso a la memoria, cualquier ciclo de reloj adicional dedicado por el procesador del dominio de aplicación para engañar al dominio de confianza podría detectarse y la verificación fallaría.

Estos serían los pasos necesarios para acceder a los binarios de la figura 4 (basados en los pasos propuestos por primera

vez en [26]):

1. El dominio de confianza genera una semilla aleatoria.
2. El dominio de confianza calcula una dirección de memoria dentro de la sección de memoria que almacena los binarios que se desea verificar a partir de esta semilla, pide al dominio de aplicación que cargue el contenido de esta dirección y pone en marcha un temporizador.
3. El dominio de aplicación envía el contenido de esta dirección exacta al dominio de confianza.
4. El dominio de confianza repite este proceso hasta que se leen todas las direcciones requeridas (en un orden aleatorio) y se calcula el estado actual.
5. El dominio de confianza detiene el temporizador.

Si el dominio de la aplicación ha sido comprometido y los accesos a la memoria han sido redirigidos en cualquier punto de este procedimiento, el recuento del temporizador sería demasiado alto (porque las instrucciones para comparar direcciones y para reemplazarlas se habrían ejecutado en el dominio de la aplicación comprometido). Esto tiene que resultar en una verificación fallida.

Hay que señalar que la aleatorización de los accesos a la memoria (en lugar de realizarlos secuencialmente como en el caso en que es posible utilizar DMA) tiene un impacto en el cálculo del estado y esto puede dificultar la verificación del *hash* en el verificador (la comparación con el *golden-state*). Para evitar estos problemas, la respuesta al reto debe ser un poco diferente en este caso, concretamente:

$E_K(semilla||estadoActual||nonce||firmaConAIK)$

Por lo tanto, si el verificador tiene los binarios originales (los válidos) almacenados en su base de datos (en lugar de estados o *hashes*), será posible calcular el *golden-state* a partir de la semilla recibida y estos binarios válidos. Los resultados obtenidos pueden ser comparados con el estado recibido para decidir si la verificación tiene éxito o si falla. Esta solución no debería suponer un problema (desde el punto de vista de consumo de recursos) ya que el verificador en este caso es un servicio en la nube que se ejecuta en una plataforma potente en cuanto a recursos.

IV-B3. Verificación basada en software: Como se muestra en la figura 5, cuando la parte de la solución de seguridad adaptativa que se ejecuta en la nube necesita realizar la verificación remota de uno o más controles de seguridad asociados a un dispositivo de borde (paso 1), delega la verificación en esa parte de la solución que se ejecuta en dispositivo de borde y por lo tanto está más cercana a estos controles. Por ejemplo, este sería el reto para verificar los controles 1, 2 y 3:

$E_K(nonce||Id1||Id2||Id3)$

Este dispositivo realiza la verificación remota por software (pasos 3 y 4) a menos que el control pueda realizar la verificación por hardware (porque incluya una raíz de confianza de hardware), el mecanismo específico dependerá del tipo de control verificado, de los recursos de los que dispone y del tipo de conexión de red entre el dispositivo de borde y este control. Las principales alternativas que se pueden utilizar para realizar esta verificación se han mencionado en la sección de trabajos relacionados; [26], [19] o [27] son algunos ejemplos.

Una vez medido el estado actual de uno o más controles, el dispositivo de borde verifica estos estados y genera un informe (paso 5), que consiste en su propio estado y uno o

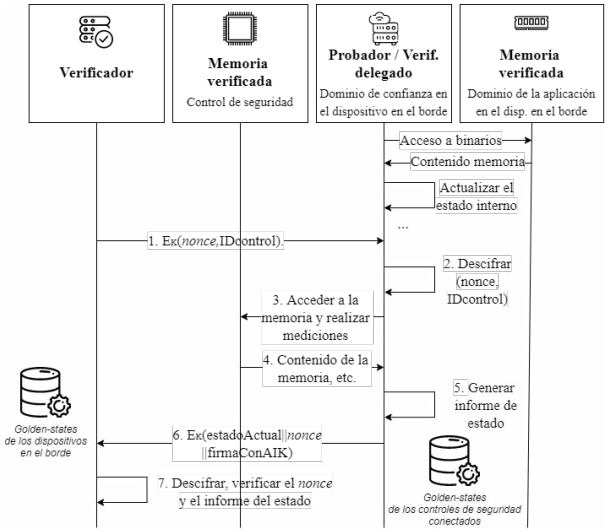


Figura 5. Verificación basada en software.

más resultados de verificación de los controles. Por ejemplo, si el servicio en la nube solicita realizar una verificación remota de los controles 1, 2 y 3, un ejemplo del informe incluido en la respuesta (paso 6) podría ser:

$E_K(estadoActual||Id1OK||Id2OK||Id3fail||nonce||firmaConAIK)$

Al recibir este informe, el verificador puede comprobar primero el *nonce* de la respuesta, la firma con el AIK y el estado actual del dispositivo de borde (paso 7). Si todo es correcto, este informe significa que el dispositivo de borde ha podido verificar con éxito los controles 1 y 2, pero la verificación del dispositivo 3 ha fallado. Si el dispositivo de borde utiliza el modelo de acceso a la memoria que se muestra en la opción 2 de la figura 3.b), se debe utilizar la solución propuesta en la sección anterior, aleatorizar los accesos a la memoria desde una semilla y almacenar los binarios que se consideran correctos en lugar de estados o resúmenes.

V. Prototipo y validación

V-A. Caso de uso

En este trabajo se ha validado el mecanismo propuesto en un proyecto de monitorización remota de personas dependientes ([28]), modificando los *routers* existentes en dicho proyecto para convertirlos en un control adaptativo y añadiendo, además, la verificación remota. En este caso de uso los controles de seguridad que se deben adaptar son los *routers* de acceso a Internet de cada uno de los domicilios monitorizados; Las adaptaciones permiten modificar sus reglas de filtrado de tráfico de red para hacerlas más permisivas o restrictivas y pasar de estrategias de *black-listing* a estrategias *white-listing*. Los dispositivos de borde (ya presentes en el proyecto al que se aplica este mecanismo), se utilizan en cada domicilio para ofrecer servicios de seguimiento de tratamiento farmacológico, para monitorizar constantes vitales, para controlar un cerrojo inteligente del domicilio y para realizar llamadas a los servicios de emergencia cuando una de las personas dependientes ha sufrido una caída o no se ha levantado de la cama. Se trata de una "caja negra" instalada por una compañía de seguros de salud, que, además de lo mencionado, permite

desplegar parte de la solución de seguridad adaptativa y el mecanismo de verificación remota propuesto en este trabajo.

V-B. Implementación

El mecanismo propuesto no está ligado a una implementación concreta, simplemente requiere que la plataforma utilizada para la implementación incluya un TPM y que se puedan desplegar los modelos soportados para las comunicaciones y el acceso a la memoria (figura 3).

El primer prototipo se ha implementado en una FPGA. A medida que los precios de las FPGAs continúan bajando y los lenguajes de alto nivel evolucionan, estas plataformas se han vuelto más accesibles, permitiendo una fácil integración con diferentes tipos de dispositivos de borde y la reconfiguración remota si se necesita realizar cambios. Se ha utilizado una placa SMARTmpsoc (que incluye un ARM Cortex-A53 de cuatro núcleos, una FPGA Xilinx Zynq Ultrascale, 2 GB de DDR4 y un TPM Infineon [29]). El dominio de aplicación (específico para cada caso de uso, en nuestro caso todo lo relacionado con la monitorización de las personas dependientes) se ejecuta en el procesador ARM, mientras que el dominio de confianza se ejecuta en la FGPA y en el chip TPM. Se pueden implementar diferentes alternativas para resolver las comunicaciones y los accesos a la memoria (opciones 1 y 2 en las figuras 3.a) y 3.b)). Cuando el dispositivo de borde es verificado, se verifica la memoria de 2 GB (o la sección ejecutable dedicada a almacenar binarios).

La longitud del *nonce* y del *hash* que resume el estado interno de los dispositivos de borde y de los controles de seguridad se ha fijado en 20 bytes (tamaño de los PCR en el TPM) y la firma basada en los AIK tiene una longitud de 256 bytes. Para este primer prototipo, la longitud de las claves utilizadas para cifrar el reto y la respuesta son de 256 bits, se utiliza TinyAES [30] como esquema de cifrado simétrico y se ha implementado [26] y [31] para realizar la verificación basada en software de los controles de seguridad (del *router)*.

V-C. Resultados experimentales

El primer conjunto de experimentos se ha realizado para validar los protocolos propuestos y sus propiedades requeridas. La verificación remota de los dispositivos de borde (asignando 1 GB a la memoria ejecutable) y de los controles de seguridad (asignando 256 MB a la memoria ejecutable) se ha realizado con un periodo de auto medición entre 50 y 250 ms en los dispositivos de borde (hay que señalar que este periodo es aleatorio para evitar ataques de sincronización). La verificación del software de los controles de seguridad (de toda su memoria) se ha realizado utilizando SWATT [26].

Estas son algunas de las conclusiones más importantes que se pueden extraer del análisis realizado:

- Latencia: Se ha medido el tiempo necesario para realizar las tareas de verificación en el probador y en el verificador. Cuando el componente central de la solución de seguridad adaptativa solicita una verificación de una capa (para chequear la memoria de un dispositivo de borde) tiene que preparar el *nonce*, cifrarlo y, al recibir la respuesta del verificador, descifrarlo, verificar su firma y compararlo con el *golden state* almacenado para ese dispositivo. Se han construido 1000 retos dentro de cada experimento para calcular una media de 8.9 ms para el verificador. La verificación de la respuesta (comprobar la firma y compararla con el patrón dorado) es la tarea que más tiempo consume en este procedimiento.
  Por otro lado, el probador (el dispositivo de borde) tiene que descifrar el *nonce*, acceder a la última auto medición, construir la respuesta, firmarla, cifrarla y enviarla de vuelta al verificador. La latencia más corta se ha observado cuando se dispone de comunicación directa y DMA es de 5.4 ms. Si se utilizan comunicaciones indirectas (el dominio de confianza en el dispositivo de borde no tiene su propia pila TCP/IP y necesita utilizar los mecanismos de comunicación del dominio de aplicación), la latencia media de la respuesta es algo peor, 6.8 ms. Por último, con comunicación directa pero acceso indirecto a la memoria verificada (con aleatorización de acceso a la memoria), el tiempo medio de respuesta es de 5.6 ms. Estas latencias se mantienen prácticamente constantes durante los diferentes conjuntos de experimentos, ya que no dependen del tamaño de la memoria verificada. La explicación se encuentra en la auto medición proactiva y periódica realizada en el probador. De hecho, esta medición evita que el modelo de acceso a la memoria tenga un impacto en las latencias: al recibir un nuevo reto, el probador construye una respuesta con la última auto medición disponible. Esta auto medición fuera de línea, que implica lecturas de memoria y un cálculo de un *hash*, es la parte del procedimiento afectada por el tamaño de la memoria y por el modelo de acceso a esta. De media, la auto medición para 1 GB tarda 0.5 ms con DMA y 6.1 ms sin DMA. Hay que destacar que la naturaleza aleatoria de la verificación de la memoria en este último caso crece como $O(n - ln[n])$ siendo n el tamaño de la memoria verificada (número de palabras).
- Consumo de ancho de banda: Dada la estructura de los retos y respuestas propuestos, la utilización de nuestro mecanismo añade una sobrecarga mínima a la red, cada verificación sólo consume unos pocos bytes (alrededor de 300) de ancho de banda: *nonces* de 20 B, estados de 20 B, firmas de 256 B, IDs de controles (dependiendo del esquema de identificación) y resultados de verificación (OK/fallo) de unos pocos bits.
- Tiempo de respuesta: El tiempo de respuesta para realizar una verificación de una capa (de un dispositivo de borde) se ha medido como el tiempo que transcurre desde que el componente central de la solución de seguridad adaptativa inicia el procedimiento de verificación de este dispositivo de borde hasta que este procedimiento se completa con un resultado exitoso o fallido. Se han construido 1000 retos dentro de cada experimento para calcular una media. Con comunicación directa y DMA, el tiempo de respuesta es de 33.1 ms (se suman a las latencias ya comentadas otras como las de comunicación).
  El tiempo de respuesta para realizar una verificación de dos capas (atestación de un dispositivo de borde y de un *router*) se ha medido como el tiempo que transcurre desde que el componente central inicia el procedimiento de verificación de estos dispositivos hasta que este procedimiento se completa con un resultado

satisfactorio o fallido. El tiempo medio de respuesta (otra vez con experimentos con 1000 retos) es un orden de magnitud mayor, con comunicación directa y DMA en el dispositivo de borde y atestando los 256 MB de memoria en el *router*, es de 235 ms.

En cuanto a la propiedad de flexibilidad, cabe destacar que el prototipo desarrollado podría actualizarse o adaptarse fácilmente para trabajar con nuevos protocolos de verificación basados en software o para soportar nuevos modelos de comunicación o de acceso a la memoria, por ejemplo.

VI. Conclusiones y trabajo futuro

Este trabajo ha presentado un mecanismo de verificación remota que permite que una solución de seguridad adaptativa realice sólo las adaptaciones de los controles de seguridad cuando la situación de partida de estos es segura y la información utilizada para tomar la decisión de adaptación, fiable.

La principal desventaja del mecanismo propuesto podría ser el coste que añade a los dispositivos de borde, ya que es necesario que incluyan un TPM. Esto es algo asumible hoy en día, ya que la mayor parte de estos dispositivos incorporan un chip que puede hacer de raíz de confianza (en cuyo caso, el precio no se incrementaría), o pueden hacerlo fácilmente; en arquitecturas basadas en una FPGA suelen venir incorporado (no teniendo incremento en el coste) y, si estas se basan en Arduino o Raspberry Pi (como suele ser el caso de los microprocesadores o algunas de las placas FPGA), añadir un TPM puede suponer un incremento de entre $30 y $50). Estamos trabajando en el enrolamiento dinámico de los controles a nuestro mecanismo, así como en la recuperación remota de dispositivos de borde y controles en caso de detectar que han sido comprometidos.

Referencias

[1] L. Sion, K. Yskout, D. Van Landuyt, and W. Joosen, "Risk-based design security analysis," in *Proceedings of the 1st International Workshop on Security Awareness from Design to Deployment*, 2018, pp. 11–18.

[2] E. Lara, L. Aguilar, M. A. Sanchez, and J. A. García, "Adaptive security based on mape-k: A survey," in *Applied Decision-Making*. Springer, 2019, pp. 157–183.

[3] V. Varadharajan, K. Karmakar, U. Tupakula, and M. Hitchens, "A policy-based security architecture for software-defined networks," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 4, pp. 897–912, 2019.

[4] F. Graur, "Dynamic network configuration in the internet of things," in *2017 5th International Symposium on Digital Forensic and Security (ISDFS)*, 2017, pp. 1–4.

[5] B. Djoudi, C. Bouanaka, and N. Zeghib, "Model checking pervasive context-aware systems," in *2014 IEEE 23rd International WETICE Conference*, 2014, pp. 92–97.

[6] I. Supriana, K. Surendro, Aradea, and E. Ramadhan, "Self-adaptive cyber city system," in *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, 2016, pp. 1–6.

[7] M. Sokovic, D. Pavletic, and K. K. Pipan, "Quality improvement methodologies–pdca cycle, radar matrix, dmaic and dfss," *Journal of achievements in materials and manufacturing engineering*, vol. 43, no. 1, pp. 476–483, 2010.

[8] K. Shibata, H. Nakayama, T. Hayashi, and S. Ata, "Establishing pdca cycles for agile network management in sdn/nfv infrastructure," in *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, 2015, pp. 619–625.

[9] J. R. Boyd, "The essence of winning and losing," *Unpublished lecture notes*, vol. 12, no. 23, pp. 123–125, 1996.

[10] P. Arcaini, E. Riccobene, and P. Scandurra, "Modeling and analyzing mape-k feedback loops for self-adaptation," in *2015 IEEE/ACM 10th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, 2015, pp. 13–23.

[11] E. Psarra, Y. Verginadis, I. Patiniotakis, D. Apostolou, and G. Mentzas, "A context-aware security model for a combination of attribute-based access control and attribute-based encryption in the healthcare domain," in *Web, Artificial Intelligence and Network Applications*, L. Barolli, F. Amato, F. Moscato, T. Enokido, and M. Takizawa, Eds., 2020, pp. 1133–1142.

[12] M. Fazeen, G. Bajwa, and R. Dantu, "Context-aware multimedia encryption in mobile platforms," in *Proceedings of the 9th Annual Cyber and Information Security Research Conference*. New York, NY, USA: Association for Computing Machinery, 2014, p. 53–56.

[13] L. Fernández Maimó, L. Perales Gómez, F. J. García Clemente, M. Gil Pérez, and G. Martínez Pérez, "A self-adaptive deep learning-based system for anomaly detection in 5g networks," *IEEE Access*, vol. 6, pp. 7700–7712, 2018.

[14] T. Liu, J. Tian, Y. Gui, Y. Liu, and P. Liu, "Sedea: State estimation-based dynamic encryption and authentication in smart grid," *IEEE Access*, vol. 5, pp. 15 682–15 693, 2017.

[15] F. Martinelli, C. Michailidou, P. Mori, and A. Saracino, "Too long, did not enforce: A qualitative hierarchical risk-aware data usage control model for complex policies in distributed environments," in *Proceedings of the 4th ACM Workshop on Cyber-Physical System Security*, ser. CPSS '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 27–37. [Online]. Available: https://doi.org/10.1145/3198458.3198463

[16] B. K. Tripathy, D. P. Das, S. K. Jena, and P. Bera, "Risk based security enforcement in software defined network," *Computers Security*, vol. 78, pp. 321 – 335, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167404818301913

[17] G. Coker, J. Guttman, P. Loscocco, A. Herzog, J. Millen, B. O'Hanlon, J. Ramsdell, A. Segall, J. Sheehy, and B. Sniffen, "Principles of remote attestation," *International Journal of Information Security*, vol. 10, no. 2, pp. 63–81, 2011.

[18] R. V. Steiner and E. Lupu, "Attestation in wireless sensor networks: A survey," *ACM Computing Surveys*, vol. 49, no. 3, pp. 51:1–51:31, 2016.

[19] T. AbuHmed, N. Nyamaa, and D. Nyang, "Software-based remote code attestation in wireless sensor network," in *GLOBECOM 2009-2009 IEEE Global Telecommunications Conference*, 2009, pp. 1–8.

[20] C. Basile, S. Di Carlo, and A. Scionti, "FPGA-based remote-code integrity verification of programs in distributed embedded systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 2, pp. 187–200, 2012.

[21] K. Eldefrawy, G. Tsudik, A. Francillon, and D. Perito, "SMART: Secure and minimal architecture for (establishing dynamic) root of trust," in *NDSS*, vol. 12, 2012, pp. 1–15.

[22] K. Eldefrawy, N. Rattanavipanon, and G. Tsudik, "HYDRA: Hybrid design for remote attestation (using a formally verified microkernel)," in *Proceedings of the 10th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 07 2017, pp. 99–110.

[23] S. Xin, Y. Zhao, and Y. Li, "Property-based remote attestation oriented to cloud computing," *Seventh International Conference on Computational Intelligence and Security*, pp. 1028–1032, 2011.

[24] L. Gu, X. Ding, R. H. Deng, Y. Zou, B. Xie, W. Shao, and H. Mei, "Model-driven remote attestation: Attesting remote system from behavioral aspect," in *2008 The 9th International Conference for Young Computer Scientists*, 2008, pp. 2347–2353.

[25] "TCG trusted attestation protocol (TAP) information model for TPM families 1.2 and 2.0 and DICE family 1.0," https://trustedcomputinggroup.org/resource/tcg-trusted-attestation-protocol-tap-information-model-for-tpm-families-1-2-and-2-0-and-dice-family-1-0/.

[26] A. Seshadri, A. Perrig, L. Van Doorn, and P. Khosla, "SWATT: Software-based attestation for embedded devices," in *IEEE Symposium on Security and Privacy, 2004. Proceedings. 2004*, 2004, pp. 272–282.

[27] X. Jin, P. Putthapipat, D. Pan, N. Pissinou, and S. K. Makki, "Unpredictable software-based attestation solution for node compromise detection in mobile WSN," *IEEE Globecom Workshops*, pp. 2059–2064, 2010.

[28] M. Beltrán, "Identifying, authenticating and authorizing smart objects and end users to cloud services in Internet of Things," *Computers & Security*, vol. 77, pp. 595–611, 2018.

[29] "SMARTmpsoc board," https://soc-e.com/products/smartmpsoc-module-zynq-ultrascale-som-for-time-aware-networking/.

[30] "Tiny AES," https://github.com/kokke/tiny-AES-c.

[31] M. Shaneck, K. Mahadevan, V. Kher, and Y. Kim, "Remote software-based attestation for wireless sensors," in *Proceedings of the Second European Conference on Security and Privacy in Ad-Hoc and Sensor Networks*, 2005, pp. 27–41.

http://doi.org/10.18239/jornadas_2021.34.43

MARISMA-BiDa Pattern: Integrated Risk Analysis for Big Data

David G. Rosado
GSyA Research Group
Univ. de Castilla-La Mancha
ESI. Ciudad Real
david.grosado@uclm.es
ORCID: 0000-0003-4613-5501

Julio Moreno
NTT Data
Madrid
jmorgarc@nttdata.com
ORCID: 0000-0001-9974-1199

Luis E. Sánchez,
GSyA Research Group
Univ. de Castilla-La Mancha
ESI. Ciudad Real
LuisE.Sanchez@uclm.es
ORCID: 0000-0003-0086-1065

Antonio Santos-Olmo
I+D+i Department, Marisma
Shield S.L and Sicaman
Nuevas Tecnologías S.L.
Tomelloso (Ciudad Real)
asolmo@sicaman-nt.com
ORCID: 0000-0002-2349-3894

Manuel A. Serrano
Alarcos Research Group.
Univ. de Castilla-La Mancha
ESI. Ciudad Real
Manuel.Serrano@uclm.es
ORCID: 0000-0003-0962-5659

Eduardo Fernández-Medina
GSyA Research Group
Univ. de Castilla-La Mancha
ESI. Ciudad Real
Eduardo.FdezMedina@uclm.es
ORCID: 0000-0003-2553-9320

Abstract- **Data is one of the most important assets for all types of companies, which have undoubtedly grown their quantity and the ways of exploiting them. Big Data appears in this context as a set of technologies that manage data to obtain information that supports decision-making. These systems were not conceived to be secure, resulting in significant risks that must be controlled. Security risks in Big Data must be analyzed and managed in an appropriate manner to protect the system and secure the information and the data being handled. This work proposes a risk analysis approach for Big Data environments, which is based on a security analysis methodology called MARISMA (Methodology for the Analysis of Risks on Information System), supported by a technological environment in the cloud (eMARISMA tool) already used by numerous clients. Our proposal, called MARISMA-BiDa, is based on the main related standards, such as ISO/IEC 27000, or the NIST Big Data reference architecture or ENISA and CSA recommendations for Big Data.**

***Index Terms*- Big Data, Risk assessment, Risk analysis, Information Security, Security Standards**

Tipo de contribución: *Investigación publicada*

Publicación: *David G. Rosado, Julio Moreno, Luis E. Sánchez, Antonio Santos-Olmo, Manuel A. Serrano, Eduardo Fernández-Medina. MARISMA-BiDa pattern: Integrated risk analysis for big data. Computers & Security, Volume 102, 2021, https://doi.org/10.1016/j.cose.2020.102155.*

I. EXTENDED ABSTRACT

Data has become increasingly important in companies in any area, not only they are fundamental for organizations related to the area of information technologies, but also crucial for industries as varied as health, education, engineering, or governments. The broader use of social networks, multimedia data, and IoT produce an increasing amount of data with an unstructured format, which, together with the rapid production of them complicates their analysis using traditional systems. This set of characteristics are known as the 3 Vs (Volume, Variety, and Velocity) of Big Data. Big Data consists of extensive datasets that require a scalable architecture for efficient storage, manipulation, and analysis. Big Data arises in response to the need to analyze and better understand this data, in order to obtain valuable information for the organization.

Using Big Data not only increases the scale of traditional privacy and security issues but also adds new challenges that must be addressed [1]. These problems stem from the fact that Big Data was not initially conceived as a secure environment, but instead, the security risks to which a system of this type may be subject are very high. Therefore, it is of prime importance to have a series of guides, methodologies, and mechanisms to adequately implement both the Big Data environment and its security. In addition to that, it is also widely considered that any global information security management environment in the company should be focused on risks.

The MARISMA methodology (see Fig.1) was developed to improve the current approaches of risk analysis and management. MARISMA defines a risk analysis and management process aligned with international security standards. On the one hand, it performs the analysis identifying assets and potential threats, assessing the impact and probability of occurrence of the threats, and identifying the most appropriate safeguards to evaluate the risk. On the other hand, in the risk management, the most appropriate controls

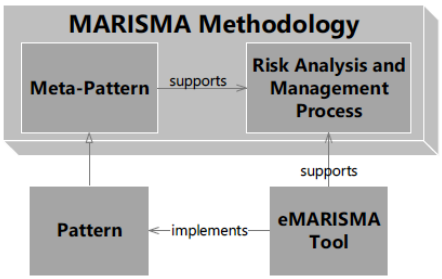


Fig. 1. General Schema of MARISMA Methodology.

are selected, the events are managed, and the evolution of the risks is monitored, as well as the treatment of the risk. MARISMA is focused on the reuse of knowledge for the process of risk analysis and management (concept of meta-pattern), which is a data model representing the key risk analysis components, and which can be specialised to any specific area with particular features, such as IoT, cyber-physical systems, critical systems, etc. through the corresponding risk pattern (see Fig.2). eMARISMA is a platform in the cloud that implements the MARISMA processes, allowing the automation of risk analysis and management process, supporting reuse, and allowing real time updating of risk indicators.

This work presents a specific risk pattern for Big Data environments to deal with risks related Big Data characteristics. To take into account the context, the organization, and the system to be protected, a context-dependent pattern must be defined. For this reason, to conduct a risk analysis and management in a specific environment such as Big Data, the elements of the meta-pattern focused on Big Data must be selected, obtaining a new specific pattern with the characteristics and particularities of this type of environment, which we have called MARISMA-BiDa.

MARISMA-BiDa can be used as a base for instantiating the particularities of any Big Data system for any context (health, finance, government, manufacturing, etc.) given that the types of assets, threats, and risk in these environments are similar from one system to another. It is of prime importance to highlight that these contexts share the same pattern since they all use a Big Data system. The elements of the MARISMA-BiDa pattern are based on the identification and definition of the specific elements for Big Data (based on international recommendations [2,3,4]), which are:

- Domains, control objectives, and controls are usually obtained from international security standards or norms such as ISO/IEC 27000 family. We have considered 14 domains, 35 control objectives and 114 controls that are easily adaptable to the features of Big Data.
- Types of Asset are selected from the taxonomies and reference architectures in Big Data such as NIST, ENISA or CSA. The pattern defines five families: Infrastructure, Data, Individuals and roles, Big Data analytics, and Security and privacy techniques.
- For the dimensions of the MARISMA-BiDa pattern, the different Vs typical of Big Data systems have been considered: Volume, Velocity, Variety, Veracity, Variability and Value.
- Definition of threats in Big Data which have been grouped into 5 types (ENISA recommendations): Unintentional damage; Eavesdropping, interception or hijacking; Nefarious activities or abuse; Legal; and Organizational.

In addition, as it can be seen from the meta-pattern in the Fig. 2, a detailed analysis of the threats, of the assets type that are vulnerable to these threats, and of the dimensions that can be affected by the materialization of this threat on the asset, is necessary. A matrix of relation between these three elements of the risk analysis has been defined.

The proposed pattern is instantiated by a typical Big Data scenario in a healthcare context (defined in [5]) showing how accessible its instantiation is for any given Big Data context. This Big Data system will be used for analysis and decision making in a medical environment as well as protecting, managing, and storing sensitive medical data that can be used in analytical processing for disease detection and decision making. To do this, the first thing to identify is the set of assets that are part of the system (from of the assets defined in the pattern). Secondly, identify which dimensions will be affected. After that, the next step is to analyze the probabilities of occurrence of possible threats (from those defined in the pattern) that may occur in this system, as well as calculate the percentage of degradation that the materialization that the threat may have in the assets associated with those threats.

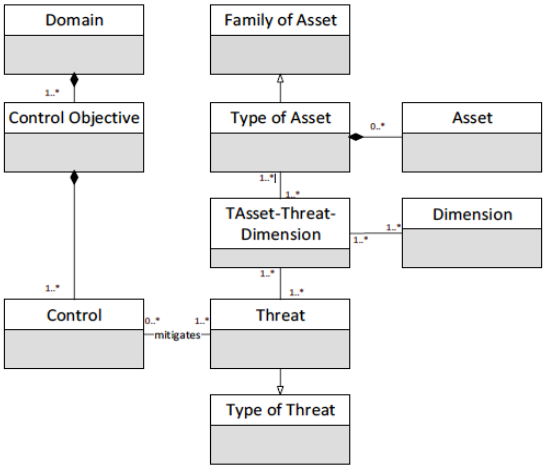


Fig. 2. Meta-pattern in MARISMA.

II. CONCLUSIONS

This work shows how the MARISMA methodology is used, to generate a security management and analysis pattern focused on Big Data aspects, which allows dynamic management of the risk associated with elements of a Big Data environment in a company (MARISMA-BiDa pattern).

The use of the eMARISMA tool helps to automate the process, define context-specific patterns, and perform automatic risk assessment and assists in making risk decisions.

AGRADECIMIENTOS

This work has been funded by the ECLIPSE project (Ministerio de Ciencia, Innovación y Universidades, RTI2018-094283-B-C31), the GENESIS Project (Consejería de Educación, Cultura y Deportes de la JCCM, SBPLY/17/180501/000202) y Fondo Europeo de Desarrollo Regional FEDER. We thank the support of the companies Sicaman Nuevas Tecnologías S.L. (www.sicaman-nt.com) and Marisma Shield S.L. (www.emarisma.com).

REFERENCIAS

[1] Moreno J, Serrano MA, Fernandez-Medina E, Fernandez EB. Towards a security reference architecture for big data. 20th International Workshop on Design, Optimization, Languages and Analytical (DOLAP) 2018.

[2] NIST. NIST Big Data Interoperability Framework: Volume 6, Reference Architecture. NIST Special Publication 1500-6r1. Gaithersburg, MD2015. p. 62. https://doi.org/10.6028/NIST.SP.1500-6

[3] Rekleitis E. Big Data Threat Landscape and Good Practice Guide. European Union Agency For Network And Information Security 2016.

[4] Murthy P, Bharadwaj A, Subrahmanyam P, Roy A, Rajan S. Big Data Taxonomy. Cloud Security Alliance, September; 2014. p. 33.

[5] NIST. Special Publication 800-37 Risk Management Framework for Information Systems and Organizations A System Life Cycle Approach for Security and Privacy 2018b. https://doi.org/10.6028/NIST.SP.800-37r2

http://doi.org/10.18239/jornadas_2021.34.44

Análisis de la Normativa sobre Seguridad de Redes y Sistemas de Información: el Real Decreto 43/2021

M. Robles Carrillo
Network Engineering and Security Group (NESG)
mrobles@ugr.es
Universidad de Granada
ORCID iD: https://orcid.org/0000-0002-6324-4665

Abstract – **La seguridad de redes y sistemas de información ha sido regulada a nivel europeo a través de la Directiva (UE) 2016/1148 y el Reglamento de Ejecución 2018/151. La Directiva es transpuesta en Derecho español mediante el Real Decreto-Ley 12/2018. Esta norma remite a un desarrollo normativo reglamentario posterior que ha tenido lugar con la adopción del Real Decreto 43/2021. El análisis de este conjunto de normas permite identificar el régimen jurídico establecido en cuanto al marco normativo, el ámbito de aplicación, el marco estratégico e institucional, las obligaciones y requisitos, los criterios de calificación y las modalidades de control del cumplimiento de la normativa NIS.**

***Index Terms*-** Seguridad de redes y sistemas, Directiva NIS, Reglamento de Ejecución, Real Decreto-Ley 12/2018, Real Decreto 43/2021

Tipo de contribución: *Investigación en desarrollo*

I. INTRODUCCIÓN

La Unión Europea (UE) adopta en 2016 la Directiva 2016/1148 relativa a las medidas destinadas a garantizar un elevado nivel común de seguridad de las redes y de los sistemas de información en la Unión Europea (en adelante, Directiva NIS) [1]. Esta norma es transpuesta en España mediante el Real Decreto-Ley 12/2018, de 7 de septiembre, de seguridad de redes y sistemas de información (en adelante, RDL-2018) [2], que prevé un desarrollo reglamentario posterior. Finalmente, el 26 de enero de 2021 se aprueba el Real Decreto 43/2021 (en adelante, RD-2021) [3] por el que se desarrolla la norma anterior. Aunque a nivel europeo ya está en marcha la revisión de esta normativa [4], con el RD-2021 se habría culminado en el ámbito interno el proceso de regulación de la seguridad de redes y sistemas de información. El análisis de este complejo entramado normativo pone de manifiesto aciertos, pero también permite advertir serios problemas formales y de fondo.

Para empezar, en el preámbulo del RD-2021 se afirma que esta norma cumple con los principios de buena regulación, necesidad y eficacia, proporcionalidad, transparencia, eficiencia y seguridad jurídica. El principio de buena regulación garantiza la seguridad jurídica, pero depende, en gran medida, del cumplimiento de todos los demás. No está claro que esta normativa responda efectivamente a esos principios.

Un primer problema se advierte cuando, según el RD-2021, se ha cumplido el *principio de transparencia*, "al haber sometido el proyecto de real decreto al trámite de audiencia, definiéndose claramente los objetivos de la iniciativa normativa y su justificación". El problema estriba en que la transparencia desde el punto de vista formal no garantiza la transparencia sustantiva o material, que es la que verdaderamente importa a los destinatarios de la norma y la que garantiza la seguridad jurídica. En este RD-2021 y en su relación con la regulación existente en esta materia, como se comprueba a lo largo de este trabajo, no se aprecia un régimen jurídico claro, transparente y comprensible. Aunque formalmente se hayan respetado los procedimientos, el resultado es un complicado marco normativo que requiere para su clarificación una ardua labor de exégesis jurídica.

Un segundo principio que plantea dudas es el de *proporcionalidad*. Según el preámbulo, este principio se ha respetado al no existir otras medidas menos gravosas para los operadores de servicios esenciales (OSE) y los proveedores de servicios digitales (PSD) en orden a cumplir las obligaciones en materia de seguridad de redes y sistemas. Más allá de que se han sucedido las críticas en distintos foros en cuanto a las cargas que implica esta regulación, en particular, la obligación de designar a un responsable de seguridad, no hay que soslayar el hecho de que la propia complejidad del marco normativo resultante es contraria a la idea de proporcionalidad que quiere predicar.

En tercer lugar, el principio de *eficiencia* se entiende cumplido, textualmente, "dado que no se establecen cargas adicionales a las contempladas en el real decreto-ley que desarrolla". Tampoco en este caso es una apreciación que se pueda compartir. Es cierto que el RDL-2018 es la fuente a partir de la cual se establece la regulación contenida en el RD-2021. Pero esta regulación se extiende considerablemente más allá de lo previsto en el RDL-2018 en cuanto al alcance de muchos de sus contenidos, mientras que otros cambian sensiblemente. No es posible afirmar que no contiene cargas adicionales, salvo que se realice una interpretación muy amplia de las previstas en el RDL-2018 o muy restrictiva de las contempladas en el RD-2021.

El objetivo de esta investigación es analizar el régimen normativo en materia de seguridad de redes y sistemas de información tras la adopción del RD-2021 a la luz de esos principios y desde una perspectiva general. Aunque se advierten ciertos avances, se identifican asimismo algunos retrocesos y carencias en aspectos significativos de una regulación que resulta extraordinariamente compleja en su diseño y organización.

II. MARCO NORMATIVO

El RD-2021 es la norma de desarrollo del RDL-2018 que, a su vez, es la norma de transposición de la Directiva NIS. Hay que incluir, asimismo, en este marco normativo el Reglamento de Ejecución (UE) 2018/151 por el que se establecen normas

para la aplicación de esta Directiva en relación con los PSD (RdE) [5]. Aunque este reglamento es menos conocido, las cuatro normas regulan a diferente nivel los distintos aspectos de una misma materia y han de ser tenidas en cuenta, con la función que les corresponde en cada caso, en el diseño global de este régimen jurídico.

Con carácter general, las normas están sujetas a unas técnicas y principios jurídicos en cuanto a su aplicación e interpretación que, en este caso, no resultan fáciles de aplicar. El principio de competencia o de jerarquía determina la norma aplicable en caso de conflicto entre dos o más preceptos jurídicos. El principio de especialidad normativa determina que lo específico prevalece sobre lo genérico.

En buena lógica, los aspectos más específicos y concretos de la normativa NIS deberían encontrarse regulados en el RD-2021. Pero hay cuestiones que se reglamentan a ese nivel en el RDL-2018 o en el RdE o en ambos, mientras que para otras hay que acudir directamente al RD-2021. A veces, incluso, como se verá en el apartado VI, los mismos criterios o requisitos apuntan a distintos sujetos o finalidades en cada una de las normas. No es fácil interiorizar la filosofía del sistema, ni sus dinámicas, ni sus contenidos cuando, en función de los sujetos o de la materia, el régimen normativo específico se encuentra en una norma, en dos o en las tres y no necesariamente la de carácter reglamentario incluye el contenido normativo más detallado o especializado.

En caso de conflicto o contradicción entre estas normas, hay que aplicar la regla de la primacía -que determina las relaciones entre derecho europeo y nacional- y la regla de la jerarquía dentro del derecho europeo, por una parte, y dentro del derecho nacional, por otra. Al tratarse de normas de Derecho de la UE, tanto la Directiva como el RdE tienen primacía sobre el conjunto del derecho interno español, esto es, tanto sobre el RDL-2018 como sobre el RD-2021. El RdE está subordinado a la Directiva. El RD-2021, a su vez, está subordinado a todas ellas, incluido el RDL-2018 (Tabla 1).

Directiva	Jerarquía	Primacía	Primacía
	Reglamento de Ejecución	Primacía	Primacía
		Real Decreto-Ley	Jerarquía
			Decreto Ley

Tabla 1: Relaciones normativas

Estas reglas generales resultan fáciles de comprender y de aplicar en términos generales y de técnica jurídica, pero no es fácil implementarlas cuando varias normas regulan, unas veces simultánea y otras alternativamente, una misma materia como ocurre en materia de seguridad de redes y sistemas. Por ello, habría resultado más fácil operar solo con dos normas, una europea y una nacional, reformando el RDL-2018 para incluir estas nuevas disposiciones de desarrollo. La adopción de una nueva norma, como el RD-2021, y el modo en que se ha diseñado complican extraordinariamente el régimen jurídico para los destinatarios de esta regulación.

III. ÁMBITO DE APLICACIÓN NORMATIVA

Los destinatarios de la normativa sobre seguridad de redes y sistemas son los OSE y los PSD. La Directiva NIS establece un modelo jurídico diferente para cada uno de ellos y, conforme al mismo, dispone unas normas específicas para los PSD en el RdE. Este reglamento europeo determina los elementos que han de tener en cuenta los PSD para gestionar los riesgos existentes para la seguridad de las redes y sistemas de información, así como los parámetros para determinar si un incidente tiene un impacto significativo. En derecho español, sin embargo, el RDL-2018 pretende definir un régimen uniforme para OSE y PSD que, finalmente, resulta ser más ficticio que real, tanto en sus propias disposiciones como, posteriormente, en las del propio RD-2021. El resultado es que no solo no sirve para homogeneizar las normas de OSE y PSD, sino que, incluso, acaba siendo un modelo de regulación más complejo que el dispuesto a nivel europeo.

El art. 16 del RDL-2018 establece las obligaciones de seguridad de OSE y PSD, optando por una regulación única. No obstante, ya en su apartado 2, marca la diferencia entre ambos porque reenvía a un reglamento posterior para concretar las medidas necesarias a esos efectos respecto de los OSE, mientras que en el apartado 6 determina esas medidas en relación con los PSD. Por su parte, el RD-2021 incluye una extensa y completa regulación de las medidas de seguridad que han adoptar los OSE, pero no hace referencia a los PSD. En términos similares, el art. 19 del RDL-2018 se ocupa de la obligación de notificación, aparentemente, con una regulación única para ambos, pero tampoco en este caso se alcanza ese resultado.

Junto con lo anterior, siguiendo lo previsto en el propio RDL-2018, el RD-2021 establece una diferencia adicional en cuanto a la aplicación subjetiva de la normativa NIS a OSE y PSD. El RD se aplica: 1) A los OSE que estén establecidos en España, esto es, aquellos cuya residencia o domicilio social se encuentren en territorio español, siempre que estos coincidan con el lugar en que esté efectivamente centralizada la gestión administrativa y la dirección de sus negocios o actividades; y 2) A los servicios esenciales que los operadores o residentes de otro Estado ofrezcan a través de un establecimiento permanente situado en España. En cambio, en el caso de los PSD, esta normativa se aplica a: 1) Aquellos que tengan su sede social en España y ésta constituya su establecimiento principal en la Unión Europea; y 2) Aquellos que, no estando establecidos en la Unión Europea, designen en España a su representante en la Unión.

En resumen, el RDL-2018 y el RD-2021 parten del principio de regulación conjunta de las obligaciones de los OSE y los PSD, pero acaban estableciendo distintas medidas y en diferentes textos jurídicos para cada uno de ellos. Hay dos aspectos principales merecedores de crítica.

El primero es precisamente el haber organizado esa operación artificial consistente en pretender unificar el estatuto de los OSE y los PSD en el RDL-2018 que se demuestra fallida o se revierte directamente en el propio RDL-2018 y, sobre todo, en el RD-2021.

El segundo es la complicada regulación resultante de todo ello en la medida en que, bajo esa pretendida apariencia de uniformidad, los PSD tienen que atender al RdE, al RDL-2018 y al RD-2021, mientras que los OSE han de seguir en unos aspectos el RDL-2018 y en otros el RD-2021. En ocasiones, la regulación principal está en el RDL-2018 y en otros se encuentra en el RD-2021.

Una sola norma nacional habría evitado esta compleja arquitectura normativa, que tampoco se habría producido si los criterios previstos para los PSD en el RDL-2018 se hubiesen incluido en el RD-2021, al igual que se ha hecho con los OSE

o si, simplemente, se hubiese mantenido la diferenciación establecida entre unos y otros a nivel europeo.

Posiblemente, al adoptar la decisión de asimilar OSE y PSD en derecho español, no se apreció suficiente o debidamente el impacto y las consecuencias de la diferenciación realizada en las normas europeas tanto en la propia Directiva como, en particular, en el RdE. Hay un dato, a este respecto, que es muy importante. Las directivas crean obligaciones para los Estados que han de adoptar la legislación interna necesaria para cumplir con ellas, siendo esta legislación nacional la que obliga a OSE y PSD. Pero el reglamento europeo es directamente obligatorio, en este caso, para los PSD. Ello significa que las obligaciones de los OSE derivan de la normativa interna -el RDL y el RD-, mientras que las de los PSD resultan tanto de estas normas nacionales como del reglamento de ejecución europeo que también tiene primacía sobre el conjunto del Derecho interno.

IV. MARCO ESTRATÉGICO E INSTITUCIONAL

El marco estratégico e institucional de seguridad de redes y sistemas de información está definido en el Título III del RDL-2018 (arts. 8 a 15) y en el capítulo II del RD-2021 (arts. 3 a 5). En el RDL-2018 se regulan las autoridades competentes, los equipos de respuesta a incidentes de seguridad informática (CSIRT), el punto de contacto único, la cooperación entre autoridades y las obligaciones de confidencialidad. En el RD-2021 se hace referencia, asimismo, a las autoridades competentes, la cooperación y coordinación de los CSIRT y el punto de contacto único. El hecho de que haya cuestiones reguladas en las dos normas obliga a integrar sus disposiciones para conocer su régimen jurídico.

A) Las autoridades competentes

El RDL-2018 establece las autoridades competentes y sus funciones en los artículos 9 y 10, respectivamente, mientras que el RD-2021, en su artículo 3, completa lo dispuesto en el art. 9 del RDL-2018 especificando las autoridades competentes para OSE no considerados operadores críticos y no incluidos en el ámbito de aplicación de la Ley 40/2015, de 1 de octubre, de Régimen Jurídico del Sector Público. Junto con ello, la Disposición Adicional Cuarta del RD-2021 estipula el régimen para OSE y PSD que dependan de proveedores externos a los que les sea de aplicación la disposición adicional novena de la Ley 34/2002, de 11 de julio, de servicios de la sociedad de la información y de comercio electrónico) [6].

El análisis de estas distintas disposiciones permite determinar las autoridades competentes en cada caso que son las siguientes:

1) Para los OSE designados como operadores críticos, con independencia del sector estratégico, la autoridad competente es la Secretaría de Estado de Seguridad, del Ministerio del Interior, a través del Centro Nacional de Protección de Infraestructuras y Ciberseguridad (CNPIC).

2) Para los OSE no designados como operadores críticos, la autoridad competente es la autoridad sectorial correspondiente por razón de la materia, según se determine reglamentariamente.

3) Para los PSD, la autoridad competente es la Secretaría de Estado para el Avance Digital, del Ministerio de Economía y Empresa.

4) Para los OSE y los PSD que no son operadores críticos y se encuentran comprendidos en el ámbito de aplicación de la Ley 40/2015, de 1 de octubre, de Régimen Jurídico del Sector Público, la autoridad competente es el Ministerio de Defensa, a través del Centro Criptológico Nacional.

5) Para los OSE que no son operadores críticos y que no se encuentran comprendidos en el ámbito de aplicación de la Ley 40/2015, las autoridades competentes están determinadas por sectores de actividad pormenorizadamente en el art. 3 del RD-2021.

5) Para los OSE y PSD, cuando dependan de proveedores externos a los que les sea de aplicación la disposición adicional novena de la Ley 34/2002, de 11 de julio, el CERT competente del proveedor externo se corresponderá con: a) El CCN-CERT, del Centro Criptológico Nacional, cuando el proveedor esté incluido en el ámbito subjetivo de aplicación de la Ley 40/2015, de 1 de octubre; o b) El INCIBE-CERT, en el resto de los casos. Esto implica la sujeción del proveedor externo en cuestión -prestadores de servicios de la Sociedad de la Información, registros de nombres de dominio y agentes registradores que estén establecidos en España- a este régimen cuando dependan de ellos tanto OSE como PSD.

Para terminar, los OSE y los PSD que no son operadores críticos y se encuentran comprendidos en el ámbito de aplicación de la Ley 40/2015, de 1 de octubre) [7], están asignados a la misma autoridad, mientras que para los que no están cubiertos por esa norma, hay una relación específica y precisa de las autoridades previstas para los OSE, pero no se hace referencia a los PSD.

El resultado de esta regulación es, como puede comprobarse, una relación extensa y no menos compleja de situaciones distinguiendo entre OSE y PSD en muchos casos y diferenciando entre los supuestos en los que son o no operadores críticos o están o no sometidos a otra legislación. Mayor puede ser la complicación en el caso de OSE y PSD que puedan operar en más de un sector o sometidos a varias normativas.

El modelo de asignación de autoridades competentes es materialmente descentralizado, así como subjetiva y funcionalmente también descentralizado en algunos supuestos. Esta opción implica una mayor especialización, pero también una mayor complejidad y dispersión. Una posibilidad alternativa sería haber optado por un sistema centralizado con una única autoridad competente en materia de seguridad de redes y sistemas de información dentro de la cual podrían existir diferentes secciones o divisiones competentes por materias, sujetos o normas. El modelo sería transparente y accesible porque todos los sujetos tendrían claramente identificada a la única autoridad en la materia y esta se encargaría, internamente, de asignar cada caso a la sección/división especializada dentro de la misma. Ello posibilitaría una mayor y mejor coherencia y coordinación de las actividades y los trabajos en esta materia.

B) Régimen jurídico de los CSIRT

Los arts. 11 y 12 del RDL-2018 regulan los CSIRT de referencia y sus requisitos y funciones, respectivamente, mientras que el art. 4 del RD-202 establece las normas relativas a su cooperación y coordinación que se instrumentarán a través de la nueva Plataforma Nacional de Notificación y Seguimiento de Ciberincidentes regulada en el artículo 11 del RD-2021. Esta plataforma será puesta a disposición de los interesados por el CCN-CERT en colaboración con el INCIBE-CERT y el ESPDEF-CERT del Mando Conjunto del Ciberespacio. No hay referencia en el art. 11 del RD-2021 al

CNPIC que, sin embargo, sí está incluido en el art. 11 del RDL-2018. No es esta la única, ni la principal disfunción identificada en relación con los CSIRT.

El art. 4 del RD-2021 se titula "Cooperación y coordinación de los CSIRT de referencia". El primer apartado, relativo a la Plataforma Nacional, sí responde a esa denominación, pero no ocurre igual con el resto. El apartado 2 se dedica a la definición de los OSE con incidencia en el ámbito de la Defensa o las Fuerzas Armadas y al procedimiento a seguir en caso de incidente. El apartado 3 define los supuestos de especial gravedad que ha de coordinar el CCN-CERT, mientras que el apartado 4 incluye el procedimiento a seguir en estos casos. No es, por tanto, una disposición que responda realmente a su denominación porque, salvo la referencia a la Plataforma, se limita a los incidentes de especial gravedad. Por otra parte, para estos casos, el RDL-2018 establece que la coordinación corresponderá al CCN-CERT y cuando las actividades puedan afectara un operador crítico, los CSIRT de referencia se coordinarán con el Ministerio de Interior a través de la Oficina de Coordinación Cibernética del Centro Nacional de Protección de Infraestructuras y Ciberseguridad (CNPIC). El RD-2021 prevé la información inmediata al Consejo Nacional de Ciberseguridad y tampoco hace referencia al CNPIC en este aspecto concreto.

C) Punto de contacto único

El art. 13 del RDL-2018 dispone que el Consejo de Seguridad Nacional, a través del Departamento de Seguridad Nacional, ejercerá la función de enlace para garantizar la cooperación transfronteriza. El art. 5 del RD-2021 desarrolla, en su apartado 1, esta función. En el apartado 2 prevé otras funciones de enlace que realizará el Consejo de Seguridad Nacional a través de su Comité especializado en materia de ciberseguridad. En el apartado 3 extiende su competencia a los supuestos del art. 18 del RDL-2018, esto es, a los sectores con normativa específica equivalente.

En este punto, el problema estriba en que el RD-2021 atribuye al Comité especializado de ciberseguridad del Consejo de Seguridad Nacional funciones que correspondería ejercer, según el RDL-2018, al Departamento de seguridad Nacional. Ello implica que hay, en realidad, dos interlocutores dentro del punto de contacto único: el Departamento de seguridad Nacional y el Comité especializado en ciberseguridad. Al ser distintos por su composición, naturaleza y funciones, esta previsión desvirtúa en cierta medida el sentido del punto de contacto único.

V. OBLIGACIONES Y REQUISITOS

La Directiva NIS, el RdE, el RDL-2018 y el RD-2021 tienen disposiciones destinadas a establecer las obligaciones y requisitos de seguridad y de notificación de incidentes, pero hay algunas divergencias en cuanto a su alcance y contenido. Las diferencias que implican un aumento o una mejora del nivel de protección en la normativa nacional están permitidas por el art. 3 de la Directiva que recoge el principio de armonización mínima. Conforme a ese principio, los Estados miembros podrán adoptar o mantener disposiciones con el objeto de alcanzar un mayor nivel de seguridad de las redes y sistemas de información. En cambio, las diferencias de otra naturaleza no están autorizadas. La normativa nacional no puede contrariar la europea, ni el RD-2021 puede contrariar el RDL-2018.

Consecuentemente con ello, para determinar obligaciones y requisitos de seguridad, hay que aplicar la Directiva NIS, el RDL-2018 y el RD-2021 en el caso de los OSE y la Directiva NIS, el Reglamento de Ejecución, el RDL-2018 y el RD-2021 en el caso de los PSD. El análisis de esa normativa permite identificar cuatro categorías de obligaciones.

A) Obligación de adoptar medidas para gestionar los riesgos de seguridad

La obligación de adoptar las medidas técnicas y de organización adecuadas y proporcionadas para gestionar los riesgos que afecten a la seguridad de las redes y sistemas de información está recogida en los arts. 14 y 16 de la Directiva para OSE y PSD, respectivamente, pero con apreciables diferencias entre ellos. El art. 14.4 especifica los criterios para determinar la importancia de los efectos de un incidente en relación con los OSE conforme a tres indicadores: el número de usuarios afectados, la duración del incidente y la extensión geográfica de la zona afectada. Por su parte, el art. 16.4 relativo a los PSD incluye, además de esos tres, otros dos indicadores: el grado de perturbación del funcionamiento del servicio; y el alcance del impacto sobre las actividades económicas y sociales. Pero, en este caso, esos criterios se utilizan para determinar si el impacto de un incidente es significativo y no exactamente, como respecto de los OSE, la importancia de sus efectos.

Aunque pueda parecer lo mismo, en términos jurídicos, no es igual determinar la importancia de los efectos de un incidente que calificar un incidente como significativo. La norma europea distingue los supuestos de hecho -importancia de los efectos del incidente frente a calificación del incidente como significativo- y los indicadores usados en cada caso -tres para OSE y cinco para PSD-. Pero es que, además, solo para los PSD, el art 16.1 de la Directiva dispone los criterios a seguir para garantizar que las medidas adoptadas son adecuadas y apropiadas para asegurar el requerido nivel de seguridad, a saber: a) la seguridad de los sistemas e instalaciones; b) la gestión de incidentes; c) la gestión de la continuidad de las actividades; d) la supervisión, auditorías y pruebas; y e) el cumplimiento de las normas internacionales.

Estas diferencias en el régimen jurídico diseñado en la Directiva NIS para OSE y PSD se amplían y consolidan en el RdE. En sus arts. 3 y 4 desarrolla y explica los criterios para la determinación del alcance significativo de un incidente respecto de los PSD. En la medida en que, en su ámbito subjetivo de aplicación, no incluye a los OSE, incluso los tres criterios comunes a PSD y OSE -el número de usuarios afectados, la duración y la extensión geográfica del incidente- pueden ser interpretados y aplicados de modo diferente a cada una de esas categorías de sujetos. Además, este reglamento europeo cuantifica el impacto del incidente conforme a unos criterios, que solo se aplican a los PSD y que son distintos a los establecidos en el RD-2021. Por todo ello, aunque la normativa española pretenda unificar sus regímenes jurídicos, la situación de ambos tipos de sujetos es jurídicamente distinta. Para alcanzar efectivamente ese objetivo, habría sido necesario introducir en la normativa española respecto de los OSE las obligaciones creadas para los PSD en el RdE europeo que, como norma, es obligatorio y directamente aplicable. No ha sido así.

El RDL-2018 establece conjuntamente para OSE y PSD, en su art.16, las obligaciones reguladas para cada uno de ellos, por separado, en los art. 14 y 16 de la Directiva NIS. A pesar de

ello, el propio art. 16 del RDL-2018 ya reconoce diferencias entre ambas categorías de sujetos. El aspecto más relevante es que determina los criterios a seguir para adoptar esas medidas en el caso de los PSD, pero reenvía a la normativa reglamentaria -que será el RD-2021- para hacer lo propio respecto de los OSE.

Por su parte, el art. 6 del RD-2021 reproduce los términos del art. 16 del RDL-2018 para OSE y PSD, pero añade que esas obligaciones se aplicarán "tanto si se trata de redes y sistemas propios, como de proveedores externos". Esto es una novedad más importante de lo que, en principio, podría parecer porque supone una ampliación de la responsabilidad y de las obligaciones de los OSE y los PSD y una complicación para la ejecución de las mismas.

Es una ampliación de obligaciones porque no se trata solo de adoptar las medidas técnicas y organizativas necesarias en su ámbito interno, sino de extender esa obligación hasta el ámbito de sus relaciones con proveedores que no necesariamente han de ser sujetos obligados directamente por esta normativa, pero acaban siéndolo indirectamente o por mediación de las propias obligaciones de OSE y PSD tal y como están definidas en el RD-2021.

Es una complicación porque una cosa es establecer medidas de seguridad internas y otra muy distinta obligar a terceros, que no son destinatarios obligados por la normativa NIS, a asumir esas obligaciones de seguridad. Ni los OSE ni los PSD cuentan, en principio, con la capacidad o el poder coactivo para imponer obligaciones a terceros. Por ello, habría tenido más lógica y resultado más efectivo incluir a esos terceros directamente y con las particularidades que requiriesen dentro del régimen jurídico de NIS que hacerlo indirectamente y a través de OSE y PSD convirtiéndolo en una obligación para estos últimos.

Para terminar, el RD-2021 incorpora tres aspectos especialmente destacables para el cumplimiento de estas obligaciones pero que solo se aplican a los OSE. El primero es la obligación de aprobar *políticas de seguridad de redes y sistemas de información* conforme a los principios de seguridad integral, gestión de riesgos, prevención, respuesta y recuperación, líneas de defensa, reevaluación periódica y segregación de tareas. El art. 6.2 específica, además, los aspectos concretos que, como mínimo, han de contener dichas políticas. El segundo es la *Declaración de Aplicabilidad* de medidas de seguridad que es un documento que ha de suscribir el responsable de seguridad de la información y que se remitirá a la autoridad competente habiendo de ser revisado, al menos, cada tres años (art. 6.4). El tercero es la figura del *responsable de seguridad de la información* que constituye una garantía adicional y fundamental para el cumplimiento de las obligaciones de seguridad. El art. 7 especifica pormenorizadamente tanto sus funciones como los requisitos necesarios par desempeñar esa responsabilidad.

B) Obligación de prevenir y reducir el impacto de los incidentes

Los arts. 14.2 y 16.2 de la Directiva establecen que los Estados velarán por que los OSE y los PSD adopten las medidas necesarias para prevenir y reducir al mínimo los efectos de los incidentes a fin de garantizar la continuidad de sus servicios. El RDL-2018 incluye esa obligación para ambos en el segundo párrafo del art. 16.1. Por su parte, el RD-2021 solo hace mención a esta obligación colateralmente al enumerar las funciones del responsable de seguridad de la información en su art.7.3.a.

C) Obligación de notificar incidentes de seguridad

Los arts. 14.3 y 16.2 de la Directiva establecen que los Estados miembros velarán por que los OSE y los PSD notifiquen sin dilación indebida a la autoridad competente o al CSIRT cualquier incidente que tenga un impacto significativo en la prestación de los servicios. La Directiva requiere que la notificación incluya la información necesaria a efectos de determinar cualquier efecto transfronterizo del incidente. Asimismo, dispone que la notificación no sujetará al notificante a una mayor responsabilidad [8].

Los criterios para determinar el impacto son, como se ha visto con anterioridad, distintos para OSE y PSD. En el caso de los OSE, habrá que atender al número de usuarios afectados, la duración del incidente y la extensión geográfica de la zona afectada. En el caso de los PSD, los parámetros para la definición del impacto del incidente los tres anteriores, los dos siguientes: el grado de perturbación del funcionamiento del servicio y el alcance del impacto sobre las actividades económicas y sociales. En el caso de los PSD, todos esos indicadores están claramente especificados en el RdE.

Una segunda diferencia entre el régimen de ambos se encuentra en el hecho de que solo respecto de los PSD está previsto que la obligación de la notificación únicamente se aplicará cuando tengan acceso a la información necesaria para valorar el impacto de un incidente en función de esos parámetros.

A pesar de esas diferencias, el procedimiento de información pública es el mismo en ambos casos: tras consultar al OSE o al PSD autores de la notificación, la autoridad competente o el CSIRT podrán informar al público sobre determinados incidentes, cuando la concienciación pública sea necesaria para evitar un incidente o gestionar uno que ya se haya producido.

El RDL-2018 dedica un título completo, el Título V, a la notificación de incidentes y regula en la misma disposición, el art. 19, esta obligación para OSE y PSD, pero con algunas mínimas diferencias como la posibilidad abierta a los OSE de notificar sucesos o incidencias que aún no han tenido lugar. Aparte de este caso, los incidentes sujetos a la obligación de notificación son más amplios en el RDL-2018 al especificar que se aplica "tanto si se trata de redes y servicios propios como si lo son de proveedores externos, incluso si éstos son proveedores de servicios digitales sometidos a este real decreto-ley". Esto supone que hay una diferencia a esos efectos entre PSD y OSE porque estos últimos no se mencionan en esa previsión. Pero implica asimismo que, si un PSD no notifica, bien otro PSD o algún OSE con los que tenga relación puede notificar y, con ello, evidenciar el incumplimiento por parte del primero de su obligación. Esa opción también está prevista en el art. 24 del RDL-2018 si se tiene conocimiento de incidentes que afecten a servicios ofrecidos en España por PSD establecidos en otro Estado de la UE.

El RDL-2018 establece los criterios de valoración del impacto para OSE y PSD incluyendo, además de los cinco establecidos en la Directiva para los PSD, los dos siguientes: la importancia de los sistemas afectados o de la información afectada por el incidente para la prestación del servicio esencial y el daño a la reputación. Aparte de estas diferencias, el RDL-2019 establece disposiciones sobre la protección del denunciante, la posibilidad de notificaciones iniciales, intermedias y finales, así como sobre información pública.

Sorprendentemente -o quizás, después de lo visto hasta ahora, ya no tanto-, el RD-2021 no dedica un apartado específico a la notificación como el RDL-2018, sino que incluye la notificación dentro del Capítulo IV sobre gestión de incidentes de seguridad y se ocupa exclusivamente de las obligaciones de los OSE en su art. 9 y de los procedimientos de notificación de los OSE en el art. 10. No resulta fácil entender, si es los que hay, los motivos que justifican este artificial y confuso proceso de unificación y diversificación de regímenes jurídicos de los OSE y los PSD. Si la Directiva NIS dispone un régimen distinto y es eso lo que finalmente se establece en la mayoría de las disposiciones, no había necesidad alguna de operar con la ficción de una equiparación entre ambos.

En este punto, merece una valoración positiva la Plataforma Nacional de Notificación y Seguimiento de Ciberincidentes establecida en el art. 11 del RD.

La Plataforma Nacional de Notificación y Seguimiento de Ciberincidentes es responsabilidad del CCN-CERT en colaboración con el INCIBE-CERT y el ESPDEF-CERT del Mando Conjunto del Ciberespacio. El RD-2021 introduce este importante instrumento en materia de notificación y seguimiento de incidentes con las siguientes atribuciones y caracteres: 1. Permite el intercambio de información y el seguimiento de incidentes por parte de OSE, PSD, autoridades competentes y CSIRT de referencia de manera segura y confiable; 2. Debe asegurar la disponibilidad, autenticidad, integridad y confidencialidad de la información; 3. Garantiza el acceso de las autoridades competentes a toda la información y les permite efectuar en todo momento el necesario seguimiento y control de la situación; 4. Puede utilizarse para cumplir con la obligación de notificación; y 5. Implementa el procedimiento de notificación y gestión de incidentes, estando disponible permanentemente, y disponiendo de las siguientes capacidades: a) Capacidad de gestión de ciberincidentes; b) Capacidad de intercambio de información sobre ciberamenazas; c) Capacidad de análisis de muestras; d) Capacidad de registro y notificación de vulnerabilidades; e) Capacidad de comunicaciones seguras; f) Capacidad de intercambio masivo de datos; y g) Generación de estadísticas e informes agregados.

Una valoración positiva merece, asimismo, la Instrucción nacional de notificación y gestión de ciberincidentes incluida en el Anexo del RD-2021 donde se encuentra una taxonomía pormenorizada de los mismos, los criterios de determinación de su peligrosidad, los indicadores del nivel de impacto y la relación de la información a notificar en cada caso.

D) Obligación de resolver incidentes de seguridad

La normativa española introduce una obligación que no está prevista como tal en la Directiva NIS. El Título V del RDL-2018, titulado "Notificación de incidentes" incluye una disposición que, a pesar de su importancia, podría pasar inadvertida precisamente por esa localización. Se trata del art. 28 del RDL-2018 en virtud del cual los OSE y los PSD "tienen la obligación de resolver los incidentes de seguridad que les afecten, y de solicitar ayuda especializada, incluida la del CSIRT de referencia, cuando no puedan resolver por sí mismos los incidentes". El RD-2021 incluye esta misma obligación, pero de un modo diferente y más lógico.

El Capítulo IV del RD-2021, titulado "Gestión de incidentes de seguridad", establece primero, en su art. 8, esa obligación de gestionar y resolver los incidentes de seguridad y, a continuación, en el art. 9, la obligación de notificación de incidentes.

Conforme al art. 8.1 del RD-2021, los OSE y los PSD "deberán gestionar y resolver los incidentes de seguridad que afecten a las redes y sistemas de información utilizados para la prestación de sus servicios. En el caso de redes y sistemas que no sean propios los operadores deberán tomar las medidas necesarias para garantizar que dichas acciones se lleven a cabo por los proveedores externos". Consecuentemente con ello, la obligación de *resolver* del RDL-2018 se convierte en obligación de *gestionar y resolver* en el RD-2021 y mientras que en el RDL-2018 se limita a los propios incidentes, en el RD-2021 se extiende a las acciones que lleven a cabo sus proveedores externos. Siguiendo el tenor de esta disposición, deben tomar las medidas necesarias para garantizar que sus proveedores externos gestionen y resuelvan sus incidentes de seguridad. Una vez más, se traslada a los OSE y PSD una responsabilidad que no les corresponde y difícilmente pueden gestionar. Ello es así en mayor medida, si cabe, porque en este caso se trata de una obligación no solo de comportamiento, sino de resultado. Jurídicamente, la diferencia es muy importante.

Por otra parte, el apartado 4º del art. 8 establece en este punto otra diferencia entre OSE y PSD porque se refiere exclusivamente a los OSE para estipular que "aplicarán los aspectos pertinentes de la política de gestión de la seguridad de las redes y sistemas de información a la que se refiere el artículo 6, así como las obligaciones específicas que en su caso establezcan las autoridades competentes". Podría pensarse que ello se debe a que en el caso de los PSD no existen las mismas obligaciones y que respecto de ellos se aplicaría, en su lugar, la normativa del RdE. Pero este reglamento se dedica a establecer los elementos de seguridad y los parámetros que han de ser tenidos en cuenta para determinar si el impacto de un incidente es significativo. No se refiere a las modalidades de resolución del incidente.

Para terminar con este punto, hay disfunciones entre ambas normativas que no tienen justificación. El RD-2021 hace referencia a los proveedores externos respecto de la obligación de adoptar medidas para gestionar los riesgos de seguridad y la obligación de gestionar y resolver incidentes. En cambio, el RDL-2018 solo hace referencia a los proveedores externos respecto de la obligación de notificar que, sin embargo, no es mencionada en el RD-2021. Como consecuencia de ello, en realidad, será efectiva respecto de las tres obligaciones. Pero es difícil entender el porqué de esa regulación innecesariamente incoherente y complicada.

VI. CRITERIOS DE CALIFICACIÓN

La Directiva, el RdE, el RDL-2018 y el RD-2021 incluyen en sus distintas disposiciones diversos criterios de calificación en relación con la determinación del efecto perturbador significativo de un incidente, la importancia de sus efectos o su impacto significativo. El análisis de estas normas pone de manifiesto, una vez más, la complejidad de este marco normativo. Más allá del hecho de que la Directiva y el RdE distinguen a esos efectos entre OSE y PSD, la relación de criterios y su diferente articulación muestran un panorama desalentador no solo en términos jurídicos sino, sobre todo, desde una perspectiva práctica y operativa. Sin necesidad de entrar en la tipología específica del RD-2021 a partir de los

parámetros identificados en el anexo, que complica aun más la situación, el resultado puede apreciarse en la Tabla 2.

Norma	Concepto	Criterio
Directiva	Efecto perturbador significativo del incidente (art. 6)	Nº de usuarios Dependencia Repercusión en grado y duración Cuota de mercado Extensión geográfica Importancia de la entidad
	Importancia de los efectos del incidente para OSE (art. 14)	Nº de usuarios Duración del incidente Extensión geográfica
	Impacto significativo de un incidente para PSD (art. 16)	Nº de usuarios Duración Extensión geográfica Grado de perturbación del servicio Alcance del impacto en las actividades
RdE 2018/151	Parámetros a tener en cuenta para determinar el impacto significativo de un incidente para PSD (art. 3)	Nº de usuarios: - nº de personas físicas y jurídicas con las que existe un contrato de prestación de servicios - nº de afectados atendiendo al tráfico previo Duración: plazo desde la perturbación hasta el restablecimiento Extensión geográfica: Estados afectados Grado de perturbación del servicio: atendiendo a la disponibilidad, autenticidad, integridad o confidencialidad Alcance del impacto en las actividades: atendiendo a las relaciones contractuales, usuarios y pérdidas
	Impacto significativo de un incidente para PSD (art. 4)	Servicio indisponible en términos de horas de usuario: 5.000.000 horas Pérdida de autenticidad, integridad o confidencialidad que afecte a más de 100.000 usuarios Riesgo para la seguridad o pérdida de vidas humanas Daños materiales a un usuario de la UE superiores a 1.000.000 Euros
RDL-2018	Importancia de los efectos del incidente (art. 21)	Nº de usuarios Duración Extensión o áreas geográficas Grado de perturbación del servicio Alcance del impacto en actividades Importancia de los sistemas afectados o de la información Daño a la reputación
RD-2021	Parámetros para determinar el nivel de impacto del incidente (Anexo)	Impacto en la seguridad nacional o ciudadana Efectos en la prestación de un servicio esencial o infraestructura crítica Tipología de la información o sistemas afectados Grado de afectación de las instalaciones Posible interrupción en la prestación del servicio normal Tiempo y costes para la recuperación Pérdidas económicas Extensión geográfica Daños reputacionales
	Nivel de impacto del incidente (Anexo)	Crítico Muy alto Alto Medio Bajo Sin impacto
	Nivel de peligrosidad del incidente (Anexo)	Crítico Muy alto Alto Medio Bajo

Tabla 2: Criterios de calificación

La lectura de estas disposiciones pone de manifiesto la complejidad de un modelo basado en distintos conceptos mezclados con distintos criterios que, incluso, cuando pueden ser similares o idénticos en su contenido, han de ser objeto de una aplicación y una interpretación distintas para OSE y para PSD por estar recogidos de diverso modo y en diferentes normas.

VII. EL MODELO DE SUPERVISIÓN

El control del cumplimiento de las obligaciones de seguridad y notificación por parte de OSE y PSD es fundamental para garantizar la consecución de los objetivos de la normativa NIS.

La Directiva NIS establece la necesidad de supervisión de las obligaciones para los OSE en su art. 15 y para los PSD en su art. 17. Además de formalizarlo en diferentes disposiciones, en el caso de los PSD ese control está limitado a los casos en los que fuera necesario, mediante actividades de supervisión a posteriori y cuando se tengan pruebas de que un PSD no cumple los requisitos. Por su parte, el RDL-2018 establece el régimen de supervisión de los OSE en su art. 32 y el de los PSD en el art. 33. En este caso, la diferencia principal entre ambos estriba en que, en el caso de los PSD, la autoridad competente solo inspeccionará el cumplimiento de las obligaciones cuando tenga noticia de algún incumplimiento.

En cambio, el RD-2021 regula en una misma disposición, en su art. 15, el régimen de supervisión de OSE y PSD. En el apartado 1 establece que las autoridades competentes supervisarán el cumplimiento de las obligaciones por parte de OSE y PSD sin mencionar la diferenciación recogida a esos efectos en el RDL y precisando, además, la obligación de colaboración existente en ambos casos. A pesar de ello, en el resto de sus apartados se encuentra la diferencia entre el régimen de los OSE y el de los PSD. Como cabía esperar, solo respecto de los OSE está prevista la verificación del cumplimiento de las funciones del responsable de seguridad de la información, así como de la política de seguridad y de la Declaración de aplicabilidad de medidas de seguridad. Solo a los OSE se les podrá requerir la remisión de un informe de auditoría. Solo respecto a los PSD está prevista la supervisión coordinada con las autoridades competentes de otros Estados miembros.

En definitiva, la Directiva establece un régimen diferenciado para OSE y PVD. El RDL-2018 mantiene esa regulación separada y ese régimen jurídico diferente. Pero el RD-2021 regula conjuntamente ambas categorías de sujetos, sin especificar el régimen especial de los PSD -supervisión a posteriori y en caso de tener noticia de algún incumplimiento- pero estableciendo obligaciones especiales para los OSE.

VIII. CONCLUSIONES

El objetivo de alcanzar un elevado nivel de seguridad de redes y sistemas de información justifica la adopción de la Directiva NIS, el RdE, el RDL-2018 y el RD-2021. No ha sido un proceso fácil, ni los resultados son realmente satisfactorios (Tabla 3).

Concepto	Acierto	Error	Motivo	Solución
Marco normativo		4 normas	Complejo, poco transparente y redundante	1) Una única norma interna 2) Una regulación coherente y homogénea
Ámbito de aplicación		Régimen jurídico de OSE y PSD	Diferenciación europea y pretendida unificación nacional	1) Un régimen diferenciado siguiendo el modelo europeo 2) Un régimen uniforme real
Autoridades competentes		Multiplicidad de autoridades	Un modelo descentralizado material, subjetiva y funcionalmente es complejo, disfuncional y poco transparente.	La creación de un modelo centralizado con una única autoridad
Régimen jurídico de los CSIRT		Falta de coherencia en la regulación	Incoherencias innecesarias en la regulación	Una mejor y más coherente regulación
Punto de contacto único		Asignación de funciones a los órganos	Duplicidad de órganos	Una regulación acorde con la naturaleza del punto único de contacto
Obligaciones y requisitos		Desajustes entre la regulación europea y la nacional y entre las normas nacionales	Una regulación de la normativa inferior compatible con la superior	Una regulación coherente
	Políticas de seguridad		Los OSE deben establecer estas políticas	
	Declaración de aplicabilidad		Es un documento que garantiza la existencia de la política de	

			seguridad y su actualización	
	Responsable de la seguridad de la información		Garantiza el cumplimiento de las políticas y medidas de seguridad	
	Obligación de resolver los incidentes		Es una obligación de resultado	
	Plataforma Nacional de Notificación y Seguimiento de Ciberincidentes		Es un modelo centralizado de notificación	
	Instrucción nacional de notificación y gestión de ciberincidentes		Es un instrumento útil y necesario	
Criterios de calificación		Relación de conceptos y criterios	El uso incoherente de conceptos y criterios	Un modelo uniforme en cuanto a los conceptos y criterios
Modelo de supervisión		Régimen jurídico de OSE y PSD	Regulación diferenciada y conjunta	Una regulación -diferenciada o conjunta- pero racional y homogénea

Tabla 3: Valoración

Con carácter general, el primer problema se encuentra en la regulación desde el punto de vista subjetivo. La Directiva NIS establece normas diferentes para OSE y PSD. El RdE de la Comisión consolida esa opción al aplicarse solo a PSD. En cambio, el RDL-2018 y el RD-2021 parten de la regulación conjunta de las obligaciones y requisitos para OSE y PSD. Esta opción podría considerarse positiva en la medida en que puede homogenizar y equiparar su régimen jurídico. Pero, en realidad, el régimen jurídico establecido del RDL y en el Capítulo III del RD no solo es diferente, sino que resulta complicado y menos transparente de lo que habría de ser en términos de seguridad jurídica.

Un segundo problema general se encuentra en la regulación desde el punto de vista objetivo o normativo. Hay cuestiones reguladas innecesariamente por duplicado en el RDL-2018 y en el RD-2021. Hay cuestiones específicamente reguladas en el primero y no en el segundo y a la inversa. Hay cuestiones reguladas de un modo en el primero y de manera distinta en el segundo y a la inversa. Es un problema de técnica y coherencia normativa que se podría y debería haber evitado.

Un tercer problema de orden general se identifica en el plano funcional. La operatividad del sistema puede verse cuestionada por la complejidad del diseño en aspectos como la determinación de las autoridades competentes o la definición de las obligaciones, entre otros.

El análisis realizado sobre las disposiciones en materia de seguridad de redes y sistemas, en particular, tras la aprobación del RD-2021 no permite suscribir la afirmación de que se trata de una normativa respetuosa de los principios de transparencia, proporcionalidad y eficiencia salvo desde una perspectiva formal. Material o substancialmente, son muchos los aspectos dentro de esta regulación susceptibles de mejora lo que impide, lamentablemente, reconocer que se ajuste a los principios de una buena regulación.

AGRADECIMIENTOS

Este trabajo ha sido financiado parcialmente por el Gobierno de España, con fondos FEDER, a través del Proyecto TIN2017-83494-R.

REFERENCIAS

[1] *DOUE*, L 194, de 19 de Julio de 2016, p. 1. Puede verse sobre esta norma M. Robles Carrillo, "Seguridad de redes y sistemas de información en la Unión Europea: ¿Un enfoque integral?", *Revista de Derecho Comunitario Europeo*, vol. 60, 2018, pp. 563-600. Sobre su relación con la normativa PIC, puede verse D. Sánchez Cabello, A.L. Sandoval Orozco y L.J. García Villalba, "El efecto de la transposición de la Directiva NIS en el sector estratégico TIC de la ley 8/2011", *Actas de las V Jornadas Nacionales de Investigación en Ciberseguridad*, Cáceres, 2019, pp. 300-301.

[2] Real Decreto-Ley 12/2018, de 7 de septiembre, de seguridad de redes y sistemas de información, *BOE* nº 218, de 8 de septiembre de 2018, p. 87675. Puede verse sobre esta norma: M. Robles Carrillo, "El proceso de transposición de la Directiva sobre seguridad de redes y sistemas de información en el derecho español", *Instituto Español de Estudios Estratégicos*, nº 78/2018, 2018, pp. 1-22; M. Robles Carrillo, "Seguridad de redes y sistemas de información: de la Directiva 2016/1148 al Real Decreto-Ley 12/2018". *Actas de las V Jornadas Nacionales de Investigación en Ciberseguridad*, Cáceres, 2019, pp. 167-169.

[3] Real Decreto-Ley 43/2012, de 26 de enero, por el que se desarrolla el Real Decreto-ley 12/2018, de 7 de septiembre, de seguridad de redes y sistemas de información, *BOE* nº 24, de 28 de enero de 2021, p. 8187.

[4] European Commission. Proposal for a Directive of the European Parliament and of the Council on measures for a high common level of cybersecurity across the Union, repealing Directive (EU) 2016/1148. COM (2020) 823, final, Brussels, 16.12.2020.

[5] Reglamento de Ejecución (UE) 2018/151 de la Comisión de 30 de enero de 2018 por el que se establecen normas para la aplicación de la Directiva (UE) 2016/1148 del Parlamento Europeo y del Consejo en lo que respecta a la especificación de los elementos que han de tener en cuenta los proveedores de servicios digitales para gestionar los riesgos existentes para la seguridad de las redes y sistemas de información, así como de los parámetros para determinar si un incidente tiene un impacto significativo (*DOUE*, L 26, de 31 de enero de 2018, p. 48). Puede verse sobre esta norma, M. Robles Carrillo y P. García Teodoro, "Medidas de Aplicación de la Directiva NIS a Proveedores de Servicios Digitales: Alcance y Limitaciones", *Actas de las IV Jornadas Nacionales de Investigación en Ciberseguridad*, Donostia-San Sebastián, 2018, pp. 151-158

[6] BOE, nº 166, de 12 de julio de 2002. Referencia: BOE-A-2002-13758.

[7] BOE, nº 236, de 2 de octubre de 2015. Referencia: BOE-A-2015-10566.

[8] Puede verse: CCN, *Guía de Seguridad de las TIC CCN-STIC 817,* Junio 2018.

Sesión de Investigación B2: Gobierno y gestión de riesgos
http://doi.org/10.18239/jornadas_2021.34.45

Automatic Verification and Diagnosis of Security Risk Assessments in Business Process Models

Ángel Jesús Varela-Vaca
IDEA Research Group - Universidad de Sevilla
Orcid:0000-0001-9953-6005
ajvarela@us.es

Rafael M. Gasca
IDEA Research Group - Universidad de Sevilla
Orcid:0000-0003-2348-7424
gasca@us.es

Luisa Parody
IDEA Research Group - Universidad Loyola Andalucía
Orcid:0000-0001-6096-8564
mlparody@uloyola.es

María Teresa Gómez-López
IDEA Research Group - Universidad de Sevilla
Orcid:0000-0002-3562-875X
maytegomez@us.es

***Abstract*—Organisations execute daily activities to meet their objectives. The performance of these activities can be fundamental for achieving a business objective, but they also imply the assumption of certain security risks that might go against a company's security policies. A risk may be defined as the effects of uncertainty on the achievement of the goals of a company, some of which can be associated with security aspects (e.g., data corruption or data leakage). The execution of the activities can be choreographed using business processes models, in which the risk of the entire business process model derives from a combination of the single activity risks (executed in an isolated manner). In this paper, a risk assessment method is proposed to enable the analysis and evaluation of a set of activities combined in a business process model to ascertain whether the model conforms to the security-risk objectives. To achieve this objective, we use a business process extension with security-risk information to (1) define an algorithm to verify the level of risk of process models; (2) design an approach to diagnose the risk of the activities that fail to conform to the level of risk in the objectives; and (3) the implementation of a tool that supports the described proposal in automatic way. In addition, a real case study is presented, and a set of scalability benchmarks of performance analysis are carried out in order to check the usefulness and suitability of automation of the algorithms.**

***Index Terms*—Business Process Management, Business Process Model, Security-Risk Assessment, Model-based Diagnosis, Constraint Programming**

Tipo de contribución: *Investigación ya publicada - IEEE Access journal, vol. 7, pp. 26448-26465, 2019. DOI: 10.1109/ACCESS.2019.2901408. JCR Impact Factor (2019): 3.745.*

Summary of the contribution

Organisations carry out several activities to meet their objectives. The execution of each of these activities can imply tackling certain security risks. In fact, the activities are not executed in an isolated manner; they can be choreographed by using business processes models formed of a set of activities [1] whose execution can imply tackling more complex security risks. Therefore, the analysis of the business processes [2] is crucial to understanding the impact of the possible risks. In this work, risks refer to IT security risks. A risk is defined as the effects of uncertainty on the achievement of the goals (e.g.; data corruption or unauthorised access). The decision to execute a business process in a company to achieve its objective implies the acceptance of the derived risk (or even operational risks). This derived risk is a combination of single risks associated with the activities that conform to the process. The risk assessment of business process models is crucial to detecting potential security risks. For instance, business process compromise attacks [3] are used in an attempt to understand a business process behaviour with the aim of manipulating it and generating specific profits for attackers. After understanding a process behaviour, an attacker might deploy malware within certain tasks that are executed unintentionally by an employee. This malware can allow the attacker the unauthorised access to the systems, such as access to confidential information. Therefore, the organisation must measure and assess the isolated risks and combined activity risks that can cause this type of threat in their business process models (e.g.; data leakage caused by staff). The main objective should be determine which tasks might potentially be affected by these threats.

Our approach focuses on combining together business process management (BPM) and security-risk descriptions. More concretely, determining how to obtain the level of risk for the entire process and how to identify the risk responsible for a nonconformity are paramount to this relationship. To achieve it, the proposals of the paper include (1) a framework for a risk-aware design and the development of business process models; (2) a verification algorithm that checks whether every trace of the annotated process model satisfies the risk objectives determined by the organisation; (3) an algorithm to diagnose the risk responsible in the case of a nonconformity; and (4) the integration of the previous proposals into an implemented tool that supports graphical design, verification and diagnosis.

Risk-aware Business Process Models

The standards ISO/IEC 27005:20018 [4] and UNE 71504:2008 [5] point out business processes and information as relevant assets to be measured in organisations. Therefore, a formalisation of the risk elements according to process models is included. These elements are supported as a BPMN 2.0 extension of risk information. We defined a lightweight extension of the BPMN 2.0 meta-model [6]. This extension derives from two main models UML Profile for Modelling Quality of Service and Fault Tolerance (hereinafter *QFTP*) [7]

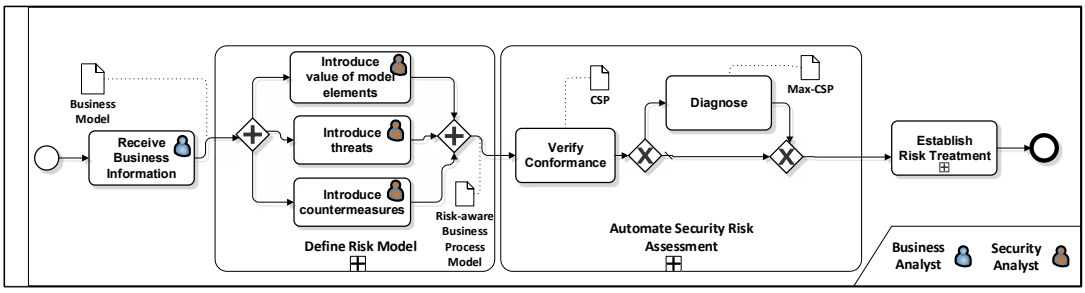


Fig. 1. Process for verification and diagnosis of security-risks.

and Business Motivation Model (*BMM*) [8]. *QFTP* provides a set of generic concepts to develop risk assessment capabilities within IT systems. In contrast, *BMM* provides a model for defining and developing business plans. Business plans are carried out in a final stage by business processes. Thus, *BMM* enables both the identification of factors and relations to define business plans and determine how to achieve and assess these plans.

Risk Assessment of Business Process Models

We introduced [9] how the partial risk of the activities within a process model can be combined in a formulation to carry out the estimation of risk of an entire business process model. This formulation is based on a set of patterns according to the control-flows included in the process, which are used afterwards to verify the conformance of the process. In our approach, independence on the activities risk is assumed.

Automatic Verification of Conformance and Diagnosis by using Constraint Programming

To automate both the verification of conformance and the diagnosis, we propose a two-step subprocess called *Automate Security Risk Assessment* (see Figure 1). First, the *Verify Conformance* activity runs the risk estimation of the business process model and verifies its conformance with regard to the *acceptable risks* that are included as a part of the *CSP*. This verification task is responsible for the calculation of the truth values of the verification. The truth values of verification risk conformance problem *VRC* are obtained by means of a constraint satisfaction problem (CSP) (cf. *CSP* in the figure) created automatically by the transformation of the risk-aware business process model into a *CSP* model. The resolution of the *CSP* obtains those truth values. If any risk formula is nonconforming, the fault diagnosis is executed (cf. *Diagnose* in the figure) by creating and solving a *Max-CSP* to identify the activities whose risks are directed toward the VRC.

Tooling

We developed a tool (called OPBUS-Risk) [10], an Eclipse plug-in based on model-driven architecture technologies that integrates (1) a business process modeller that supports the specification of the extension presented; (2) a transformation engine that enables business process models extended with risk information to be translated into CSP and COP models; (3) a mechanism to support various constraint solvers; and (4) the automation of the verification and diagnosis processes through a set of algorithms that create the CSPs and Max-CSP and solve them.

Conclusions

In this paper, the problem of automatic security risk management in the current BPMS is addresses. First, a formalisation of the risk elements according to process models is included. These elements are supported as a BPMN 2.0 extension of risk information that is analysed to determine non conformance regarding risk goals. In addition, a diagnosis of the risk associated with the activity responsible for the non-conformance is also carried out. To this end, the proposal applies mechanisms based on the model-based diagnosis in which activities are in non-conformance with regard to the acceptable level of risk. The automation of diagnosis is carried out using artificial intelligence techniques based on constraint programming. The proposal is supported by the implementation of a plug-in that enables the graphical specification of the extension and the automation of the verification and diagnosis process. To the best of our knowledge, this is the first published work that addresses the risk-aware design of business processes with automatic techniques.

References

[1] M. Weske, *Business Process Management: Concepts, Languages, Architectures*. Springer, 2007.

[2] S. Sakr, Z. Maamar, A. Awad, B. Benatallah, and W. M. P. van der Aalst, "Business process analytics and big data systems: A roadmap to bridge the gap," *IEEE Access*, vol. 6, pp. 77 308–77 320, 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2881759

[3] Trend Micro. (2017) Business Process Compromise (BPC). [Online]. Available: https://www.trendmicro.com/vinfo/us/security/definition/business-process-compromise

[4] ISO, "ISO/IEC 27005:2018 Information technology – Security techniques – Information security risk management," *Online: https://www.iso.org/standard/75281.html*, 2018.

[5] AENOR, "UNE 71504:2008 - Metodología de análisis y gestión de riesgos para los sistemas de información." *Online: https://www.aenor.es*, 2008.

[6] A. Varela-Vaca, R. Gasca, and A. Jimenez-Ramirez, "A model-driven engineering approach with diagnosis of non-conformance of security objectives in business process models," in *RCIS, 2011 Fifth International Conference on*, may 2011, pp. 1 –6.

[7] Object Management Group (OMG). (2009) UML Profile for Modeling QoS and Fault Tolerance Characteristics and Mechanisms. [Online]. Available: http://www.omg.org/spec/QFTP/1.1

[8] OMG. (2007) Business motivation model (BMM) 1.1. [Online]. Available: http://www.omg.org/spec/BMM/1.1

[9] A. J. Varela-Vaca, L. Parody, R. M. Gasca, and M. T. Gómez-López, "Automatic verification and diagnosis of security risk assessments in business process models," *IEEE Access*, vol. 7, pp. 26 448–26 465, 2019.

[10] A. J. Varela-Vaca. (2018) OPBUS tools. [Online]. Available: http://www.idea.us.es/portfolio-item/opbus-tool/

Sesión de Investigación B2: Gobierno y gestión de riesgos
http://doi.org/10.18239/jornadas_2021.34.46

Enforcing cloud security controls

Javier Jerónimo Suárez
Santander Global Tech
Av. De Cantabria S/N. 28660 Boadilla del Monte (Madrid)
javier.jeronimo@gruposantander.com
https://orcid.org/0000-0001-9701-2662

Jorge López Hernández-Ardieta
Santander Global Tech
Av. De Cantabria S/N. 28660 Boadilla del Monte (Madrid)
jorge.lopez@gruposantander.com

Abstract- **As cloud adoption increases, so does the threats. Cloud security has become a key pillar to deliver critical business applications. In this incipient research we explore what technologies we count on to shift-left the security in cloud applications, helping to enforce security controls in a preventive manner and hence avoid deploying exploitable cloud resources. We classify and compare these technologies according to a proposed taxonomy, and reason about the limitations found. We identify what features a security control-enforcing technology should exhibit, and outline future research direction.**

***Index Terms*- cloud security, guardrails, prevention**

Tipo de contribución: *Investigación en desarrollo*

I. Introduction and Motivation

Cloud adoption has increased at a steady pace over the past years. Infrastructure simplification, cost optimisation, the ability to scale transparently and dynamically, and the agility for the business are just a few of the benefits that cloud computing offers. However, recent events show that cloud also brings the attention of cyber criminals, either to abuse cloud applications and use them as a source of further attacks (e.g. phishing, cloud malware delivery), or simply to compromise the data and services hosted in those applications. A research report by Netskope shows that 61% of all malware is delivered via a cloud app, up from 48% year-over-year [1]. Data breaches in cloud are also on the rise: 70% of companies that host data in public cloud have suffered a breach over a one year period [2]. Some recent high-profile data breaches prove that misconfiguration of cloud applications or poor access control policies can lead to high-impact incidents (CapitalOne [3], relevant breaches in other AWS customers [4]). Delivering secure cloud applications has therefore become an essential objective for businesses to succeed in the digital economy [5].

Efforts around system and software security assurance have produced important contributions over the past years. Today, we count on mature processes and tools that help DevOps teams to shift-left the security and deliver more secure applications [6]. Most of the contributions, however, focus on improving and verifying the security of the application code, whereas verifying the security of the overall cloud architecture is still an error-prone manual activity that leads to exposing vulnerabilities. We are particularly interested in ways to enforce the security of the cloud resources (architectural elements) that the application is built upon, considering applicable security policies (e.g. corporate policies). Enforcing means preventing any inconsistent or non-compliant cloud resource from being deployed in production, hence avoiding deployment of exploitable resources. We believe that enforcing the security at all layers, system-wise, from the cloud infrastructure up to the application layer is essential to meet the abovementioned objective.

After reviewing the bibliography and other published papers on cloud security controls we found either no substantial contributions or others like [14] dating 10 years ago, something we believe is too long for the speed of evolution of cloud computing.

In this paper, we first review the options that we count on to help deliver a secure cloud infrastructure by enforcing the security controls that need to be in place. As this layer has traditionally been less observed by security assurance practices, less mature technologies and practices exist. We introduce a taxonomy where we classify and compare these options according to five parameters, namely type, cloud resource assessed, context, language and expressivity. Then, we explore the limitations found that prevent a granular and system-wide security controls enforcement. Finally, we outline some promising ideas to overcome these limitations and where we will devote our future efforts.

II. Cloud Security Controls

A cloud workload is essentially the set of IT assets running in cloud and that support business processes. These assets typically include business applications, data, virtual machines, and managed services (e.g. managed database, managed key store, managed data analytics framework). Depending on the cloud service model, the workload may be implemented mostly using virtual machines (Infrastructure-as-a-Service, IaaS) as the core infrastructure element, may leverage managed services (Platform-as-a-Service, PaaS), or use a hybrid approach (IaaS/PaaS). Software-as-a-Service (SaaS) models are not part the scope of this paper.

Cloud infrastructure-wise, our interest resides in being able to enforce the required security controls (countermeasures implemented through security capabilities and/or secure configurations) for any virtual machine and PaaS service part of the workload. We assume that the business application has gone through a mature security assurance process, but we must also assume some level of divergence between the intended software security architecture requirements and its actual implementation. This is often called architecture erosion or architectural drifts [7], and is the consequence of the refinement steps from high-level security reference

Table I
TECHNOLOGIES TO ENFORCE SECURITY CONTROLS FOR CLOUD RESOURCES

ENVIRONMENT	TECHNOLOGY	TYPE
Terraform Enterprise	Sentinel Policies [8]	Prevention, Detection (1)
Azure	Azure Policy [9]	Prevention, Detection (2)
Amazon Web Services (AWS)	Config Rules [10]	Detection (3)
	Service Control Policies (SCP) [11]	Prevention
Google Cloud Platform (GCP)	Organisation Policies [12]	Prevention
Multi-cloud	CSPM	Detection (4)

(1) Before actual deployment, out of the CSP.

(2) During the deployment inside the *cloud resource manager*.

(3) After actual deployment (event-driven), inside the CSP.

(4) After actual deployment (polling), outside of the CSP.

architectures to more detailed designs and implementations themselves. The goal of any SecDevOps team is to detect and prevent a wrong implementation from being deployed to production systems.

Due to the shared responsibility model in public cloud, the ability to deepen into the PaaS configuration and actual implementation is limited. In short, in an IaaS model the organisation must take care of the (guest) operating system security and the rest of the stack up to the application, including users' identities and access management (IAM), and protecting the data. On the other hand, in PaaS, the scope is restricted to (part of) the configuration of the PaaS service (e.g. enable encryption at rest in the managed SQL database), the IAM as well as protecting the data as required.

Assuming this shared responsibility model, security controls should be assessed and enforced by the cloud customer organisation as part of its workload security assurance processes. Other controls that fall under the cloud service provider's (CSP) responsibility cannot be overseen by the cloud customer.

To illustrate this situation, *Figure 1* shows a security control that has to be enforced through well-configured security attributes on various cloud resources. However, one of the security attributes in a cloud resource is on the CSP side of the shared responsibility model. If the design and/or implementation of the CSP for that kind of resource does not fulfil such *security attribute 2*, there is nothing the cloud customer organisation can do, and the overall compliance of the security control will not be fulfilled.

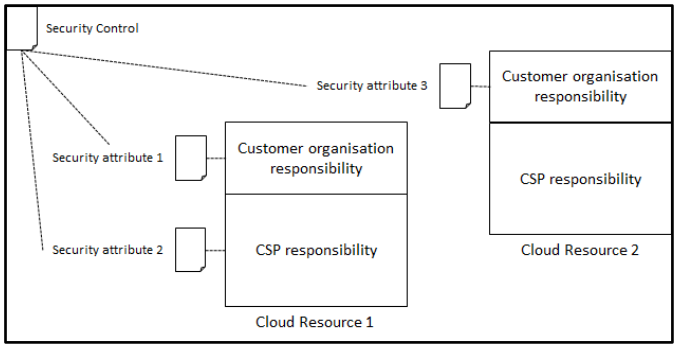


Figure 1: Security Control assessing different attributes of several cloud resources.

For the security controls that fall under the cloud customer responsibility, there is a number of options available. *Table 1* summarises the main technologies that can be used to enforce security controls, for three main cloud providers and a well-known *Infrastructure-as-Code* framework. These technologies automatically verify and enforce security controls for cloud resources at scale. They work either in a preventive mode, meaning that they can prevent a non-compliant cloud resource from being deployed, or in detection mode, so they detect resources already deployed and correct potential misconfigurations or gaps.

Preventive technologies are usually referred to as guardrails as they help DevOps teams from misconfiguring security attributes of cloud resources. Some of these guardrails define properties of cloud resources that must be fulfilled and enforce the desired values on them. For example, *Azure Policy* are JSON rules coded by the organisation and that are assessed and enforced via the *Azure Resource Manager* (ARM). Every time a cloud resource is to be [re-]deployed, ARM requests the *Azure Policies* engine to determine if the resource configuration is compliant, not deploying it otherwise. In AWS, this is implemented through *Service Control Policies* (SCPs), whereas in GCP the technologies include *Organization Policies*. In a step before in the *SecDevOps* pipeline, *Sentinel* policies in *Terraform* can check whether *Terraform* modules violate any security property.

Detective technologies, on the contrary, are able to perform continuous security audit of running cloud environments, checking whether deployed resources violate any security control that should be in place. This capability is built-in in every cloud provider, and given the fact that all of them expose rich management APIs, a plethora of vendors already offer products, known as *Cloud Security Posture Management* (CSPM) tools, to audit security compliance of cloud resources. The downside of CSPM is that they detect a posteriori, opening a window of opportunity for exploitation.

Security control frameworks implemented by cloud providers allow in most cases to implement both detective and preventive controls, and they can be grouped in two categories depending on the main resource type they assess, as depicted in *Table 2*.

An example of the first design is *Azure Cloud* (*Figure 3*). It has *Azure Policy* as an engine embedded in the *cloud resource manager* (i.e. ARM), so that every request to manage a cloud resource is processed by the ARM and assessed by the *Azure Policy* engine. This allows *Azure Policy* to provide the

Table 2
POLICY ENGINE TYPE

TECHNOLOGY	ENGINE TYPE
Terraform Sentinel	Cloud resource centric (1)
Azure Policy	Cloud resource centric
AWS SCPs	IAM principal centric
AWS Config Rules	Cloud resource centric
GCP Organization Policies	Cloud resource centric

1: All resources in a deployment definition

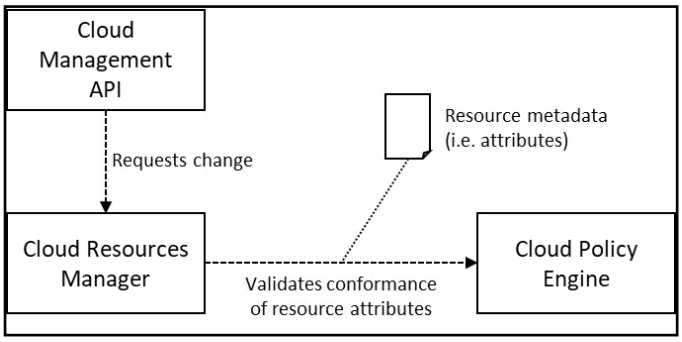


Figure 3: Azure Cloud *policy engine* is resource centric.

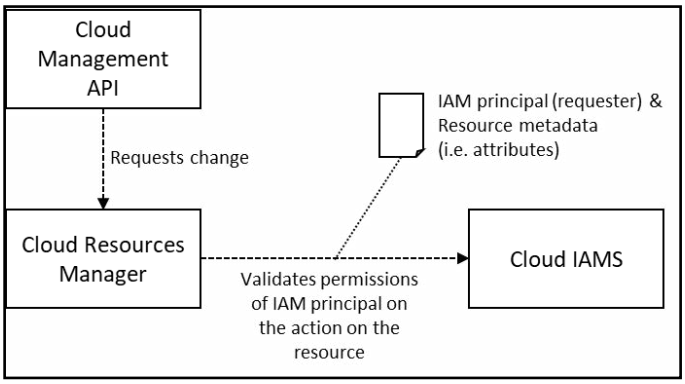


Figure 2: AWS Cloud *policy engine* is IAM principal centric.

capability to the cloud customer to implement detective, preventive and corrective controls using the same technology, as it is embedded in the *cloud resource manager*.

An example of the second design is *AWS Cloud* (Figure 2). Validation of change requests to AWS resources is assessed mainly depending on the *IAM principal* requesting the change. This implies several limitations. For example, if external principals are allowed to manage cloud resources (e.g. principals from accounts in a different organisation), then they can bypass *Service Control Policies* defined by the organisation hosting the cloud resource.

In addition to this internal design of the *policy engine* (resource centric / IAM principal centric) there is another fundamental design decision that will have an impact on the options available to the cloud customer when implementing security guardrails. It is the context information provided by the *cloud resource manager* to the *policy engine*, in addition to the key resource data they already share. *Table* 3 lists this information that these two components of the CSP share.

Table 3
CLOUD CONTEXT AVAILABLE IN POLICY ENGINE*

TECHNOLOGY	MAIN RESOURCE ASSESSED	CONTEXT AVAILABLE
Terraform Sentinel	Cloud resource	Several resources
Azure Policy	Cloud resource	Parent (1)
AWS SCPs	IAM principal	Resource (1)
AWS Config Rules	Cloud resource	Any resource (2)
GCP Org. Policies	Cloud resource	None

(1) Limited scope.
(2) Via API

* In addition to the key resource assessed by the engine, which is:
- IAM principal requesting the change, in AWS SCPs.
- Resource being [re-]deployed, in all others.

III. LIMITATIONS

Security guardrails help organisations to enforce a vast number of security controls and validate many properties of cloud resources, but there are some limitations today that impede a system-wise approach.

First, there are security controls that can be enforced by just looking at properties and configuration of a single cloud resource, e.g. enforcing TLS 1.2 in any HTTPS interface. However, other security controls need a system-wide reasoning, performing complex and broad security checks covering several cloud resources, e.g. checks on routing in a virtual network and rules configured in a network firewall to assess ingress network connectivity from Internet to the workload. Most technologies described in *Table 1* work very well for the former, but dramatically fail for the latter.

Second, the expressivity of the language used to implement security controls in the different CSPs may bring further limitations [13] – See *Table 4*. General purpose languages with specific SDKs (e.g. *AWS Config*) allow the implementation of complex conditions and checks on groups of resources. It allows also to use well-known software development practices and tooling like encapsulation or composition, leading to software components that are easier to maintain and test. On the contrary, document formats with specific schemas like JSON in *Azure* and *Google* limit the customer to implement only basic Boolean conditions on attributes of resources.

Table 4
POLICY ENGINE LANGUAGE TYPE

TECHNOLOGY	LANGUAGE
Terraform Sentinel	DSL
Azure Policy	JSON (1) *
AWS SCPs	JSON (1)
AWS Config Rules	Programming language SDK
GCP Org. Policies	JSON (2)

*: Azure Kubernetes Service policies (i.e. OPA REGO rules) are out of scope in this paper, as well as other container PaaS technologies.
1: Basic Boolean logic and resource attributes.
2: Resource attributes.

In between we have the approach followed by Terraform Sentinel, which uses a *Domain Specific Language* (DSL) that provides some of the flexibility of a general purpose programming language, but constrained by the limitations imposed by the designer of the framework/DSL. Generally, the possibilities of implementing custom logic using a DSL are quite similar to those using a general purpose programming language, but with less degree of freedom, which can be beneficial (i.e. simpler implementation). The drawback is that a custom toolset is required when developing using a DSL.

Table 5 shows a summary of the different dimensions we can use to compare the technologies cloud customer organisations can use to implement security controls.

Table 5
CLOUD POLICY ENGINE COMPARISON

TECHNOLOGY	CONTROL TYPE	ELEMENT ASSESSED	CONTEXT AVAILABLE	LANGUAGE	EXPRESSIVITY
Terraform Sentinel	P*, D	Cloud resource	Several resources	DSL	Medium
Azure Policy	P, D	Cloud resource	Parent resource (limited)	Boolean logic in JSON doc.	Low
AWS SCPs	P	IAM Principal (1)	Cloud resource (limited)	Boolean logic in JSON doc.	Low
AWS Config Rules	D	Cloud resource	Any resource through SDK	Programming language SDK	High
GCP Org. Policies	P, D	Cloud resource	None	Boolean logic in JSON doc.	Low

Control types: P *preventative*; D *detective*
*: deployment is prevented by the tool, but it can be still run from outside of Terraform controlled environment (i.e. out of the ALM system).
(1): IAM principal requesting the change.

IV. DIRECTION AND FUTURE WORK

Our aim when enforcing security controls is not only to assess specific security attributes of cloud resources to deploy, but also the relations they have with others in its context. This requires a system-wide approach rather than only leveraging the single-resource analysis performed by the *cloud resource manager*.

In our opinion, to unleash the potential of enforcing security controls in cloud workloads we need to have technologies that meet the next features:

1. **Preventive** assessment of cloud resources, built-in in the *cloud resource manager*, allowing to prevent (i.e. deny) non-compliant deployments.
2. **High expressivity** through a well-defined, rich DSL, or better, using a general purpose programming language with specific SDKs.
3. **Context** information available in addition to the resource being assessed, to allow for system-wide security controls checks. E.g. information on other cloud resources in the same environment of the cloud resource being [re]-deployed.

Future work will deepen into these properties and the design constraints that CSPs may have imposed to their *cloud resource manager* and *policy engine* that can explain their current limitations, such as lowering the latency on requests to the *cloud resource manager* because of the processing performed inline in the *policy engine*. As described in this paper, initial analysis suggests that limitations found preclude cloud customers from required level of control. We are working closely with main cloud providers to devise joint areas of collaboration with this regard, and that could be beneficial for the wider cloud community. Other lines of research include standardising APIs, which can help both to have better preventive capabilities as well as to implement continuous auditing of the cloud security posture.

V. CONCLUSIONS

Enforcing security controls in a preventive manner is paramount to significantly reduce the risk of deploying vulnerable cloud applications. This can be achieved today by using technology that cloud providers offer. However, technologies available bring important limitations (lack of a rich expressive language and the ability to contextually assess multiple resources together) that impede a system-wide and thorough enforcement.

Azure Policy allows to implement both preventive and detective controls, but expressivity is low due to the language used (JSON documents). *AWS Config* is the best possible API the customer organisation may use to implement detective controls (i.e. general purpose programming language with specific SDK), but it cannot implement preventive controls. *AWS SCPs* implement preventive controls but they are *IAM principal* centric and as such have a limited scope and also the same limitation as Azure Policy has, due to the same language used (JSON documents).

We expect that our current collaboration with main public cloud providers will help to research and improve current technology, and help other organisations to securely navigate their public cloud adoption.

REFERENCES

[1] "Cloud and Threat Report". Netskope. February 2021 Edition.

[2] "The state of cloud security 2020". Sophos. July 2020.

[3] https://www.capitalone.com/digital/facts2019/

[4] "The Capital One/AWS Breach". Deutsche Bank Research. July 2019.

[5] "Top Threats to Cloud Computing: Egregious Eleven". Cloud Security Alliance. 2019.

[6] R. A. Khan, S. U. Khan, M. Ilyas, M. Y. Idris: "The State of the Art on Secure Software Engineering: A Systematic Mapping Study", in Evaluation and Assessment in Software Engineering, pp. 487–492. April 2020.

[7] Sebastian Herold, Andreas Rausch: "A Rule-Based Approach to Architecture Conformance Checking as a Quality Management Measure", in Relating System Quality and Software Architecture, Chapter 7, Pages 181-207. 2014

[8] https://www.terraform.io/docs/cloud/sentinel/index.html

[9] https://docs.microsoft.com/en-us/azure/governance/policy/overview

[10] https://docs.aws.amazon.com/.../developerguide/WhatIsConfig.html

[11] https://docs.aws.amazon.com/.../orgs_manage_policies_scps.html

[12] https://cloud.google.com/.../docs/organization-policy/overview

[13] Matthias Felleisen: "On the expressive power of programming languages", in *Science of Computer Programming*, Volume 17, Issues 1–31, Pages 35-75, 1991.

[14] Ulrich Lang, Rudolf Schreiner: "Analysis of recommended cloud security controls to validate OpenPMF 'policy as a service'", in *Information Security Technical Report*, Volume 16, Issues 3–4, Pages 131-141, 2011.

Sesión de Investigación B2: Gobierno y gestión de riesgos
http://doi.org/10.18239/jornadas_2021.34.47

Un resumen de "Un modelo de evaluación de la calidad de los datos y su aplicación a las fuentes de datos de ciberseguridad"

Enrique Pinto
Research Institute of Applied Science in Cybersecurity
Edificio MIC. Campus de Vegazana s/n 24071 León
eping@unileon.es
ORCID: 0000-0002-7253-9203

Noemí DeCastro-García
Departamento de Matemáticas. Universidad de León
Campus de Vegazana s/n 24071 León
ncasg@unileon.es
ORCID: 0000-0002-5610-0153

***Resumen*—La proliferación de grandes sistemas de almacenamiento implica que las empresas y las instituciones públicas deben evaluar la calidad de los datos que almacenan, para asegurarse de que las decisiones que se tomen a partir de ellos sean las más adecuadas. En particular, los equipos de respuesta a ciberincidentes trabajan con una gran cantidad de datos de diversas fuentes, y es vital asegurar su calidad para poder tomar las medidas adecuadas frente a incidentes de ciberseguridad.**

En este artículo, presentamos un resumen de un estudio ya publicado, en el que definimos una metodología de evaluación de la calidad de los datos, adaptada a un sistema de almacenamiento CERT. Teniendo en cuenta la naturaleza multidimensional de la calidad de los datos, proponemos un conjunto de dimensiones de calidad y establecemos la metodología para evaluarlas. También presentamos una herramienta de software que permite realizar la evaluación de forma automática.

***Palabras clave*—Calidad de los datos, Big data, CERT**

Tipo de contribución: *Investigación ya publicada. "A Data Quality Assessment Model and Its Application to Cybersecurity Data Sources", 13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020).*

I. Introducción

Los CSIRT/CERT (*Computer Security Incident Response Team* / *Computer Emergency Response Team*) reciben, desde diversas fuentes, grandes cantidades de datos relacionados con eventos e incidentes de ciberserguridad. A partir de la información extraída de estos datos, se toman decisiones y se establecen acciones de respuesta. Si la calidad de los datos no es adecuada, tampoco lo serán las decisiones ni las respuestas que se lleven a cabo. Además, los datos registrados relativos a los ciberincidentes son utilizados posteriormente para el desarrollo de sistemas inteligentes de detección. Por lo tanto, es de vital importancia analizar la calidad de los datos en sistemas de gestión y respuesta de ciberincidentes ([3], [4]).

En [1] se establece un método para evaluar la calidad de los datos que recibe un CSIRT/CERT y se aporta una herramienta [2] desarrollada en Python para automatizar el proceso.

La estructura de este artículo es la siguiente: la sección II presenta el modelo de evaluación de calidad de los datos, en la sección III se presenta el software desarrollado para ello y en la sección IV se detalla el caso de estudio para la validación de la metodología y el software. Finalmente, en la sección V, se establece el trabajo futuro y se exponen las conclusiones.

II. Modelo de evaluación de la calidad de los datos

La mayoría de estudios coinciden en que la calidad de los datos es un concepto multidimensional, es decir, debe ser analizada desde diversos puntos de vista, lo que sugiere el uso de diferentes dimensiones de calidad. Para nuestro sistema, basádonos en [5] y [6], hemos definido el conjunto de dimensiones $\mathfrak{D}$ = {Cantidad, Completitud, Nivel de información, Veracidad, Veracidad desconocida, Frecuencia, Consistencia, Relevancia, Precio, Evaluación manual}.

Dado que los datos proceden de diversas fuentes, estas dimensiones de calidad se evaluarán para cada una de ellas. Denominaremos $D = \cup D_j$ al conjunto de datos original, siendo D_j el conjunto de datos que proporciona la fuente S_j. Además, suponemos que tenemos datos sobre diferentes eventos E_h tales que $D = \cup D_h$.

Los datos proporcionados por las diversas fuentes son evaluados de acuerdo a las dimensiones de calidad $\mathfrak{D}_i$. Para cada dimensión, fijamos unos umbrales $I_i = [a_i, b_i]$ para determinar si la puntuación es buena, aceptable o mala. Las puntuaciones de cada dimensión son mostradas de dos formas:

- Datos en bruto: número de registros que cumplen los criterios de calidad.
- Datos normalizados: datos en bruto comparados con un valor de referencia. Esto nos da un número $\mathcal{V}(\mathfrak{D}_i) \in [0, 1]$ para cada dimensión, que podemos comparar con los umbrales de referencia I_i, definidos previamente.

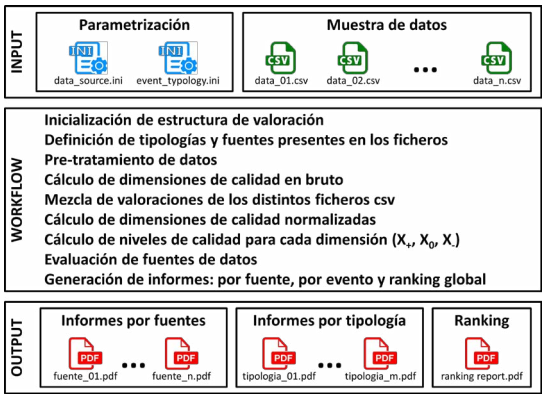


Figura 1. Flujo software evaluación calidad

Event typology	Quantity	Completeness	Level of information	Veracity	Unknow veracity	Frecuency	Consistency	High relevance	Medium relevance	Low relevance	Unknow relevance	Price per data
[redacted]	1,0	0,474	0,979	0,01	0,99	06:00:00	Medium	0.0	0.01	0.0	0.99	0,0
[redacted]	1,0	0,474	0,976	0,0	1,0	06:00:00	Medium	0.0	0.0	0.0	1.0	0,0
[redacted]	0,591	0,161	0,937	0,511	0,489	06:00:00	Medium	0.257	0.254	0.0	0.489	0,0
[redacted]	1,0	0,128	0,92	1,0	0,0	06:00:00	Medium	0.0	0.0	1.0	0.0	0,0

Figura 2. Informe de evaluación de calidad de una fuente: sección de puntuaciones normalizadas

Una vez que hemos evaluado cada dimensión, aplicamos la siguiente función sobre los datos normalizados:

$$\text{Color}_{S_j}^{E_h}(\mathfrak{D}_i) = \begin{cases} \text{Verde} & \text{si } \mathcal{V}_{S_j}^{E_h}(\mathfrak{D}_i) > b_i \\ \text{Amarillo} & \text{si } \mathcal{V}_{S_j}^{E_h}(\mathfrak{D}_i) \in I_i \\ \text{Rojo} & \text{si } \mathcal{V}_{S_j}^{E_h}(\mathfrak{D}_i) < a_i \end{cases} \quad (1)$$

donde $\text{Color}_{S_j}^{E_h}(\mathfrak{D}_i)$ denota el color (nivel) de la dimensión $\mathfrak{D}_i$ para la fuente de datos S_j en el evento E_h, y $\mathcal{V}_{S_j}^{E_h}(\mathfrak{D}_i)$ indica el valor de cada dimensión de calidad para la fuente de datos S_j en el evento E_h.

Para establecer una evaluación global para cada fuente se otorga una puntuación mediante la aplicación de la fórmula:

$$\mathtt{V}_{S_j} = w_+ \cdot \frac{|\mathrm{V}_{S_j}(\mathfrak{D})|}{|\mathfrak{D}|} + w_0 \cdot \frac{|\mathrm{A}_{S_j}(\mathfrak{D})|}{|\mathfrak{D}|} + w_- \cdot \frac{|\mathrm{R}_{S_j}(\mathfrak{D})|}{|\mathfrak{D}|} \quad (2)$$

donde $|\mathfrak{D}|$ denota el número total de dimensiones de calidad, $|\mathrm{V}_{S_j}(\mathfrak{D})|$, $|\mathrm{A}_{S_j}(\mathfrak{D})|$ y $|\mathrm{R}_{S_j}(\mathfrak{D})|$ son los cardinales de puntuaciones verdes, amarillas y rojas obtenidas por la fuente S_j, respectivamente, y w_+, w_0 y w_- son los pesos aplicados. Por defecto, w_+=1, $w_0 = \frac{1}{2}$ y $w_- = -1$, aunque estos valores son parametrizables.

A esa puntuación se le suma un coeficiente de diversidad que se calcula como el número de tipologías de eventos que proporciona la fuente entre el total de tipologías de eventos.

III. Herramienta de evaluación de la calidad

Para la aplicación del modelo de evaluación de calidad de una manera automática, se ha desarrollado un software en Python [2]. El programa utiliza como entrada:

- Dos ficheros .ini, que contienen la parametrización de las diferentes fuentes de datos y tipologías de eventos. En estos ficheros se definen, entre otras cosas, los umbrales para considerar una determinada dimensión de calidad es buena (verde), aceptable (amarillo) o mala (rojo).
- Un conjunto de ficheros .csv con la muestra de datos.

El *workflow* del software de evaluación, se puede consultar en la figura 1.

Como salida del programa se generan una serie de informes de calidad en formato pdf:

- Un informe por cada una de las fuentes, mostrando los resultados para cada una de las tipologías que proporciona, y en global. Las puntuaciones de cada tipología proporcionada por una fuente se muestran en bruto y normalizados (con los niveles indicados con la escala de colores de la fórmula 1).
- Un informe para cada una de las tipologías de eventos, detallando los resultados de cada una de las fuentes que los proporciona y proporcionando un ranking con las más adecuadas. Las puntuaciones de cada fuente que proporciona datos para la tipología también se muestran en bruto y normalizadas.
- Un ranking global de fuentes de datos, detallando su valoración de calidad, de diversidad y global.

IV. Caso de estudio

El software de evaluación de la calidad de los datos ha sido validado sobre una muestra de datos real, cedida por el Instituto Nacional de Ciberseguridad (INCIBE), bajo acuerdo de confidencialidad. Consta de 48 ficheros csv, con un total de 25.446.964 registros (filas) con 113 *features* (columnas). Los datos provienen de 28 fuentes y hacen referencia a 55 tipos de evento.

En la figura 2 puede observarse un ejemplo de la tabla de puntuaciones normalizadas obtenidas por una fuente de datos, para las cuatro tipologías de evento que proporciona.

V. Conclusiones

En conclusión, nuestro sistema proporciona una forma sencilla y rápida de evaluar la calidad de los datos, considerando diferentes dimensiones de calidad. Permite evaluar la calidad de cada fuente que proporciona datos al sistema. El sistema es flexible, ya que los archivos de configuración permiten incluir o eliminar fuentes de datos en los estudios, y permiten configurar los umbrales y referencias para las cuales una dimensión de calidad se considera buena, aceptable o mala.

Como trabajo futuro, la herramienta de evaluación de calidad sigue en evolución con la adición de nuevas dimensiones de calidad. Además se generarán nuevos informes agrupando las fuentes por su tipo (propias, públicas y privadas) y agrupando los distintos eventos. También está previsto generar otros rankings de fuentes aplicando metodología AHP (*Analytic Hierarchy Process*).

VI. Agradecimientos

Este trabajo se enmarca dentro de los contratos de investigación con clave orgánica X50 y X54 financiados por el Instituto Nacional de Ciberseguridad (INCIBE), y realizado en RIASC (ULE).

Referencias

[1] Noemí DeCastro-García y Enrique Pinto: "A Data Quality Assessment Model and Its Application to Cybersecurity Data Sources", en *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020)*, pp. 263–272, 2021.

[2] Enrique Pinto: "RIASC Data Quality Evaluation tool", en *https://github.com/amunc/DataQualityEvaluation*, 2020.

[3] J. J. Gonzalez: "Towards a Cyber Security Reporting System – A Quality Improvement Process", en *International Conference on Computer Safety, Reliability, and Security*, vol. 3688, pp. 368-380, 2005.

[4] G. Grispos, W. Bradley Glisson, T. Storer: "How Good is Your Data? Investigating the Quality of Data Generated During Security Incident Response Investigations", en *https://arxiv.org/abs/1901.03723v1*, 2019.

[5] Askham N, Cook D, Doyle M, Fereday H, Gibson M, Landbeck U, et al.: "The six primary dimension for data quality assessment- defining data quality dimensions", en *DAMA UK, https://www.whitepapers.em360tech.com/wp-content/filesmf/*, 2013.

[6] "Norma ISO/IEC 25012", en *https://iso25000.com/index.php/normas-iso-25000/iso-25012*, 2018.

Sesión de Investigación B3: Tecnologías emergentes y entrenamiento en ciberseguridad

http://doi.org/10.18239/jornadas_2021.34.48

IoT-as-a-Service: definición y retos tecnológicos

Santiago de Diego de Diego
TECNALIA, Basque Research and Technology Alliance (BRTA)
0000-0002-8823-7509
Bizkaia Science and Technology Park, 700,
E-48160 Derio, Bizkaia, Spain
santiago.dediego@tecnalia.com

Gabriel Maciá-Fernández
Dpto. Teoría de la Señal, Telemática y Comunicaciones.
CITIC-UGR
0000-0001-9256-453X
Universidad de Granada
España
gmacia@ugr.es

Cristina Regueiro Senderos
TECNALIA, Basque Research and Technology Alliance (BRTA)
0000-0002-6031-9449
Bizkaia Science and Technology Park, 700,
E-48160 Derio, Bizkaia, Spain
cristina.regueiro@tecnalia.com

Abstract- **El internet de las cosas (*Internet of Things,* IoT) está tremendamente presente en nuestro mundo actual, por lo que es probable que surjan numerosos modelos de negocio a partir de este paradigma tan sumamente ubicuo. Uno de ellos es el modelo de negocio IoT-as-a-Service (IoTaaS), que ofrece dispositivos IoT bajo demanda, con un considerable ahorro de costes y optimización de recursos. Sin embargo, se debe considerar la seguridad cuidadosamente cuando se trata del IoT debido a sus particularidades. Este artículo presenta formalmente el modelo de negocio del IoT como servicio (IoTaaS) y analiza el problema de la seguridad mostrando algunos retos que afectan a este modelo de negocio y sus implicaciones de seguridad.**

***Index Terms*- IoT, as-a-Service, IoTaaS, DLTs**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

La etiqueta "as-a-Service" se ha vuelto increíblemente popular hoy en día. En el contexto de Internet, "as-a-Service" significa que no es necesario poseer un recurso para poder utilizarlo. En los modelos "as-a-Service", los usuarios de un servicio pueden reducir costes, ya que evitan ser propietarios del material que no están interesados en poseer, mientras que los propietarios del servicio pueden aprovechar un material del que sí disponen, ofreciendo su uso a aquellos que lo necesitan y, por ende, monetizándolo. En otras palabras, "as-a-Service" suele implicar una situación en la que ganan tanto el usuario como el propietario del servicio. El caso de uso de Netflix [1] es un ejemplo perfecto de una exitosa solución "as-a-service" que redefine un caso de uso existente y tradicional (ver películas en Internet) con un modelo de negocio basado en la suscripción. En este caso, Netflix posee material (películas) que no necesita y ofrece a los usuarios, los cuales le pagan por poder verlas, en este caso mediante una cuota mensual. Sin embargo, también podrían mencionarse otros modelos de negocio, como el de pago por uso, en el que los usuarios pagan por el tiempo que utilizan el servicio. Otra solución "as-a-service" ampliamente conocida es Amazon Web Services (AWS) [2], donde cualquier usuario puede utilizar las capacidades de computación y almacenamiento en la nube pagando sólo por la cantidad de tiempo o el tráfico utilizados. De nuevo, Amazon proporciona el hardware y los usuarios pagan por el derecho a utilizarlo.

Por otro lado, el IoT se está integrando en numerosas soluciones existentes gracias a su ubicuidad. Según las cifras en [3], el número de dispositivos conectados en todo el mundo pasará de 20.4 billones en 2020 a 75.0 en 2025. Teniendo en cuenta la enorme cantidad de dispositivos IoT que se espera que existan en los próximos años, es probable que muchos de ellos se infrautilicen o incluso queden sin utilizar. Por ello, es conveniente pensar en ideas y soluciones que permitan reutilizar los dispositivos existentes, en lugar de fabricar otros nuevos, evitando las consecuencias negativas de la superpoblación de nuestro entorno con dispositivos IoT [4]. Además, los dispositivos del IoT a menudo se fabrican ad-hoc para un propósito particular, por lo que las soluciones que permitan reutilizar eficazmente estos dispositivos para ser empleados para otras aplicaciones serán muy bienvenidas.

En este contexto, el modelo de negocio IoT-as-a-Service (IoTaaS) se está acuñando como una solución a estos problemas, añadiendo un valor especial a la hora de permitir que diferentes aplicaciones reutilicen diferentes dispositivos IoT.

El presente trabajo de investigación en desarrollo realiza dos aportaciones principales: a) Presentar formalmente el modelo de negocio IoT-as-a-Service (IoTaaS) con sus características propias y b) proponer algunos retos que afectan a este modelo de negocio, con especial énfasis en la seguridad. Las secciones se organizan como sigue: La Sección II muestra los avances de la comunidad investigadora en relación con el modelo IoT-as-Service y analiza brevemente el estado actual de la seguridad para el IoT. La sección III y la sección IV presentan formalmente el modelo de negocio IoT-as-a-Service y sus retos asociados. Por último, la sección V extrae las conclusiones y presenta las posibles líneas futuras de investigación.

II. Estado del arte

A pesar de que algunos CEOs de diferentes empresas han sugerido el término IoTaaS, así como también lo ha hecho de forma preliminar la comunidad investigadora [5][6], estas menciones no profundizan en la descripción y el funcionamiento del modelo de negocio IoT-as-a-Service.

Paralelamente al aumento de la popularidad y la ubicuidad del IoT, han surgido numerosos problemas y retos que deben abordarse cuidadosamente. En particular, la seguridad de los dispositivos IoT es un aspecto crucial. Como afirman numerosos autores [7][8], no es trivial llegar a alcanzar un nivel de seguridad aceptable en los ecosistemas IoT. Algunos dispositivos IoT, como los *wearables*, pueden almacenar información sensible de sus usuarios y, por esta razón, los atacantes tienen cada vez más en cuenta a estos dispositivos

como objetivo de sus ataques. Por desgracia, algunos incidentes de seguridad muy conocidos, como Stuxnet [9] o la *botnet Mirai* y otros ataques relacionados con el IoT [10], que han infectado a millones de dispositivos IoT han puesto de manifiesto la necesidad de adoptar un enfoque que tenga en cuenta la seguridad a la hora de desarrollar soluciones relacionadas con el IoT. Otros ataques más generalistas son el robo de datos, el mal funcionamiento de los dispositivos y el control remoto de los mismos [11]. El conocido como ataque *Sybil* también afecta al IoT [12] y se produce por una deficiente gestión de la identidad. En este sentido, las soluciones que sean capaces de abordar adecuadamente los problemas de seguridad asociados a los ecosistemas de IoT serán muy bienvenidas.

III. Definición del modelo de negocio IoTaaS

En esta sección se analiza el modelo de negocio IoT-as-a-Service (IoTaaS) para identificar los principales retos tecnológicos que conlleva. Como paso previo, se presenta una formalización de este modelo.

Consideremos una persona o entidad denotada como *propietario*, que posee un conjunto de *dispositivos* IoT. Cada *dispositivo* ha sido fabricado por diferentes *fabricantes* y puede ser "contratado" por cualquier *consumidor* (persona o entidad) interesado durante un periodo de tiempo. Esta nomenclatura puede extrapolarse considerando varios *propietarios*, cada uno con su propio conjunto de *dispositivos*.

En este escenario, los *consumidores* requerirán determinados servicios del conjunto de *dispositivos*. La cantidad de dinero que cada *consumidor* pagará por el servicio dependerá de la cantidad de tiempo o recursos que consumirá de estos *dispositivos*, así como del tipo de servicio solicitado, siguiendo un modelo de negocio as-a-service.

Así, un *consumidor* podría estar interesado en los datos proporcionados por un *dispositivo* IoT, como pueden ser la temperatura o las lecturas de un generador de energía eólica, pero también en utilizar algunos servicios de este dispositivo. Por ejemplo, esos servicios podrían incluir una prueba de conectividad con otros sistemas, o una notificación en respuesta a determinados eventos. Los ejemplos de aplicaciones reales para el IoTaaS son lo suficientemente abundantes como para no detallarlos en este documento.

El modelo de negocio de IoTaaS estaría incompleto sin el concepto de *entidad de certificación*. Cada *consumidor* estará interesado en obtener servicios o datos de los *dispositivos*, si y sólo si esos *dispositivos* cumplen algunos requisitos de calidad. Las *entidades de certificación* atestiguan que los *dispositivos* cumplen realmente estos requisitos de calidad. En otras palabras, aportan confianza a los *consumidores*. Tanto los *propietarios* como los *fabricantes* son *entidades de certificación* en nuestro modelo, pero también pueden serlo muchos otros actores, como un auditor externo, entre otros, ya que pueden certificar algunos atributos de los dispositivos.

IV. Retos Tecnológicos del IoTaaS

A continuación, describimos brevemente algunos de los principales retos tecnológicos que aparecen a la hora de afrontar una implantación real del modelo de negocio IoTaaS.

A. Acceso a los servicios

Para empezar, uno de los aspectos relevantes a abordar es la definición de cómo se van a ofrecer los servicios proporcionados por los *dispositivos* y cómo van a acceder a ellos los diferentes *consumidores*. Con respecto al modelo de acceso a los servicios, el patrón *publicación-suscripción* es una opción para la comunicación entre *dispositivos* y *consumidores*, pero podrían seguirse otros modelos. En cuanto a la manera de obtener los servicios ofrecidos por los *dispositivos*, resulta bastante evidente que se necesitarán mecanismos para conectar a los *consumidores* con los *dispositivos*. Una alternativa, basada en otras soluciones as-a-Service, es el uso de un *marketplace*, es decir, un mecanismo para conectar los *dispositivos* IoT con sus potenciales *consumidores*. En el caso del IoTaaS, este *marketplace* será una interfaz en la que se listarán los *dispositivos* IoT para que los *consumidores* puedan elegir aquel que mejor se ajuste a sus necesidades. El *marketplace* tiene el objetivo de evitar que los *dispositivos* IoT se saturen respondiendo a las peticiones de los *consumidores*, lo que podría terminar en una Denegación de Servicio (DoS), especialmente si estos *dispositivos* tienen recursos limitados. Por lo tanto, actúa como un directorio donde los *consumidores* pueden comprobar qué *dispositivos* están disponibles, sin interferir en el normal funcionamiento de los mismos. Sin embargo, este reto es en sí mismo un tema abierto de investigación. Por ejemplo, centrándose en la seguridad, [13] propone un esquema de consulta que preserva la privacidad para el IoT utilizando un cifrado homomórfico. Además, el diseño final de este *marketplace* podría variar y hacerse más complejo en función de los requisitos finales de implementación [14]. La transferencia de datos entre los propios *dispositivos* IoT o entre dichos *dispositivos* y el *marketplace* puede realizarse mediante el uso de protocolos específicos para IoT, por ejemplo, MQTT-S, como es habitual en las redes de sensores inalámbricas [15], o cualquier otro protocolo relacionado con IoT. Asimismo, deben considerarse versiones seguras de los protocolos MQTT y MQTT-S (SMQTT y SMQTT-S respectivamente) [16] si los requisitos de seguridad son elevados. Además, la solución elegida debe incorporar un mecanismo que permita realizar consultas para que los *consumidores* puedan buscar *dispositivos* por diferentes parámetros, facilitando la selección del más adecuado a sus necesidades. De nuevo, esas consultas deben realizarse contra el *marketplace* para evitar la saturación de dispositivos y evitar una situación de DoS involuntaria. Además, el mecanismo elegido debe estar estandarizado y los *propietarios* deberán conocer este estándar para publicar correctamente sus dispositivos en el *marketplace*.

Hay algunos atributos de los *dispositivos* IoT que son importantes para ser mostrados y consultados en el IoTaaS, que son:

- Tipo de dato o servicio proporcionado por el dispositivo
- Certificaciones de la calidad del dispositivo
- Coste
- Localización y ámbito geográfico
- Propietario

A pesar de que este es el conjunto mínimo de atributos necesarios para implementar el modelo de negocio, dependiendo de la implementación final del modelo de negocio, algunos atributos más podrían ser incluidos en esta lista.

B. Optimización

Otro reto por analizar es el de la optimización, de forma que futuros trabajos sobre el IoTaaS podrían estudiar diferentes problemas de optimización asociados a dicho modelo de negocio. Por ejemplo, suponiendo el coste de un dispositivo IoT y el beneficio esperado en función de algunos aspectos como pueden ser el tipo de datos que proporciona el dispositivo, la ubicación o la calidad, un *propietario* puede estar interesado saber cómo distribuir sus *dispositivos* entre diferentes ubicaciones, o cuántos certificados de calidad deben tener sus dispositivos, a fin de maximizar el beneficio. Sin embargo, también podría estar interesado en reducir el coste lo máximo posible. A partir de esta fórmula se pueden construir modelos económicos más complejos, como considerar el coste implícito de la depreciación, lo cual repercutiría en el coste total, o un beneficio variable en función de algunos aspectos, como la temporada o la demanda.

C. Sistema de pagos

Otro reto a tener en cuenta es el hecho de proporcionar mecanismos que permitan realizar pagos. Estos mecanismos deben diseñarse cuidadosamente para evitar que sufran problemas de seguridad. Por ejemplo, los micropagos podrían realizarse de los *consumidores* a los *propietarios,* quizás a través del *marketplace* (esquema tradicional), o directamente a los *dispositivos* IoT siguiendo un esquema machine-to-machine (M2M). Estos pagos pueden gestionarse y rastrearse utilizando una Distributed Ledger Technology (DLT), como puede ser una blockchain. Con esta última solución, el proceso puede realizarse sin la intervención del propietario, ya que un contrato inteligente puede enviar al *propietario* el dinero directamente desde sus *dispositivos* IoT cuando se alcance cierta cantidad de dinero. En este enfoque, sería posible descentralizar la solución de pago, preservando así la privacidad de las partes involucradas mediante el uso de cualquier tipo de direcciones blockchain como las implementadas en Bitcoin [17] o Ethereum [18]. Estas direcciones son pseudo-anónimas, lo que significa que la privacidad no puede ser totalmente garantizada, pudiendo producirse algunos ataques como los ataques de desanonimización [19], que tratan de revelar la identidad existente detrás de una dirección.

Es por ello por lo que los micropagos no están exentos de riesgos de seguridad. Los sistemas blockchain tienen sus propios riesgos de ciberseguridad, como se analiza en [20]. Por ejemplo, focalizando en los relacionados directamente con el proceso de pago, el problema del doble gasto puede afectar a algunos sistemas de micropagos basados en blockchain [21] sin autoridad central, aunque este ataque tiene pocas probabilidades de éxito. Sin embargo, de llevarse a cabo satisfactoriamente, este ataque podría afectar a los *consumidores*, haciéndoles pagar más de una vez la cantidad de dinero que están pagando por un determinado *dispositivo*. Por lo tanto, es necesario tener esto en cuenta a la hora de diseñar el sistema de micropagos.

D. Identidad

Finalmente, el último reto a considerar está relacionado con la gestión de la identidad. La suplantación es un riesgo de seguridad común y también está presente en el IoTaaS ya que los diferentes actores (*propietarios, dispositivos, fabricantes, marketplace*) pueden ser suplantados. Dicha suplantación se puede prevenir con una sólida gestión de identidad. Los autores en [22] dan una amplia visión sobre el concepto de identidad y cómo ha evolucionado con el tiempo. En resumen, la identidad ha sufrido varias transformaciones durante los últimos años, empezando por un modelo aislado y centralizado y pasando después a un modelo federado. Sin embargo, estos modelos adolecen de graves problemas de escalabilidad y seguridad. El último modelo sugerido es el de la identidad autosoberana (SSI). La SSI ofrece muchas ventajas [23] frente al resto de modelos de gestión de identidad: modelos aislados, centralizados y federados. En primer lugar, elimina los llamados "centros de identidad" permitiendo a los usuarios gestionar sus propias identidades de forma descentralizada. En segundo lugar, y más importante cuando se trata del IoT, escala mejor como consecuencia de la eliminación de dichos centros de identidad, lo cual cobra especial relevancia en el entorno del IoT [22], donde la ubicuidad y la existencia de una miríada de dispositivos son características esenciales de los dispositivos IoT [8]. Este hecho complica claramente el proceso de gestión de la identidad con los enfoques tradicionales. Por último, los autores de [24] reconocen algunos beneficios adicionales al aplicar SSI al IoT, que son: se incrementan los ingresos y se reducen los costes y los riesgos. Este trabajo también analiza qué amenazas de ciberseguridad que afectan al IoT pueden mitigarse utilizando un sistema de gestión de identidad basado en SSI.

Adicionalmente, una cuestión relevante es si los dispositivos del IoT tienen suficientes recursos para ejecutar un esquema de identidad basado en SSI. Los autores en [25] analizan con precisión los requisitos mínimos que deben tener los dispositivos IoT para ejecutar una solución SSI y concluyen que la mayoría de los dispositivos son capaces de hacerlo. Además, en caso de que esto no se cumpla, proponen una aproximación basada en un proxy para dispositivos extremadamente restringidos.

La protección de las claves criptográficas, que son la piedra angular de los procesos de firma, es crucial para el proceso de gestión de la identidad. Un *Trusted Execution Environment* (TEE) [26] se conoce comúnmente como un entorno de procesamiento aislado en el que se pueden ejecutar aplicaciones, separadas del resto del sistema y, lo que es más relevante para el presente trabajo, se puede utilizar para proteger el proceso de firma, así como los datos (claves criptográficas) implicados en el mismo. El proceso de firma también puede protegerse empleando claves criptográficas integradas utilizando un *Trusted Platform Module* (TPM) para sistemas embebidos [27].

E. Resumen de los mecanismos de seguridad

En resumen, la Tabla I incorpora las medidas de seguridad propuestas con el resultado esperado en términos de

Tabla I
REQUISITOS DE SEGURIDAD PROPORCIONADOS POR ALGUNAS MEDIDAS DE SEGURIDAD PROPUESTAS

	Privacidad	Integridad	Disponibilidad	Autenticación
Marketplace	No	No	Si	No
SMQTT [16]	Si	Si	No	Si
DLTs	Parcial	Si	Si	Solo privadas
SSI	Si	Si	Si	Si
TEE	Si	Si	No	No
TPM	Si	Si	No	No

preservación de la privacidad, integridad, disponibilidad y autenticación.

En cuanto a los sistemas propuestos, el *marketplace* no es en sí mismo una medida de seguridad, pero se ha incluido porque puede evitar situaciones de DoS, y por tanto proporcionar una mejora de la disponibilidad. SMQTT permite obtener privacidad, integridad y autenticación gracias a la aplicación de *Key/Cipher text Policy-Attribute Based Encryption* (KP/CP-ABE) con *Elliptic Curve Cryptography* (ECC) al protocolo MQTT. Las tecnologías DLT por su parte aportan integridad por diseño y una mayor disponibilidad a medida que aumenta el número de nodos. Asimismo, no todas proporcionan privacidad, siendo algunas de ellas pseudo-anónimas, como se ha mencionado con anterioridad, y aquellas que son privadas necesitan de un proceso de autenticación para interactuar con ellas. La mejora en los cuatro requisitos expuestos de seguridad gracias a SSI puede ser consultada en [24][25]. Un TEE proporciona funciones de privacidad e integridad a los datos, ya que las aplicaciones y sus datos asociados se ejecutan separadas del resto del sistema (privacidad) con mecanismos que preservan la integridad; y de igual manera los TSM se emplean generalmente para preservar la privacidad e integridad de las claves criptográficas.

V. Conclusiones

En este trabajo se ha analizado el modelo de negocio IoT-as-a-Service (IoTaaS), así como sus retos tecnológicos asociados, con un claro enfoque en la seguridad. Esto abre nuevas vías de investigación en cuanto a sistemas de micropagos, protocolos de gestión de identidad u otras medidas relacionadas con la seguridad, todas ellas factibles de ser aplicadas en el IoTaaS. El problema de la identidad es especialmente crítico en el IoT, y por ello en el IoTaaS, por lo que enfoques como el de la Identidad AutoSoberana (SSI) pueden ser interesantes a la hora de implementar un sistema de gestión de identidad para el IoTaaS.

Agradecimientos

Este trabajo ha sido parcialmente apoyado por el Gobierno del País Vasco bajo el programa ELKARTEK, proyecto TRUSTIND (KK-2020/00054) y por el MINECO (Ministerio de Economía y Competitividad) a través del proyecto TIN2017-83494-R.

Referencias

[1] AU-YONG-OLIVEIRA, Manuel; MARINHEIRO, Miguel; TAVARES, João A. Costa. The Power of Digitalization: The Netflix Story. En World Conference on Information Systems and Technologies. Springer, Cham, 2020. p. 590-599.

[2] VARIA, Jinesh, et al. Overview of amazon web services. Amazon Web Services, 2014, vol. 105.

[3] Review42. Internet of Things Statistics, Facts & Predictions. [Online]. Available in: https://review42.com/internet-of-things-stats/.

[4] GREU, Victor, et al. Facing IoT-The New Giant Wave of the Information and Communications Technologies Development. Romanian Distribution Committee Magazine, 2015, vol. 6, no 4, p. 18-25.

[5] DENG, Der-Jiunn; PANG, Ai-Chun; HANZO, Lajos. Recent Advances in IoT as a Service (IoTaaS 2017). Mobile Networks and Applications, 2019, vol. 24, no 3, p. 721-723.

[6] BRINCAT, Alberto Attilio; PACIFICI, Federico; MAZZOLA, Francesco. IoT as a service for smart cities and nations. IEEE Internet of Things Magazine, 2019, vol. 2, no 1, p. 28-31.

[7] KHAN, Minhaj Ahmad; SALAH, Khaled. IoT security: Review, blockchain solutions, and open challenges. Future Generation Computer Systems, 2018, vol. 82, p. 395-411.

[8] ZHOU, Wei, et al. The effect of iot new features on security and privacy: New threats, existing solutions, and challenges yet to be solved. IEEE Internet of Things Journal, 2018, vol. 6, no 2, p. 1606-1616.

[9] LANGNER, Ralph. Stuxnet: Dissecting a cyberwarfare weapon. IEEE Security & Privacy, 2011, vol. 9, no 3, p. 49-51.

[10] KAMBOURAKIS, Georgios; KOLIAS, Constantinos; STAVROU, Angelos. The Mirai botnet and the IoT zombie armies. En MILCOM 2017-2017 IEEE Military Communications Conference (MILCOM). IEEE, 2017. p. 267-272.

[11] LU, Yang; DA XU, Li. Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 2018, vol. 6, no 2, p. 2103-2115.

[12] RAJAN, Anjana; JITHISH, J.; SANKARAN, Sriram. Sybil attack in IOT: Modelling and defenses. En 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2017. p. 2323-2327.

[13] LU, Rongxing. A new communication-efficient privacy-preserving range query scheme in fog-enhanced IoT. IEEE Internet of Things Journal, 2018, vol. 6, no 2, p. 2497-2505.

[14] ZHENG, Weijun. The business models of e-marketplace. Communications of the IIMA, 2006, vol. 6, no 4, p. 1.

[15] HUNKELER, Urs; TRUONG, Hong Linh; STANFORD-CLARK, Andy. MQTT-S—A publish/subscribe protocol for Wireless Sensor Networks. En 2008 3rd International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE'08). IEEE, 2008. p. 791-798.

[16] SINGH, Meena, et al. Secure MQTT for internet of things (IoT). En 2015 fifth international conference on communication systems and network technologies. IEEE, 2015. p. 746-751.

[17] SATOSHI, Nakamoto. Bitcoin: A peer-to-peer electronic cash system. Consulted, 2008, vol. 1, no 2012, p. 28.

[18] BUTERIN, Vitalik, et al. A next-generation smart contract and decentralized application platform. white paper, 2014, vol. 3, no 37.

[19] BIRYUKOV, Alex; TIKHOMIROV, Sergei. Deanonymization and linkability of cryptocurrency transactions based on network analysis. En 2019 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2019. p. 172-184.

[20] DASGUPTA, Dipankar; SHREIN, John M.; GUPTA, Kishor Datta. A survey of blockchain from security perspective. Journal of Banking and Financial Technology, 2019, vol. 3, no 1, p. 1-17.

[21] CHOHAN, Usman W. The double spending problem and cryptocurrencies. Available at SSRN 3090174, 2017.

[22] ZHU, Xiaoyang; BADR, Youakim. Identity management systems for the internet of things: a survey towards blockchain solutions. Sensors, 2018, vol. 18, no 12, p. 4215.

[23] TOBIN, Andrew; REED, Drummond. The inevitable rise of self-sovereign identity. The Sovrin Foundation, 2016, vol. 29, no 2016.

[24] S. FOUNDATION. Self-Sovereign Identity and IoT [Online]. Available: https://sovrin.org/wp-content/uploads/SSI-and-IoT-whitepaper.pdf. 2020. Accedido el: 1 Enero, 2021

[25] KORTESNIEMI, Yki, et al. Improving the privacy of IoT with decentralised identifiers (DIDs). Journal of Computer Networks and Communications, 2019, vol. 2019.

[26] SABT, Mohamed; ACHEMLAL, Mohammed; BOUABDALLAH, Abdelmadjid. Trusted execution environment: what it is, and what it is not. En 2015 IEEE Trustcom/BigDataSE/ISPA. IEEE, 2015. p. 57-64.

[27] KINNEY, Steven L. Trusted platform module basics: using TPM in embedded systems. Elsevier, 2006.

http://doi.org/10.18239/jornadas_2021.34.49

Análisis de Seguridad y Privacidad en dispositivos de la Internet de las Cosas usados por jóvenes

Sonia Solera-Cotanilla[1], Mario Vega-Barbas[1], Manuel Álvarez-Campana Fernández-Corredor[1], Cayetano Valero Amores[2], Jaime Fúster de la Fuente[2], Gregorio López López[2]
[1]ETSI Telecomunicación, Universidad Politécnica de Madrid
{sonia.solera, mario.vega, manuel.alvarez-campana}@upm.es
[2]Instituto de Investigación Tecnológica, ICAI, Universidad Pontificia Comillas
{201507104, jaimeff}@alu.comillas.edu, gllopez@comillas.edu

***Resumen*—Bajo el paradigma de la Internet de las Cosas, en los últimos años ha proliferado la aparición y el uso de una amplia variedad de dispositivos conectados. En este artículo se presentan los resultados del análisis de problemas de seguridad y privacidad de dichos dispositivos efectuado dentro del Proyecto Europeo RAYUELA, cuyo principal objetivo es fomentar el uso seguro de Internet por parte de los jóvenes. Los resultados del análisis presentan un conjunto de vulnerabilidades asociadas al uso de los dispositivos conectados más utilizados por los menores. Se pone en evidencia que, con frecuencia, los fabricantes de los dispositivos priorizan las funcionalidades y los servicios a proporcionar, dejando en un segundo plano los aspectos de seguridad. Del mismo modo, el estudio revela que, en ocasiones, las empresas tratan de explotar los datos ligados al uso de estos dispositivos con diversos fines, obviando el derecho a la privacidad de los usuarios.**

***Index Terms*—Internet de las Cosas, Seguridad, Privacidad, Vulnerabilidad, Dispositivos Conectados.**

Tipo de contribución: *Contribución científica original en estado preliminar y/o en desarrollo*

I. Introducción

La computación ubicua, la tecnología pervasiva y la inteligencia ambiental representan tres tecnologías que han condicionado en gran medida el avance y el desarrollo de la actual Sociedad de la Información. El auge de este ecosistema digital de dispositivos y objetos cotidianos conectados entre sí, conocido como Internet de las Cosas (IoT) [1], ha facilitado la creación de servicios novedosos y personalizados. La utilidad de dichos dispositivos está estrechamente relacionada con su adopción en la vida cotidiana de las personas [2].

Conforme aumenta el uso de los dispositivos conectados, mayor es el grado de personalización y la utilidad de los servicios y las aplicaciones que la IoT pone a nuestra disposición. De este modo, el concepto de ecosistema de dispositivos interconectados ha tenido, en los últimos años, un gran impacto en la consolidación de, por ejemplo, los servicios de telemedicina y salud electrónica [3] o la industria 4.0 [4]. Ello ha facilitado la creación de nuevos contextos de mercado como el de los juguetes y el desarrollo infantil [5] o el de la industria del automóvil [6].

No obstante, a pesar de los beneficios que ofrece el paradigma de la IoT, continuamente se detectan numerosas debilidades en los dispositivos conectados y la aplicaciones subyacentes, comprometiendo la privacidad y la seguridad de los usuarios finales. Estas debilidades se deben, en gran parte, a la deficiencia o carencia de mecanismos de seguridad robustos que les protejan frente ciberataques, así como, en muchas ocasiones, a los propios usuarios que, por descuido o desconocimiento, no toman las debidas precauciones de uso (por ejemplo, no es extraño encontrar usuarios que utilizan contraseñas por defecto o débiles). Este tipo de factores hace que los dispositivos conectados se encuentren actualmente en el punto de mira de los ciberataques más frecuentes [7], [8].

El presente trabajo se enmarca en el proyecto RAYUELA [9] (*empoweRing and educAting YoUng pEople for the internet by pLAying*) financiado por el programa H2020 de la Unión Europea. En el proyecto, liderado por Comillas, participa un consorcio multidisciplinar formado por diecisiete socios de nueve países europeos, incluyendo universidades, centros de investigación, fuerzas y cuerpos de seguridad del Estado y compañías tecnológicas. El principal objetivo de RAYUELA es el desarrollo de un juego de tipo aventura interactiva, para los menores de entre doce y diecisiete años, orientado a concienciar a los jóvenes sobre los riesgos y amenazas que entraña el uso de Internet y los dispositivos conectados, así como educarles sobre las buenas prácticas para protegerse frente a ellos. Además de los aspectos tecnológicos asociados a los riesgos de uso de los dispositivos conectados, el proyecto aborda los factores psicológicos, antropológicos y sociológicos que influyen en ciberdelitos como el *online grooming*, el *cyberbullying* y la trata de personas con fines de explotación sexual.

Dentro del proyecto, el presente trabajo se centra en el análisis del paradigma tecnológico de la IoT y las potenciales amenazas que, desde el punto de vista de la seguridad y la privacidad, conlleva su uso por parte de los menores.

Este colectivo resulta de especial interés puesto que, según el informe de UNICEF "Estado Mundial de la Infancia 2017: Los niños en el mundo digital", se estima que los niños y adolescentes menores de dieciocho años suponen más de un tercio de los usuarios de Internet en todo el mundo [10] y Eurostat estima que el 96 % de los adolescentes (de dieciséis a diecinueve años) utilizan Internet a diario en Europa [11]. Estos jóvenes, nacidos entre 2003 y 2008, son los denominados nativos digitales y destacan por una alta adopción e incluso dependencia de la tecnología en su vida cotidiana, por lo que el presente estudio trata de entender esta relación innata entre los jóvenes y cualquier tipo de dispositivo conectado del

ecosistema de la IoT [12].

II. Metodología

El objetivo fundamental del trabajo es analizar las principales vulnerabilidades y deficiencias de los dispositivos conectados más utilizados por los jóvenes. Con ello, se pretende obtener una visión clara de los principales vectores de ataque o puntos débiles de estas tecnologías, sentando las bases para concienciar y educar a los jóvenes sobre los riesgos y amenazas que su uso entraña, así como sobre las buenas prácticas de uso para evitarlos.

La metodología de análisis se ha basado en una revisión sistemática de la literatura. La búsqueda inicial se centró en determinar cuáles son los dispositivos conectados más utilizados por el segmento de población objetivo. A continuación, se procedió a recopilar un conjunto significativo de artículos de las bases de datos que se han considerado de principal interés, como son IEEE Xplore Digital Library, ScienceDirect, ACM Digital Library, Web of Science y Google Scholar. En todas ellas se ha realizado una búsqueda inicial con palabras clave como "Internet of Things", "children", "security", "privacy", etc. Y se ha refinado dicha búsqueda con palabras clave específicas, como por ejemplo "smartwatch", en caso de querer obtener informes relacionados con *wearables*.

Esta primera búsqueda se complementó con otra más amplia incluyendo conferencias, noticias e informes tecnológicos industriales y, para garantizar que las referencias consideradas fueran oportunas, de esta búsqueda con cientos de informes encontrados solo se tuvieron en cuenta aquellos fechados a partir del 2017. Como resultado de este filtrado y considerando solo aquellos que abordan aspectos concretos de seguridad y/o privacidad de uso de los dispositivos IoT por parte de los jóvenes, se obtuvieron 61 artículos especialmente relacionados con el estudio. Además, de este conjunto resultante se descartaron 16 por no aportar información adicional respecto al resto, resultando los 45 artículos en los que se basa finalmente el estudio.

En el siguiente apartado se proporciona un resumen de los resultados más relevantes obtenidos en el estudio.

Las referencias obtenidas en la primera fase se agruparon en base a las categorías principales de dispositivos utilizados por los menores. Dentro de cada grupo de dispositivos, se analizaron y compararon con detalle las correspondientes referencias, tratando de identificar los problemas de seguridad y privacidad comunes. Durante esta tarea, se llevó a cabo la definición de un glosario de cuestiones de seguridad y privacidad. Por último, se evaluó cuantitativamente la relevancia de los distintos términos del glosario en cada categoría de dispositivo, en función del número de referencias en las que aparecían.

III. Resultados

III-A. Categorización de dispositivos conectados

Como se ha indicado, el análisis se centra en los dispositivos conectados más utilizados por los jóvenes. Para determinar cuáles son estos, se optó por la realización de una encuesta orientada a los menores que, desafortunadamente, no ha podido ser completada aún. En ella se incluyen preguntas relativas a las aplicaciones y dispositivos conectados que más integrados se encuentran en su vida cotidiana. Para obtener esta información, se pregunta por la frecuencia de uso de determinados dispositivos así como por la concienciación de seguridad y privacidad que ofrecen y las amenazas que pueden acarrear. Todas estas preguntas se han desarrollado desde un punto de vista no técnico para que la encuesta pueda ser respondida por usuarios con poco o nulo conocimiento técnico del ámbito que nos concierne. Como alternativa, y a la espera de obtener los resultados de dicha encuesta, se ha recurrido a la información proporcionada por varios informes recientes [13], [14], [15].

Según los citados informes, el dispositivo más utilizado por los jóvenes es el *smartphone*, seguido del *smart TV* y el ordenador de sobremesa o portátil, y, con una incidencia similar, las *tablets* y videoconsolas. Por otro lado, los juguetes inteligentes y los *wearables* ocupan posiciones menos significativas, aunque su uso ha crecido significativamente en los últimos años. Con un uso aún marginal también se hace referencia a los asistentes personales inteligentes, los altavoces inteligentes, los dispositivos para gestionar aparatos eléctricos o de climatización, las pizarras inteligentes, etc. A partir de esta información se han definido las siguientes categorías de dispositivos conectados a considerar para el análisis de vulnerabilidades (teniendo en cuenta que los ordenadores se han dejado fuera del estudio al no considerarse parte de ninguna de las categorías):

1. *Smartphones* y *tablets*.
2. *Smart TVs* y videoconsolas.
3. Juguetes inteligentes.
4. *Wearables*.
5. Dispositivos IoT domésticos inteligentes.
6. Asistentes personales inteligentes.
7. Altavoces inteligentes.
8. Otros, como drones o cámaras.

III-B. Identificación de problemas de seguridad y privacidad

El análisis de los problemas de seguridad y privacidad de los dispositivos conectados expuesto ha tenido en cuenta, tras el filtrado de informes realizado y mencionado en la metodología, más de cuarenta trabajos de investigación enfocados a la categorización mencionada. La Fig. 1 muestra la distribución de las referencias estudiadas por tipo de dispositivo categorizado.

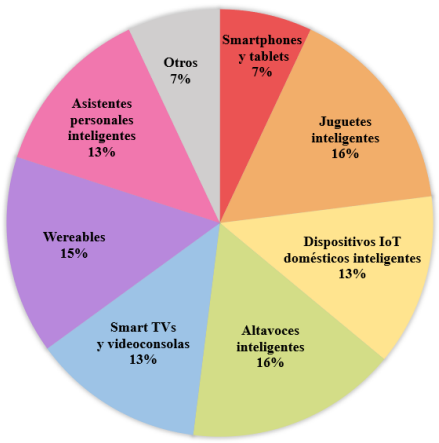


Figura 1. Referencias analizadas por categoría de dispositivo.

En función de este análisis sistemático de vulnerabilidades y deficiencias de los dispositivos pertenecientes a las categorías expuestas, se ofrece como resultado principal una clasificación de los problemas de seguridad y privacidad analizados en diez grupos diferentes, siete referidos a la seguridad y tres a cuestiones de privacidad. Cada uno de estos problemas agrupa vulnerabilidades o ataques de seguridad o riesgos de privacidad similares o relacionados. Con respecto a los problemas de seguridad, los siete grupos establecidos son:

1. Suplantación de la identidad del usuario o del dispositivo conectado.
2. Ausencia o debilidad de cifrado, consiste en la exposición de datos durante la transferencia de información entre pares debido a que estos se intercambian en texto plano o están protegidos con medios de cifrado poco fiables u obsoletos.
3. Ausencia o debilidad de autenticación debido a mecanismos de autenticación obsoletos o nulos que permiten el acceso al dispositivo con un rol específico.
4. Interacción de voz no controlada que permite la ejecución de comandos de voz por parte usuarios no autorizados, así como ataques de canal lateral. Este problema de seguridad está relacionado con los ataques de enmascaramiento e invasión de voz (VMA y VSA respectivamente por sus siglas en inglés), así como los comandos de voz ocultos.
5. Inyección de código que da lugar a la ejecución de comandos maliciosos preparados para modificar el funcionamiento común del sistema o facilitar el acceso no autorizado a partes o datos protegidos, como, por ejemplo, inyección de comandos SQL.
6. Intersección de datos mediante la escucha activa o pasiva (*sniffing*) de las comunicaciones entre dispositivos interconectados que pasa desapercibida para los usuarios comunes, como, por ejemplo, *Man-in-the-Middle*.
7. Toma de posesión, consiste en el control total del dispositivo para acceder a los datos o realizar ataques que requieren la cooperación entre los dispositivos conectados, como, por ejemplo, el ataque de Denegación de Servicio Distribuido (DDoS).

Con respecto a los problemas de privacidad identificados en los informes, destacan los siguientes:

1. Compromiso de los datos del usuario debido a pérdida y/o acceso indebido o no autorizado de dichos datos por un funcionamiento inesperado del dispositivo, el servidor subyacente o las aplicaciones de terceros.
2. Violación de las leyes de privacidad mediante el uso indebido de datos sensibles y/o personales que implique una violación total o parcial de leyes específicas de privacidad como el Reglamento General de Protección de Datos (GDPR, en inglés), la Ley de Protección de la Privacidad en Línea para Niños (COPPA, en inglés), etc.
3. Falta de control y comprensión sobre sobre la gestión de los datos del usuario por parte de los dispositivos y aplicaciones o servicios subyacentes. Esta cuestión tiene en cuenta la percepción del usuario sobre lo que ocurre con los datos personales gestionados por el dispositivo o la aplicación subyacente.

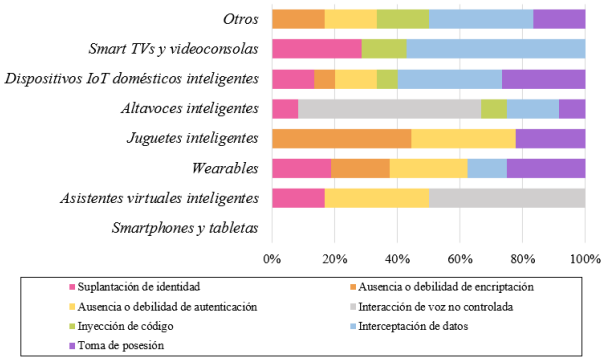


Figura 2. Problemas de seguridad.

III-C. Análisis de incidencias por tipo de dispositivo

Las Fig. 2 y 3 muestran la incidencia de cada problema de seguridad y privacidad, respectivamente, en cada tipo de dispositivo conectado. Esta incidencia se ha determinado a partir del número de informes que hacen referencia al tipo de dispositivo en cuestión y a los problemas tratados. Como puede verse en la Fig. 2, la categoría más afectada por los problemas de seguridad, según la literatura analizada, son los dispositivos IoT domésticos inteligentes, comprometida por seis de siete problemas. Le siguen los *wearables* y los altavoces inteligentes, con cinco de siete problemas. El resto de categorías se han visto afectadas por al menos tres del total de problemas de seguridad definidos. En el caso de los *smartphones* y *tablets*, debido a que el estudio se ha centrado en aplicaciones, esta incidencia se ha relacionado principalmente con los problemas de privacidad por violación de determinadas leyes, dejando de lado los de seguridad que no resultan reseñables frente a los encontrados en el resto de categorías de dispositivos.

Por su parte, la Fig. 3 nos muestra que todas las categorías, a través de los problemas de privacidad detectados, permiten comprometer los datos de los usuarios. Además, los artículos analizados indican que los usuarios de todas las categorías, excepto los juguetes inteligentes y asistentes virtuales inteligentes, muestran dudas o temores sobre la gestión que los dispositivos hacen de los datos personales. Por último, sólo los artículos orientados al análisis de aplicaciones o dispositivos especialmente diseñados para menores (juguetes inteligentes, *smartphones* y *tablets*) han detectado violaciones de las leyes de privacidad, concretamente de la COPPA.

IV. Discusión y Conclusiones

La investigación presentada en este documento ofrece un análisis de la seguridad y privacidad del paradigma tecnológico de la IoT orientado a jóvenes de entre doce y diecisiete años. El objetivo principal del estudio es advertir sobre las vulnerabilidades asociadas a los dispositivos conectados más utilizados por el colectivo objetivo en su vida diaria.

En los últimos años, el desarrollo de las Tecnologías de la Información y la Comunicación y la IoT ha adquirido tal nivel de madurez que se ha favorecido la creación de nuevos servicios telemáticos más eficientes y eficaces. Esto se debe, principalmente, a una alta penetración de este ecosistema

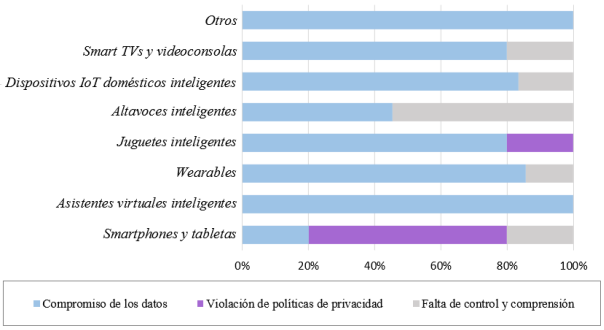


Figura 3. Problemas de privacidad.

de dispositivos interconectados en nuestra vida cotidiana, lo que facilita una mayor personalización de dichos servicios. Sin embargo, aunque los beneficios de la IoT son claros, el conjunto de dispositivos que lo componen se enfrenta a multitud de vulnerabilidades asociadas a su uso. Por ello, se ha realizado una revisión sistemática de la literatura sobre los tipos de dispositivos conectados más utilizados por el grupo objetivo (menores de doce a diecisiete años). Dicho análisis parte de una categorización de los dispositivos compuesta por ocho grupos de interés: *smartphones* y *tablets*, *smart TVs* y videoconsolas, juguetes inteligentes, *wearables*, dispositivos IoT domésticos inteligentes, asistentes personales inteligentes, altavoces inteligentes, y otros, como cámaras web o drones. En este sentido, es importante aclarar que, aunque los dispositivos más utilizados por los jóvenes son los *smartphones* y las *tablets*, el análisis realizado se ha centrado en aquellos informes que ofrecen información de seguridad y privacidad sobre las aplicaciones de uso extendido en lugar de la propia tecnología.

Como resultado principal de esta investigación se ha generado un glosario de incidencias y problemas de seguridad y privacidad. Este glosario está compuesto por siete vulnerabilidades principales de seguridad y tres de privacidad, que son: suplantación de identidad, ausencia o debilidad de cifrado, ausencia o debilidad de autenticación, interacción de voz no controlada, inyección de código, intersección de datos y toma de posesión, en lo que respecta a la seguridad; y compromiso de los datos del usuario, violación de las leyes de privacidad y falta de control y comprensión, en lo que respecta a la privacidad. El conjunto de vulnerabilidades identificadas en el glosario cubre todas las áreas de seguridad y privacidad que suponen amenazas de ataques a los distintos dispositivos contemplados en el estudio.

Mediante el estudio de la incidencia de cada problema sobre cada categoría de dispositivo (ver Fig. 2 y 3), se concluye que la mayoría de los problemas de seguridad están generados por una débil o inadecuada autenticación y/o encriptación de los datos gestionados. Esto favorece otros problemas como la interceptación de datos, la suplantación de identidad o la toma de control por terceros de los dispositivos. Además, este problema de seguridad tiene incidencia sobre la privacidad ofrecida por los dispositivos. La mayoría de las referencias analizadas que analizan problemas de privacidad remarcan que los datos gestionados por los dispositivos y aplicaciones subyacentes pueden estar comprometidos. Sin embargo, sólo aquellos trabajos centrados en el análisis de juguetes inteligentes y aplicaciones para *smartphone* y *tablets* especifican una violación concreta de regulaciones o leyes de privacidad. En este sentido, destaca la falta de análisis del GDPR, ya que la totalidad de referencias analizadas en relación a ello han utilizado COPPA como base de estudio.

V. Trabajos futuros

Tomando como punto de partida los resultados de esta investigación, se plantea como trabajo futuro dos líneas de investigación claras. En primer lugar, se pretende definir una metodología de evaluación del nivel de seguridad y privacidad de los dispositivos de la IoT orientados a menores. En segundo lugar, se plantea un análisis de los posibles factores humanos que afectan la vulnerabilidad y los riesgos detectados. En base a los principales hallazgos realizados en esta fase de investigación, se desarrollarán ciberaventuras como parte de un juego serio para ayudar a que los menores adquieran buenas prácticas en el uso de este tipo de dispositivos.

Agradecimientos

Este trabajo ha sido financiado por el programa Horizon 2020 de la Unión Europea a través del proyecto RAYUELA (nº de contrato 882828). El contenido del artículo refleja solo el punto de vista de sus autores. La Comisión Europea no es responsable del uso que se pueda hacer de la información que contiene.

Referencias

[1] K. Ashton: “That Internet of Things”, en *RFID Journal*, 22, 2000.
[2] M. Vega-Barbas, I. Pau y F. Seoane: “From general services to pervasive and sensitive services”, en *Encyclopedia of Information Science and Technology, Fourth Edition*, IGI Global, pp. 7754-7764, 2018.
[3] M.M. Alam et al.: “A Survey on the Roles of Communication Technologies in IoT-Based Personalized Healthcare Applications”, en *IEEE Access*, 6, pp. 36611-36631, 2018.
[4] Y. Lu: “Industry 4.0: A Survey on Technologies, Applications and Open Research Issues”, en *Journal of industrial information integration*, 6, pp. 1-10, 2017.
[5] D. Holloway y L. Green: “The Internet of Toys”, en *Communication Research and Practice*, 2, pp. 506-519, 2016.
[6] F. Yang, S. Wang, J. Li, Z. Liu y Q. Sun: “An Overview of Internet of Vehicles”, en *China communications*, 11, pp. 1-15, 2014.
[7] H.F. Atlam, A. Alenezi, M.O. Alassafi, A.A. Alshdadi y G.B. Wills: “Security, cybercrime and digital forensics for IoT. In Principles of Internet of Things (IoT) Ecosystem: Insight Paradigm”, en *Springer*, pp. 551-577, 2020.
[8] S. Hilt, V. Kropotov, F. Mercês, M. Rosario y D. Sancho: “The Internet of Things in the Cybercrime Underground”, en *Trend Micro Research*, 2019.
[9] “RAYUELA: A fun way to fight cybercrime”, 2020. [En línea]. Disponible en: https://www.rayuela-h2020.eu/. [Accedido: 23 de marzo, 2021].
[10] B. Keeley y C. Little: “The State of the Worlds Children 2017: Children in a Digital World”, en *ERIC*, 2017.
[11] Eurostat. Being Young in Europe Today–digital World. *European Commission*, 2020.
[12] S. Bennett, K. Maton y L. Kervin: “The ‘digital Natives’ Debate: A Critical Review of the Evidence”, en *British journal of educational technology*, 39, pp. 775-786, 2008.
[13] S. Livingstone, G. Mascheroni y E. Staksrud: “European Research on Children’s Internet use: Assessing the Past and Anticipating the Future”, en *New Media & Society*, 20, pp. 1103-1122, 2018.
[14] D. Smahel et al.: EU Kids Online 2020: Survey Results from 19 Countries. 2020.
[15] T. Cabello-Hutt, P. Cabello y M. Claro: “Online Opportunities and Risks for Children and Adolescents: The Role of Digital Skills, Age, Gender and Parental Mediation in Brazil”, en *new media & society*, 20, pp. 2411-2431, 2018.

ttp://doi.org/10.18239/jornadas_2021.34.50

Un resumen de: "Security information sharing in smart grids: Persisting security audits to the blockchain"

Sergio Chica Andrés Marín Florina Almenares Daniel Díaz
Departamento de Ingeniería Telemática
Universidad Carlos III de Madrid
{sergio.chica, andres.marin, florina.almenares, daniel.diaz}@uc3m.es
ORCIDiDs:0000-{0002-7208-8094, 0001-9350-0669, 0002-5232-2031, 0002-3323-6453}

David Arroyo
ITEFI / Consejo Superior de Investigaciones Científicas
david.arroyo@csic.es
ORCIDiD: 0000-0001-8894-9779

***Resumen*—En la última década ha habido un aumento considerable en el número de empresas eléctricas que han decidido evolucionar hacia una arquitectura más inteligente, denominadas redes eléctricas inteligentes o *smart grids*. A su vez, ha surgido una dependencia de las redes de datos para el transporte de información o el envío de instrucciones con el fin de optimizar las operaciones de las redes inteligentes. Debido a este crecimiento en el uso de operaciones en red, es de vital importancia realizar auditorías periódicas para mantenerlas seguras. En este trabajo, proponemos** AUTOAUDITOR**, una herramienta que permite la automatización de las auditorías.** AUTOAUDITOR **se ha diseñado teniendo en mente las necesidades de las empresas eléctricas, como la escalabilidad frente a la heterogeneidad de dispositivos que componen su red. De la misma manera que las auditorías son una componente esencial, también lo es el intercambio de información. Nuestra propuesta es el uso de una *blockchain* permisionada para el almacenamiento de los informes de las auditorías, permitiendo establecer un control de acceso a la información, así como demostrar la existencia de dicha información durante un incidente.**

***Index Terms*—Auditoría, Blockchain Permisionada, Red Inteligente, SmartGrid, Metasploit, Escalabilidad**

Tipo de contribución: *Investigación ya publicada.* ***Security Information Sharing in Smart Grids: Persisting Security Audits to the Blockchain. Electronics MDPI [1].***

I. Introducción

Las redes inteligentes son la evolución de las redes eléctricas actuales, destacan por promover una arquitectura descentralizada donde productoras, distribuidoras, comercializadoras y consumidores interactúan a través de la red para alertar de averías y proporcionar datos de consumo en tiempo real consiguiendo acomodar y optimizar la generación, distribución y consumo. Disponen de un amplio número de sensores a lo largo de la infraestructura para su evaluación, manteniéndola en condiciones óptimas y permitiendo que se adapte a las necesidades de los usuarios en todo momento. Esta monitorización en tiempo real permite un ahorro energético, lo que se traduce en una reducción de costes, y una mejora en la usabilidad y transparencia de la red.

En los últimos años ha habido un incremento en el número de países que han decidido mejorar su infraestructura de red eléctrica. Más de la mitad de los países miembros de la UE [2] han alcanzado un 10 % del despliegue de contadores inteligentes. Siete países van en una fase muy avanzada o ya han finalizado: Dinamarca (80 %, 2017), Estonia (98 %, 2017), Finlandia (100 %, 2013), Italia (95 %, 2011), Malta (85 %, 2014), España (100 %, 2018) y Suecia (100 %, 2009).

Se prevee que además se incorporen a la red generadores domésticos, dispositivos de protección inteligentes, dispositivos de almacenamiento, e incluso vehículos eléctricos.

Las redes inteligentes proporcionan beneficios a las empresas eléctricas como un aumento de la eficiencia operativa, la prevención, mantenimiento y control de reparaciones, o la optimización de gastos. Los clientes también se benefician de una mayor eficiencia energética lo que supone una disminución de costes, reducción en el número y duración de interrupciones del servicio, o acceso a nuevos tipos de energías renovables. De la misma manera, también se incrementan los riesgos: cada elemento que se añade a la red supone un punto de fallo. En este trabajo abordamos la protección de infraestructuras críticas frente a ciberataques, en concreto en las auditorías a los dispositivos que componen la red, también conocidas como análisis de vulnerabilidades o *pentesting*.

Las infraestructuras críticas necesitan llevar a cabo evaluaciones automatizadas de seguridad, tal y como se describe en el NISTIR-8011 Volumen 3 y 4 [3], [4]. Esto se realiza mediante la integración de las auditorías en el sistema de inventario para llevar un control en tiempo real de los componentes de la organización y su nivel de riesgo.

En [5] describimos AUTOAUDITOR, una herramienta capaz de realizar auditorías de manera automática o semiautomática e integrarse en el sistema de inventario de las empresas eléctricas. Estudios previos [6] han demostrado como las características de la blockchain se adaptan a los requisitos del sector energético. Esta comunicación es un resumen de [1], donde proponemos el uso de una blockchain permisionada para almacenar los informes. La descentralización que proporciona la blockchain es una característica fundamental en el principio del menor privilegio durante la cadena de custodia de evidencias. La inmutabilidad innata de la blockchain garantiza la integridad de las auditorías. Finalmente, una blockchain permisionada nos posibilita tener un control de acceso a los informes.

II. AUTOAUDITOR

El flujo de trabajo de AUTOAUDITOR empieza mediante un análisis de los dispositivos con CVEScanner [7], generando un informe con las posibles vulnerabilidades y módulos de metasploit relacionados. Se confecciona un plan de ataque

mediante AUTOAUDITOR, el cual se deberá revisar y modificar según sea necesario por el supervisor encargado. Una vez se ha aprobado el plan de ataque, AUTOAUDITOR crea una subred donde desplegará los contenedores de metasploit y un cliente VPN que establecerá una conexión segura hacia la red vulnerable donde se realizará la auditoría. Una vez se ha completado la auditoría, se generará un informe que se almacenará en la blockchain. Para información más detallada y acceso al código fuente, consultar el artículo original [1]. AUTOAUDITOR interactúa con la blockchain mediante un contrato inteligente que pone a disposición una serie de métodos para almacenar, consultar o eliminar informes:

- **NewReport():** Permite almacenar informes en la blockchain. Hace uso de *transient maps* para no dejar evidencia de los datos de entrada en el historial de transacciones.
- **GetReportById():** Permite consultar informes usando un identificador.
- **GetReportsByOrganization():** Permite consultar todos los informes de una organización.
- **GetReportsByDate():** Permite consultar todos los informes en una fecha.
- **GetReportHash():** Permite comprobar la integridad de un informe.
- **DeleteReport():** Marca un informe como eliminado, de manera que no aparezca en las consultas. No se elimina ninguna evidencia de la blockchain.

Existen dos bases de datos en la blockchain para almacenar informes:

- Base de datos A: Almacena información básica, como el año y mes del informe, y el número de máquinas afectadas por cada vulnerabilidad.
- Base de datos B: Almacena información más detallada, como una marca de tiempo precisa, los módulos de metasploit usados o las direcciones de red de las máquinas analizadas.

Ambas bases de datos almacenan información común: número de vulnerabilidades, vulnerabilidades analizadas y puntación.

III. Evaluación

Para la evaluación del sistema se ha instanciado un entorno virtual donde se simula una red con equipos vulnerables, un servidor VPN que permite la conexión con la red vulnerable, una red de Hyperledger Fabric (HLF) y la aplicación que lleva a cabo la auditoría.

Tabla I
COMPARTIVA DE LA MEDIA Y LA DESVIACIÓN ESTÁNDAR PARA CADA EXPERIMENTO.

Operation	Value	Reports			
		1	100	500	1000
GetReportById	μ	0.665	-	-	-
	σ	0.069	-	-	-
GetReportsByOrganization	μ	-	0.716	0.872	1.114
	σ	-	0.105	0.122	0.169
GetReportsByDate	μ	-	0.701	0.886	1.083
	σ	-	0.087	0.123	0.126
NewReport	μ	0.458	-	-	-
	σ	0.022	-	-	-

La Tabla I recoge los tiempos medios y la desviación estándar durante el almacenamiento (`NewReport`) y consulta (`GetReportsBy`) de informes en la blockchain. El tiempo de almacenamiento es ligeramente inferior al de consulta. A su vez, se observa como los tiempos de consulta son muy similares entre sí, incluso cuando existe una gran diferencia en el número de informes. La ligera mejora en los tiempos de consultas por fecha se debe a la indexación mediante enteros, en contraposición a las consultas por organización que usan cadenas de texto.

IV. Conclusiones y Trabajos Futuros

El intercambio de información relacionado con las auditorías de seguridad puede resultar en beneficios para todas las partes. Nuestro objetivo no es el de sustituir las herramientas actuales, sino el de ofrecer una nueva fuente de información y una plataforma de colaboración en la inteligencia de ciberamenazas. AUTOAUDITOR facilita lograr una monitorización continua al integrarse de manera semiautomática con el sistema de inventario de las empresas eléctricas. Los resultados de las auditorías se almacenan en una blockchain permisionada, consiguiendo así la inmutabilidad innata de la blockchain y la posibilidad de restringir el acceso a dicha información. En trabajos futuros se prevén definir unas políticas de control de acceso más detalladas, aprovechando mejor las herramientas a nuestra disposición, como los canales de HLF y definiendo listas de control de acceso más precisas en las *chaincodes* asociadas con AUTOAUDITOR. De la misma manera, se mejorará la metodología de intercambio de información, consiguiendo facilitar el flujo de trabajo durante un incidente de seguridad.

Agradecimientos

Este trabajo ha sido financiado por la Comunidad de Madrid (España) dentro del proyecto CYNAMON (P2018/TCS-4566), por el proyecto LINKA20216 ("Advancing in cybersecurity technologies", i-LINK+ program) del Consejo Superior de Investigaciones Científicas (CSIC), por los proyectos TEC2017-84197-C4-1-R y TIN2017-84844-C2-1-R del Gobierno Español con apoyo de fondos FSE y FEDER de la Unión Europea.

Referencias

[1] A. Marín, S. Chica, D. Arroyo, F. Almenares, D. Díaz, "Security information sharing in smart grids: Persisting security audits to the blockchain," *Electronics*, vol. 9, no. 11, p. 1865, 2020, doi: 10.3390/electronics9111865.

[2] C. Alaton, F. Tounquet, "Benchmarking smart metering deployment in the EU-28," Directorate-General for Energy (European Commission) and Tractebel Impact, Final report, 2020, doi: 10.2833/492070.

[3] K. Dempsey, P. Eavy, N. Goren, G. Moore, "Automation Support for Security Control Assessments: Software Vulnerability Management," NIST, NISTIR 8011, vol. 3, 2018, doi: 10.6028/NIST.IR.8011-3.

[4] K. Dempsey, P. Eavy, E. Takamura, G. Moore, "Automation Support for Security Control Assessments: Software Vulnerability Management," NIST, NISTIR 8011, vol. 4, 2020, doi: 10.6028/NIST.IR.8011-4.

[5] S. Chica, A. Marín, D. Díaz, F. Almenares, "On the Automation of Auditing in Power Grid Companies," in *Intelligent Environments 2020. Workshop Proc. 16th Int. Conf. Intelligent Environments*, vol. 28, 2020, pp. 331-340, doi: 10.3233/AISE200057.

[6] M. Andoni et al., "Blockchain technology in the energy sector: A systematic review of challenges and opportunities," *Renewable and Sustainable Energy Reviews*, vol. 100, pp. 143-174, 2019, doi: 10.1016/j.rser.2018.10.014.

[7] F. Alegre, "CVEscanner: NMAP tool to discover potential vulnerabilities in services with open ports," 2020. Available: https://github.com/alegr3/CVEscanner

http://doi.org/10.18239/jornadas_2021.34.51

COBRA: Cibermaniobras adaptativas y personalizables de simulación hiperrealista de APTs y entrenamiento en ciberdefensa usando gamificación

Félix Gómez Mármol[1], José A. Ruipérez-Valiente[1], Pantaleone Nespoli[1], Gregorio Martínez Pérez[1], Diego Rivera Pinto[2], Xavier Larriva Novo[2], Manuel Álvarez-Campana[2], Víctor Villagrá González[2], Jorge Maestre Vidal[3], Francisco A. Rodriguez López[3], Miguel Páramo Castrillo[3], Javier I. Rojo Lacal[3], Ramón García-Abril Alonso[4]

[1]Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, 30100, Murcia, España
{felixgm, jruiperez, pantaleone.nespoli, gregorio}@um.es

[2]ETSI Telecomunicación, Universidad Politécnica de Madrid, España
{diego.rivera, xavier.larriva.novo, manuel.alvarez-campana, victor.villagra}@upm.es

[3]Indra, Avenida de Bruselas, 35, Alcobendas, 28108 Madrid, España
{jmaestre, farodriguez, mparamo, jirojo}@indra.es

[4]Mando Conjunto del Ciber Espacio, Base de Retamares, 28223 Pozuelo de Alarcón, España
rgaralo@et.mde.es

***Resumen*—Una formación en ciberdefensa de alta calidad que permita adquirir competencias que luego sean aplicables en escenarios reales es altamente compleja. A pesar de que la mayoría de las organizaciones y cuerpos involucrados en este área están de acuerdo en afirmar que generar mecanismos para el desarrollo de estas capacidades es prioritario, aún existen importantes carencias a nivel de metodologías y competencias, así como de sistemas y entornos de entrenamiento. En este sentido, el proyecto COBRA de "Cibermaniobras adaptativas y personalizables de simulación hiperrealista de APTs y entrenamiento en ciberdefensa usando gamificación" es ambicioso en combinar diversas tecnologías para alcanzar este objetivo, teniendo intrínsecamente un carácter multidisciplinar pero con unas metas claras.**

***Index Terms*—Cyber Range, Ciberseguridad, Simulación APTs, Gamificación**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

Tal y como han anunciado una y otra vez las distintas organizaciones para la ciberdefensa, el éxito de cualquier iniciativa para su salvaguarda depende de la combinación adecuada de doctrina, organización, formación, procedimientos y comportamientos, así como de la disponibilidad de productos adecuados (infraestructuras y software).

Pero, a pesar de los esfuerzos de la Unión Europea (UE) y sus estados miembro hacia la disposición de habilitadores de formación y educación para operaciones en el ciberespacio, la revisión del estado del arte [1], [2] y la evaluación de las diferentes soluciones comerciales actuales sugiere la existencia de una lista emergente de desafíos y brechas tecnológicas a cubrir de cara a mejorar la capacidad de toma de decisiones en el ciberespacio [3], [4], destacando entre ellas:

- Las soluciones existentes a menudo tienen dificultades a la hora de proporcionar acceso a la formación bajo demanda, a través de infraestructuras de formación reutilizables, escalables y adaptables, permitiendo así la reducción de los costes de los equipos específicos de formación.
- Las soluciones existentes no suelen proveer la capacidad de realizar ejercicios de formación especializada adaptada a las Especificidades del Dominio de Usuario (UDS) en el ámbito de las operaciones en el ciberespacio, como sus capacidades Tecnológicas Operativas (OTS) o los dispositivos conectados al Internet de las Cosas (IoT) portados por los diferentes efectivos.
- El estado del arte presenta marcadas carencias a la hora de adaptar los procesos de operación a amenazas específicas con impacto en diferentes niveles del entorno operacional [5]. Esto incluye la dificultad de adiestrar la capacidad de toma de decisiones de los ciber comandos ante dichas amenazas.
- La tendencia a la virtualización y la construcción de gemelos digitales sugieren una incipiente demanda de escenarios para cibermaniobras con la capacidad de representar entornos operacionales mucho más complejos, permitiendo a los ciber comandos ejercitarse con un mayor realismo.

Desde un punto de vista educacional, uno de los grandes problemas de los actuales sistemas es la pérdida de motivación y sensación de aburrimiento por parte de los estudiantes. Esto puede suceder por diversas razones, como bajo interés por la materia que se está recibiendo, sensación de apatía ante los contenidos porque no interesen o sean muy fáciles o, por el contrario, la sensación de incapacidad de entender contenidos o completar ejercicios debido a su alta dificultad, con la consiguiente frustración.

Numerosos estudios han demostrado que cuando los alum-

nos se encuentran más motivados por su proceso de aprendizaje, los resultados finales mejoran de forma significativa. Todos estos problemas son tratados por el proyecto COBRA en el contexto específico de práctica, aprendizaje y entrenamiento en materias de ciberdefensa en un Cyber Range, a través de la implementación de cibermaniobras dinámicas y adaptativas al estudiante, en oposición a las estáticas actuales. Por lo tanto, nos alejamos de la idea de "one-size-fits-all" y nos adentramos en la dirección de una educación personalizada a las necesidades y estado del estudiante.

Por otra parte, la introducción de elementos de gamificación en el sistema también puede tener un efecto positivo en la motivación de los estudiantes [6], [7]. De este modo, mediante una educación más personalizada y haciendo uso de los distintos componentes de diseño que pueden mejorar la motivación de los estudiantes, esperamos que se puedan mejorar de forma significativa resultados previos de aprendizaje con el Cyber Range.

Así mismo, mediante el uso de las trazas de datos telemétricas y las señales biométricas generadas por los estudiantes mientras resuelven los escenarios planteados seremos capaces de evaluar las competencias clave en un entorno militar, no solo relacionadas con contenidos en ciberseguridad, sino también las habilidades transversales como capacidad de trabajo bajo presión.

Por último, y en lo que respecta a las capacidades de defensa (a saber, Defensa, Explotación y Respuesta) en el ámbito del Ministerio de Defensa español (derivadas de la OTAN), el proyecto COBRA permitirá desarrollar fundamentalmente la capacidad de Respuesta.

II. Objetivos

El objetivo principal del proyecto COBRA (Cibermaniobras adaptativas y personalizables de simulación hiperrealista de APTs y entrenamiento en ciberdefensa usando gamificación) se resume en:

> Desarrollo de un conjunto de herramientas para la simulación hiperrealista de Amenazas Persistente Avanzadas (APTs), orientada a la ejecución de cibermaniobras adaptativas y personalizables mejoradas con técnicas de gamificación.

Dicho objetivo se divide, a su vez, en cinco subobjetivos, que serán descritos con más detalle a continuación.

II-A. Simulación de topologías de red y tráfico real

Uno de los elementos necesarios para el desarrollo del proyecto es, a partir de una infraestructura virtualizada de red, permitir la generación de una serie de topologías de red simuladas que puedan servir de base para los escenarios de entrenamiento. En este sentido, se desarrollarán una serie de herramientas que permitan la definición flexible y parametrizada de topologías de red, y se integrarán estas herramientas en el Cyber Range utilizado como infraestructura base del proyecto. Esto implica por una parte ofrecer un sistema para la definición de estas topologías, y por otra, desarrollar un sistema capaz de, a partir de estas definiciones, desplegar y configurar de manera efectiva los nodos que componen la topología, comunicándose para ello con el sistema de virtualización del Cyber Range.

Por otro lado, una vez establecidas las topologías de red, y para ofrecer un escenario simulado completamente realista, es necesario generar tráfico de red que se asemeje al tráfico que pudiera darse en una red real. Este tráfico deberá modelarse adecuadamente para asegurar el realismo de la simulación, y posteriormente deberá inyectarse en la red simulada, utilizando para ello herramientas de generación e inyección de tráfico.

II-B. Simulación de amenazas avanzadas persistentes

Los ataques informáticos son una amenaza generalizada para cualquier sistema informático. Los ataques evolucionan al mismo tiempo que los sistemas. Es por ello que los entornos de formación y entrenamiento en ciberseguridad deben contar con sistemas que permitan conocer y practicar sobre las distintas tipologías de amenazas que existen actualmente y que están continuamente avanzando. Las amenazas más sofisticadas se integran en lo que se denominan "Amenazas Avanzadas Persistentes", del inglés Advanced Persistent Threats (APT), en las que los atacantes diseñan una estrategia de ataque, con múltiples etapas. Este objetivo propone la definición y diseño de Simulaciones de Amenazas Avanzadas Persistentes. Para ello, se partirá de un modelado formal de un ataque avanzado persistente APT genérico apoyándonos en modelos basados en estados (como, por ejemplo, en cadenas de Markov, STIX), de una forma realista y parametrizable siguiendo los pasos definidos en el modelo Unified Kill Chain (evolución de la Cyber Kill Chain en base al marco ATT&CK del MITRE). Estos modelos serán diseños propios que permitan acoplarse en su aplicación a los escenarios aleatorios y parametrizables. Para la simulación de APT se tomará en cuenta los patrones y herramientas de ataque de las propias amenazas con el objetivo de brindar un escenario lo más realista posible. Posteriormente se definirá una plataforma de simulación de los distintos estados y la evolución de eventos discretos temporales, para su aplicación en herramientas de simulación de APT en un Cyber Range.

II-C. Escenarios aleatorios y parametrizables

Otro objetivo de este proyecto tiene que ver con añadir funcionalidades a los sistemas Cyber Range en cuanto a la posibilidad de generar escenarios simulados aleatorios y parametrizables. Esto permitirá una mayor facilidad y flexibilidad a la hora de definir e instanciar ciberejercicios. Esta funcionalidad supondrá una mejora con respecto a la utilización del sistema por parte de sus usuarios, tanto instructores como estudiantes. Desde el punto de vista de los instructores, la generación de plantillas de escenarios y la definición de ciberejercicios será mucho más libre y automatizada. Desde el punto de vista de los estudiantes, la aleatorización y la posibilidad de contar con escenarios dinámicos permite una experiencia de aprendizaje mucho más rica, al proporcionarse una gran cantidad de posibles desafíos y una adecuación a las capacidades de cada uno.

Para la consecución de este objetivo se desarrollarán herramientas que permitan comunicarse con los sistemas subyacentes de definición de topologías de red, de generación de tráfico y de simulación de amenazas avanzadas persistentes,

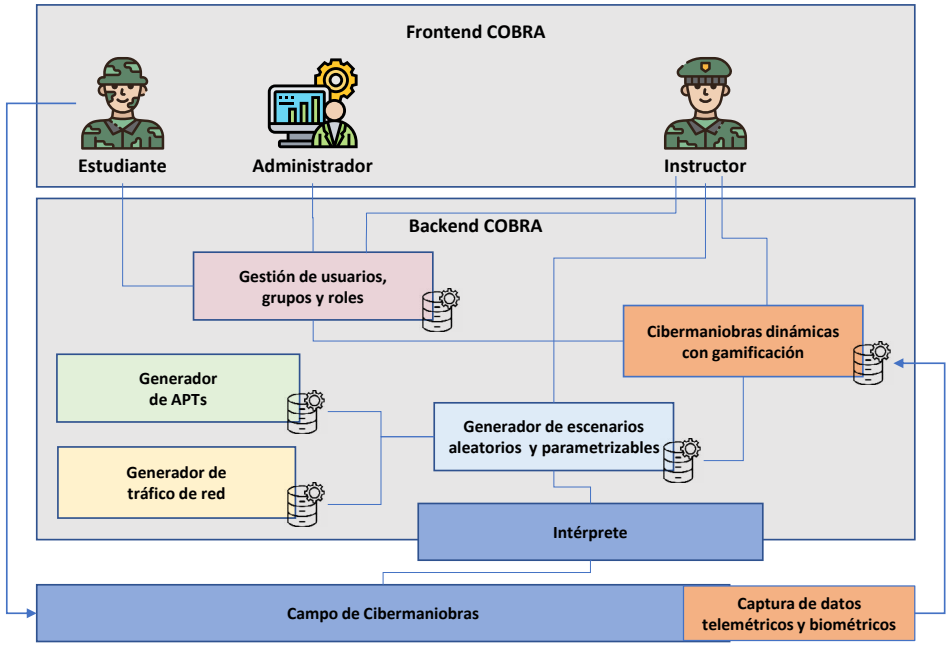


Figura 1. Arquitectura preliminar del proyecto COBRA

para transformar la definición de escenarios y plantillas en una infraestructura real sobre la que realizar los ciberejercicios. Contarán además con interfaces gráficas de usuario que permitan a los instructores realizar dichas definiciones con facilidad.

II-D. Cibermaniobras adaptativas con gamificación

A partir de los sistemas desarrollados en el proyecto para la flexibilización y la definición aleatoria y parametrizada de diversos escenarios de red sobre los que simular amenazas, en este proyecto se tiene como objetivo la incorporación de técnicas de Inteligencia Artificial con aprendizaje adaptativo para que los escenarios puedan adaptarse de forma específica a cada estudiante.

Para ello se proporcionarán sistemas de telemetría que permitan recoger información acerca del desempeño de los usuarios en los ejercicios planteados, así como diversos datos biométricos sobre el uso del sistema. Esto, junto con la aplicación de técnicas de gamificación, permitirá la adaptación de los escenarios de acuerdo a las capacidades, el desempeño y el estrés de los estudiantes.

II-E. Validación en el Cyber Range del MCCE

El quinto subobjetivo del proyecto tiene que ver con la instalación e integración de los sistemas desarrollados en la infraestructura de Cyber Range desplegada por el Mando Conjunto del Ciber Espacio (MCCE). Mediante la integración de los desarrollos se pretende evaluar la viabilidad de las soluciones definidas en el proyecto, así como determinar las competencias de los estudiantes mediante la realización de diversas pruebas diseñadas específicamente para la validación de la propuesta del proyecto.

El objetivo final del proyecto por tanto sería la generación de un demostrador operativo, instalado e integrado en el Cyber Range del MCCE, que permita la definición de entornos relevantes y cibermaniobras realistas.

III. Estado actual de COBRA

El proyecto COBRA, en el que participan la Universidad de Murcia, la Universidad Politécnica de Madrid y la empresa tecnológica Indra, comenzó el 1 de diciembre de 2020 y tiene una duración total de 24 meses. En el momento actual se han desarrollado ya una serie de trabajos iniciales entre los que destacan los que se describen a continuación.

Se ha elaborado un Plan de Gestión del Proyecto (PGP) donde se especifica toda la metodología de gestión del proyecto COBRA, la división del mismo en paquetes de trabajo y tareas, con la consiguiente planificación temporal de los mismos. También se ha incluido un Plan de Gestión de Riesgos (PGR), así como un Plan de Gestión de la Calidad (PGA) y un Plan de Gestión de la Configuración (PGC). Se han identificado además todos los entregables, tanto de tipo documentación como software, que se irán desarrollando a lo largo de la vida del proyecto, agrupados por hitos o Plazos Parciales (PP).

Por otra parte se ha diseñado un Plan de Validación y Verificación (PVV) incluyendo la metodología a seguir para la evaluación sistemática de la consecución de los objetivos planteados. También se han identificado y descrito un total de 41 requisitos funcionales y no funcionales asociados a los 5 principales objetivos del proyecto, junto con su método de verificación y su prioridad. Además, se ha trabajado para obtener una arquitectura de alto nivel del proyecto, identificando cada uno de sus principales componentes, así como las principales interconexiones e interfaces entre los mismos (véase la Figura 1). Para cada uno de dichos componentes se ha proporcionado una descripción o principios de diseño, unas premisas operacionales y unas limitaciones.

Por ultimo, cabe también destacar el comienzo de los trabajos técnicos asociados a los objetivos II-A, II-B y II-C, respectivamente. En este sentido, y en lo que respecta al objetivo II-A, se ha llevado a cabo un profundo análisis del estado del arte en cuanto a herramientas de simulación de

tráfico de red, clasificadas en función del tipo de generación de tráfico aplicada: i) Generación basada en la replicación de tráfico, ii) Generación de tráfico sintético mediante creación de paquetes, iii) Generación de tráfico basada en modelos, iv) Generación de tráfico de alto nivel y auto configurables, y v) Generación de tráfico para escenarios específicos. En esta revisión exhaustiva de la literatura se han analizado un total de 46 herramientas distintas.

En cuanto al objetivo II-B de simulación de APTs, también se ha realizado una revisión detallada del estado del arte sobre modelado y simulación de amenazas avanzadas persistentes. En este sentido, se han analizado diversas propuestas de modelado de APTs (como por ejemplo STIX), se han estudiado en detalle varios métodos de modelado de ataques multi-paso (basados en correlación, en similitud, en estructura, etc.) y se han descrito los principales pasos que componen la Cyber Kill Chain.

Y en lo que respecta al objetivo II-C de COBRA, se ha realizado otro análisis exhaustivo del estado del arte en cuanto a plataformas Cyber Range se refiere. En concreto se han estudiado los distintos tipos de Cyber Ranges, así como los posibles dominios de aplicación. También se han analizado los diferentes equipos que pueden participar en un Cyber Range (equipo rojo, equipo azul, etc.), además de las principales tipologías de ataques (explotación de vulnerabilidades, ataque a protocolos, ataques de ingeniería social, etc.) y de defensas (prevención, detección, reacción o análisis forense digital) que se pueden encontrar más comúnmente en un Cyber Range.

Con respecto al objetivo II-D, ya se han empezado a analizar las distintas arquitecturas que permitan recolectar telemetría por parte del Cyber Range y señales biométricas por parte de los usuarios que se encuentren haciendo las cibermaniobras. Se ha analizado la viabilidad de distintos dispositivos para poder recolectar dichas señales biométricas. Como próximos pasos, nos encontramos analizando los posibles indicadores de desempeño que se pueden computar, así como aquellas competencias que pueden ser valiosas de evaluar con dichos datos dentro del contexto de ciberseguridad de que deben de afrontar dichos usuarios. Por último, nos encontramos en proceso de análisis de los algoritmos de inferencia de conocimiento y aprendizaje adaptativo a aplicar en este contexto.

IV. Conclusiones y trabajo futuro

A pesar de los grandes esfuerzos hacia consolidar un marco europeo para la adecuada educación y entrenamiento en materia de ciberseguridad, a día de hoy existen importantes carencias marcadas entre otras por: 1) fallos a la hora de equipar a los profesionales de los conocimientos, competencias y aptitudes necesarios para prevenir y responder a ciberamenazas reales; 2) el acceso limitado a capacidades de análisis coste-impacto basados en pruebas; 3) dificultades a la hora de comprender la naturaleza interdisciplinar de la ciberseguridad; y 4) la necesidad de entornos de formación colaborativos y coherentes con el espacio de operaciones real.

Esta problemática se extiende al sector defensa, donde además de las restricciones inherentes a sus líneas de desarrollo (doctrina, organización, liderazgo, etc.), es necesario formar en la intersección entre las aptitudes necesarias para operar en escenarios militares, y las características de técnicas y sociocognitivas presentes en el ciberespacio. Alcanzar estas capacidades implica avanzar hacia entorno de formación colaborativos centrados en el usuario, los cuales han de ser capaces de generar dinámicamente patrones de tráfico de red y comportamientos artificiales (neutros, aliados, hostiles) suficientemente creíbles.

En este contexto, el proyecto "COBRA: Cibermaniobras adaptativas y personalizables de simulación hiperrealista de APTs y entrenamiento en ciberdefensa usando gamificación" pretende desarrollar una solución innovadora que dé respuesta a estas deficiencias desde cinco objetivos principales, a saber: i) la simulación de topologías de red y tráfico real, ii) la simulación hiperrealista de amenazas avanzadas persistentes (APTs), iii) el desarrollo de escenarios aleatorios y parametrizables, iv) el desarrollo de cibermaniobras adaptativas utilizando gamifiación, y v) la validación de toda la propuesta en el Cyber Range del Mando Conjunto del Ciber Espacio (MCCE) del Ministerio de Defensa de España.

Este ambicioso proyecto, que aún se encuentra en su primer año de vida, aún tiene bastante recorrido por delante hasta alcanzar todas las metas propuestas. Y cuando esto ocurra, se convertirá en un referente en su campo, dados los novedosos avances que proporcionará, y que hasta donde los autores conocen, no existen en ninguna otra solución actual similar.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto COBRA (10032/20/0035/00), concedido por el Ministerio de Defensa, así como por las ayudas FJCI-2017-34926 y RYC-2015-18210, concedidas por el Gobierno de España y confinanciadas por el Fondo Social Europeo.

Referencias

[1] I. Priyadarshini, "Features and architecture of the modern cyber range: A qualitative analysis and survey," Ph.D. dissertation, University of Delaware, 09 2018.

[2] E. Ukwandu, M. A. B. Farah, H. Hindy, D. Brosset, D. Kavallieros, R. Atkinson, C. Tachtatzis, M. Bures, I. Andonovic, and X. Bellekens, "A review of cyber-ranges and test-beds: Current and future trends," *Sensors*, vol. 20, no. 24, 2020.

[3] E. C. S. O. (ECSO), "Gaps in european cyber education and professional training," 2017, position paper of WG5 for Education, training, awareness, cyber ranges. [Online]. Available: https://ecs-org.eu/documents/publications/5fdb282a4dcbd.pdf

[4] ——, "Understanding cyber ranges: From hype to reality," 2020, position paper of SWG 5.1 for Cyber Range Environments and Technical Exercises. [Online]. Available: https://www.ecs-org.eu/documents/uploads/understanding-cyber-ranges-from-hype-to-reality.pdf

[5] R. Daton Medenou, V. Calzado Mayo, M. Garcia Balufo, M. Páramo Castrillo, F. González Garrido, A. Luis Martinez, D. Nevado Catalán, D. Hu, A. Sandoval Rodríguez-Bermejo, J. Maestre Vidal, G. Ramis Pasqual De Riquelme, A. Berardi, P. De Santis, F. Torelli, and S. Llopis Sanchez, "CYSAS-S3: a novel dataset for validating cyber situational awareness related tools for supporting military operations," in *Proceedings of the 15th International Conference on Availability, Reliability and Security (ARES)*, Dublin, Ireland, August 2020, pp. 1–9.

[6] F. F.-H. Nah, Q. Zeng, V. R. Telaprolu, A. P. Ayyappa, and B. Eschenbrenner, "Gamification of education: A review of literature," in *HCI in Business*. Cham: Springer International Publishing, 2014, pp. 401–409.

[7] P. Blikstein and M. Worsley, "Multimodal learning analytics and education data mining: using computational technologies to measure complex learning tasks," *Journal of Learning Analytics*, vol. 3, no. 2, pp. 220–238, 2016.

Sesión de Investigación B3: Tecnologías emergentes y entrenamiento en ciberseguridad
http://doi.org/10.18239/jornadas_2021.34.52

Exploring the Affordances of Multimodal Data to Improve Cybersecurity Training with Cyber Range Environments

Mariano Albaladejo-González, Sofia Strukova, José A. Ruipérez-Valiente, Félix Gómez Mármol
Department of Information and Communications Engineering, University of Murcia
Calle Campus Universitario, 30100 Murcia (Spain)
{mariano.albaladejog, strukovas, jruiperez, felixgm}@um.es

***Abstract*—During the last years, the constant cybersecurity breaches being reported are remarking the necessity of raising the number of cybersecurity experts that can tackle such threats. In this sense, educational technology environments can help to generate more immersive and realistic environments, and within this context, cyber range systems are one of the foremost solutions. However, these systems might not provide rich and detailed feedback to instructors and students regarding the performance in each cyberexercise. In this paper we discuss the potential of multimodal data, including clickstream, console commands, biometrics, and other sensor data, to improve the feedback and evaluation process in cyber range environments. We present the affordances that these techniques can bring to cybersecurity training as well as a preliminary architecture to implement them. We argue that these technologies can become a new generation of high-quality, realistic, and adaptive cybersecurity training that can have a dual (civil and military) impact on our society.**

***Index Terms*—Cyber range, cybersecurity training, multimodal learning analytics, educational technology.**

Tipo de contribución: *Investigación en desarrollo*

I. Introduction

The last decade has made exceptionally clear the upmost necessity of growing the number and quality of cybersecurity experts that can design secure systems and respond to potential threats. Every week we hear of new security breaches and scandals, that jeopardize entire companies and the privacy of their users. The respondents of the ISACA's State of Cybersecurity of 2020 indicated that 53% of them were expecting a cyberattack within 12 months. Moreover, Cybersecury Ventures predicted that cybercrime will produce damages totaling $6 million USD globally in 2021, a prediction which is based on recent year-over-year growth [1]. To face this problem, there is an overall agreement on the need to increase the quality of the training that these specialists receive [2]. However, a research report that interviewed over 300 cybersecurity professionals indicated that only 38% of them are happy with the level of training that they are receiving [3].

In this sense, educational technology training tools can play a pivotal role in the training quality that professionals can receive. Within this context, we are especially focused on cyber ranges, which are well-defined virtualized environments where trainees can develop practical hands-on-activities that resemble much better real cybersecurity operations. There are a good number of prominent cyber range examples in the literature [4], [5], [6]. These can represent realistic cybersecurity scenarios in safe sandbox environments where the trainees can attempt to attack a network system (red team) or defend a system against an adversary attack (blue team). These cyberexercises resemble much better the real world situations that these professionals will need to face when an actual threat emerges.

However, one of the handicaps that the current state of the art of these environments shows is a low emphasis on performing effective automatic evaluations and feedback provision based on the trainee performance in the cyberexercise. For example, a recent literature review on cyber range environments that inspected all the existing ones until today, only mentioned that the evaluation can be either done manually (with human intervention) or automatically (based on an algorithm and key variables of the cyberexercise) [7]. The majority of cyber ranges provide very limited feedback on the process that the trainee followed to solve or fail the cyberexercise. For example, a capture-the-flag cyberexercise where an attacker needs to gain admin privileges and access a hidden code [8], might provide as only feedback to the instructor that the trainee knows said hidden code. Therefore, instructors cannot provide detailed and adapted feedback, nor perform a rich evaluation of the trainee taking into account diverse factors and actions that happened during the learning process.

To face this ambitious challenge, in this paper we argue on the potential of using multimodal data to improve such evaluation within the context of cyber ranges. To do so, we collect data from multiple sources, including clickstream data, console commands, biometrics and other sensor data. Then, we apply multimodal learning analytics conducting signal processing and artificial intelligence to transform the raw multimodal data into rich information [9]. In the paper at hand, we present our current advances regarding how these multimodal data can be used to improve the evaluation and feedback of trainees in cyber range environments. More specifically, we have the following two objectives:

- To present the affordances of multimodal data to improve the training process in cyber range environments.
- To propose a preliminary architecture adapted to this specific scenario to accomplish such goal.

The remainder of this paper is organized as follows: In Section II we present an overview of the affordances of multimodal data in cyber range environments, while in Section III we discuss our preliminary architecture. We finalize the paper in Section IV with conclusions and future research lines.

II. Multimodal Data in Cyber Range Environments

The essential feature of a cyber range is the development of isolated and safe environments. For this reason, the core of a cyber range is virtualization, simulation and/or containerization technologies that support these environments. In addition to these technologies, a cyber range might also include front-end technologies to provide easy access to such environments. Their architecture isolates the users' computers and external networks from the environments that are running malware [10]. We find that in cyber range platforms, it is rare to find the presence of front-end dashboards that can monitor the results of the cyberexercises. The few cyber ranges that offer a dashboard show shallow information, measuring whether the user has completed the exercise successfully and the time required to do so. Our system aims to expand this state of the art with new measures of user skills and performance when interacting with cyber ranges.

The system requires data that can come from different sources to generate these news measures. For example, the cyber range can generate data related to the users' solution, along with the number of attempts, the typed commands, the proportion of unnecessary commands, and the quality of the solutions. Furthermore, it is easy to collect data related to user's telemetry adding keyboard and mouse monitoring tools in the front-end technologies, this is a common practice in websites and apps for real-time and asynchronous tracking [11]. These telemetry data can provide the following information:

- **Keyboard patterns**. These data are generated when the user writes commands. It includes the typing speed and the keystroke duration.
- **Clickstream**. It represents how the user interacts with the graphic interface of the environment. The clickstream includes the clicked elements, the click frequency, the click duration, and the mouse movement speed.

Our system aims to go further, including data collected by sensors and devices external to the cyber range. A camera and/or a kinetic device can get many interesting measures such as eye-tracking, the users' pose, position, and expression [12], [13]. Furthermore, we can add microphones to record the communication between the users [14].

In addition, we propose to measure physiological signals to get richer information about the users' state during the cyberexercises. Depending on the original context for which they are used, there are three types of devices to measure physiological signals: the devices used in the medical field for diagnosis purposes, the devices used for research purposes, and the commercial devices focused on the daily use of end users. Additionally, these devices can be placed in different parts of the body: for example, we can have wristbands, chest straps, and brain-computer interfaces (BCIs) that are placed as a helmet. BCIs measure the electrical activity of the brain and can estimate the emotions and moods of the users during the cyberexercises. The wristbands and chest straps can have different types of sensors to measure the heart rate, the blood pressure, the skin temperature, the oxygen saturation, the electrodermal activity (the measurement of the electrical activity of the skin), and the movement of the user measured through the accelerometers and gyroscopes, among others.

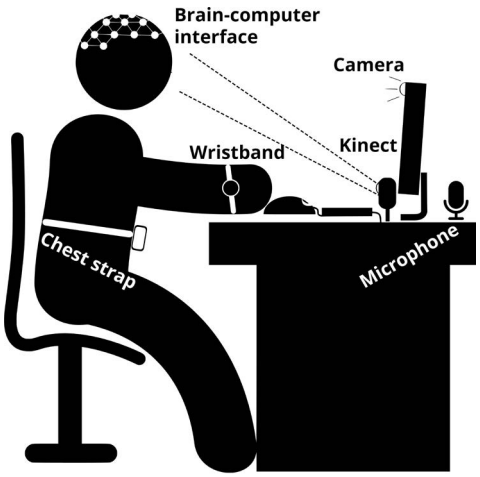


Fig. 1. Potential devices used to monitor the user during the cyberexercises.

Figure 1 represents different devices collecting data throughout the cyberexercises. The system uses all the collected data to calculate additional user performance metrics and skills. The users' emotions during the cyberexercises can be estimated and classified depending on the valence and arousal degrees [15]. The valence represents the level of positive or negative affectivity and arousal, the calming or excitement level. Thereby, we could infer states like anger, joy, sadness, and pleasure.

In addition to the user emotions, the system could measure more advanced skills closely related to the necessities of cybersecurity professionals. In real-world environments, cybersecurity professionals can be under much pressure due to the impact of their decisions; for example, failing to detect sniffers on an online shop can end up causing a data breach of 40 million card numbers and 70 million personal records stolen [16]. For this reason, it is interesting to evaluate the capacity to work under pressure, for example, through user stress or the attention level [17]. Moreover, cyberattacks and their consequences can take place over an extended period of time [18]; for this reason, it is also interesting to measure the user fatigue [19]. Teamwork skills are critical for cybersecurity professionals as they will often be part of a larger and multidisciplinary team. The proposed multimodal system can be used to evaluate teamwork skills and how teamwork affects each user. All of these metrics aim to empower instructors with additional information to provide a more nuanced feedback and assessment to the cybersecurity students. The final goal is to improve the readiness of the cybersecurity professionals to detect and resolve cybersecurity breaches.

To implement the proposed system is essential to consider how invasive the devices are and whether they can be used for extended periods of time. Devices that are too invasive might endanger data recollection by reducing user freedom of movement and making the cyber range experience more uncomfortable. Furthermore, it is important to consider the devices' cost because since some of them are quite expensive, and can be used only by a single user at a time. Microphones, cameras, kinetics, and wearables are affordable solutions with

potential to measure useful constructs.

Finally, all the aforementioned types of data generated by trainees while using cyber range environments hold the potential for being used to adjust the cyberexercises to the current status of each specific trainee. This process, which is known as *adaptive learning*, has the goal of addressing the unique needs of each user [20]. In our case, we aim to dynamically adapt current cyberexercises to the knowledge of the trainee. Commands and clickstream data along with the biometric signals of trainees will allow us to analyze and compare the results of the different simulations to progressively improve subsequent training sessions and, therefore, maintain the optimal balance between the trainees' knowledge and the difficulty of each exercise.

III. Preliminary Architecture

A. Description of the training process

Our system is used by: 1) the trainees, who interact with the learning contents and generate the raw data, and 2) the instructors, who are experienced teachers in the cybersecurity field responsible for keeping track of how the trainees are progressing and providing them with relevant feedback. Accordingly, the instructor first provides the trainees with the cyberexercises they must solve and afterward reviews the results represented in the dashboard in an easy and understandable way. This helps to build the feedback that the trainees will receive and choose the most suitable cyberexercises for the future users with similar knowledge.

The training process starts when the instructor distributes the cyberexercises across the trainees. While the latter are solving the tasks, our system collects various data types described in the previous section. Then, these data are processed and analyzed in order to visualize the dashboard with all the information about the cybersecurity development of each trainee.

B. Overview of the Architecture

Figure 2 presents the overview of the architecture of the cyber range environment with the multimodal learning analytics, and how the following components are connected within the system:

- **Cyber range**. The cyber range system is the origin of the learning process. When trainees interact in their cyber range environment, a large amount of raw multimodal data is generated, issued, collected, and stored in the webserver. We implement the event emission process using experience API schema (xAPI [1]) to make the rest of parts of the architecture agnostic of the specific cyber range system implemented.
- **Data collection**. The data collected within the cyber range include a wide variety of trainee actions, as well as the external sensors and devices. We use REpresentational State Transfer API (RESTful API) endpoints to send these data to the web server. There are several challenges regarding the ethical and security considerations of obtaining that data from the trainees. Thus, the collected personal data is encrypted and protected by applying appropriate technical and organizational measures according to the General Data Protection Regulation (GDPR).
- **Data processing and analytics**. This step aims to measure and evaluate the development of the trainees regarding their cybersecurity skills. An Extract, Transform, and Load (ETL) procedure is employed in order to extract the needed data from the database, transform them into a proper storage structure for querying and analysis, and finally load them into a final database. Due to its large size, the processing cannot be performed in real-time. Therefore, we make use of *cron jobs* to launch the data processing scripts at scheduled times.
- **Visualization dashboard**. The last step consists in providing useful and effective visualizations to both the trainees and the instructors. This is done through the dashboard that represents an activity and performance measurement interface. Specifically, we can see general statistics presenting the overall progress across the cyberexercises, or active time, to name some examples. We can also see more complex measurements such as the capacity to work under pressure and concentration level. Trainees can access only their own data, while instructors can access the information of each trainee individually or see the aggregation of the entire class. At the same time, we also develop models that can evaluate trainees' competencies based on which cyberexercises they have been able to complete.

[1] https://xapi.com/

IV. Conclusions and future work

Raising a new generation of cybersecurity professionals during the 21st century is vital to have a secure digitized world and economy. However, the specialized training of these professionals is a challenging task. Cyber range environments represent a great asset that complements more traditional cybersecurity training in order to practice hands-on cyberexercises that can resemble real scenarios where trainees need to attack and/or defend a system in real time. However, the current feedback that cyber ranges provide to instructors regarding the performance of their trainees is quite scarce. In some cases we find that the instructors do not know more than whether the cyberexercise was completed or not, with no information about the process at all. This approach is definitely not sufficient to provide a just-in-time support and feedback to the students in order to improve the learning process, specially when we want to scale cyber range case studies with entire classes getting trained simultaneously.

In this paper we have argued on the potential that multimodal data can have to improve the training process when using cyber ranges. We can collect different data in various modalities like clickstream, console commands, biometrics or audiovisual data, apply signal processing and artificial intelligence techniques, and produce measures to assess ideal solution pathways, capacity to work under pressure, or concentration, which are key capacities to become a successful cybersecurity professional. Moreover, these techniques can have a dual impact on our society. First, on the civil side, we can use them to improve the academic training of students under-taking degrees related to cybersecurity and also on professional programs training cybersecurity professionals. Second, on the military side, we can use the same approach

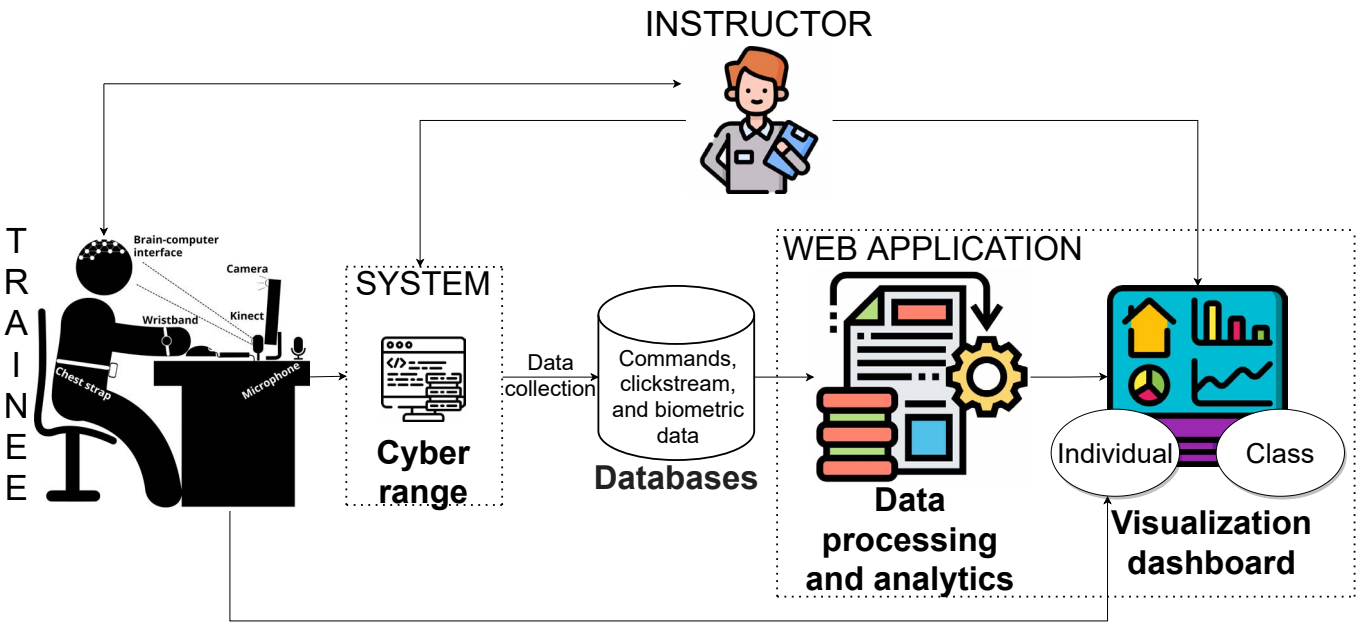


Fig. 2. Preliminary architecture of the cyber range environment with the multimodal learning analytics web application.

to improve the cyberdefence capabilities that a state may have to protect the cyberspace. Depending on the critical nature of the position of each professional getting trained, more or less invasive data collection approaches can be applied.

The future steps that we envision are multifaceted. First, we are working on developing this architecture as generic as possible, using different data sources and sensors. Then, we are planning to deploy several cyber ranges on controlled premises and make this architecture as inter-operable as possible. Then, we will conduct case studies with students undertaking security classes and with cybersecurity professionals in order to collect data and prove the viability of the architecture. Finally, we will validate that this approach is improving the overall training process.

Acknowledgments

This work has been partially funded by project COBRA (10032/20/0035/00), awarded by the Spanish Ministry of Defense, as well as the fellowships FJCI-2017-34926 and RYC-2015-18210, awarded by the Govern of Spain and co-funded by European Social Funds.

References

[1] P. Morgan, "Cybercrime facts and statistics. 2021 Report: Cyberwarfare in the C-Suite," Cybersecurity Ventures, Tech. Rep., 2021.

[2] B. E. Endicott-Popovsky and V. M. Popovsky, "Application of pedagogical fundamentals for the holistic development of cybersecurity professionals," *ACM Inroads*, vol. 5, no. 1, pp. 57–68, 2014.

[3] J. Oltsik, C. Alexander, and C. CISM, "The life and times of cybersecurity professionals," *ESG and ISSA: Research Report*, 2017.

[4] J. Vykopal, M. Vizváry, R. Oslejsek, P. Celeda, and D. Tovarnak, "Lessons learned from complex hands-on defence exercises in a cyber range," in *2017 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2017, pp. 1–8.

[5] C. Pham, D. Tang, K.-i. Chinen, and R. Beuran, "Cyris: a cyber range instantiation system for facilitating security training," in *Proceedings of the Seventh Symposium on Information and Communication Technology*, 2016, pp. 251–258.

[6] M. Rosenstein and F. Corvese, "A secure architecture for the range-level command and control system of a national cyber range testbed," in *Proceedings of the 5th USENIX conference on Cyber Security Experimentation and Test*, 2012, pp. 1–1.

[7] E. Ukwandu, M. A. B. Farah, H. Hindy, D. Brosset, D. Kavallieros, R. Atkinson, C. Tachtatzis, M. Bures, I. Andonovic, and X. Bellekens, "A review of cyber-ranges and test-beds: current and future trends," *Sensors*, vol. 20, no. 24, p. 7148, 2020.

[8] K. Leune and S. J. Petrilli Jr, "Using capture-the-flag to enhance the effectiveness of cybersecurity education," in *Proceedings of the 18th Annual Conference on Information Technology Education*, 2017, pp. 47–52.

[9] X. Ochoa and M. Worsley, "Augmenting learning analytics with multimodal sensory data," *Journal of Learning Analytics*, vol. 3, no. 2, pp. 213–219, 2016.

[10] E. Ukwandu, M. A. B. Farah, H. Hindy, D. Brosset, D. Kavallieros, R. Atkinson, C. Tachtatzis, M. Bures, I. Andonovic, and X. Bellekens, "A review of cyber-ranges and test-beds: Current and future trends," *Sensors*, vol. 20, no. 24, 2020.

[11] Whatpulse. Accessed: 2021-03-21. [Online]. Available: https://whatpulse.org

[12] P. Joshi, *OpenCV by example : enhance your understanding of computer vision and image processing by developing real-world projects in OpenCV 3*. Birmingham: Packt Publishing, 2016.

[13] J. St. Jean, *Kinect hacks*, 1st ed., ser. Hacks. Beijing ; Sebastopol, CA: O'Reilly, 2012, oCLC: ocn764382938.

[14] D. Yu and L. Deng, *Automatic Speech Recognition*. Springer London, 2015.

[15] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos)," *IEEE Access*, vol. 7, pp. 57–67, 2019.

[16] X. Shu, K. Tian, A. Ciambrone, and D. Yao, "Breaking the target: An analysis of target data breach and lessons learned," *CoRR*, vol. abs/1701.04940, 2017.

[17] S. Sriramprakash, V. D. Prasanna, and O. R. Murthy, "Stress detection in working people," *Procedia Computer Science*, vol. 115, pp. 359–366, 2017, 7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India.

[18] G. Somani, M. S. Gaur, D. Sanghi, M. Conti, and R. Buyya, "Ddos attacks in cloud computing: Issues, taxonomy, and future directions," *Computer Communications*, vol. 107, pp. 30–48, 2017.

[19] S. Huang, J. Li, P. Zhang, and W. Zhang, "Detection of mental fatigue state with wearable ecg devices," *International Journal of Medical Informatics*, vol. 119, pp. 39–46, 2018.

[20] M. Liu, E. McKelroy, S. B. Corliss, and J. Carrigan, "Investigating the effect of an adaptive learning intervention on students' learning," *Educational technology research and development*, vol. 65, no. 6, pp. 1605–1625, 2017.

Sesión de Investigación B4:
Criptografía

Sesión de Investigación B4: Criptografía
http://doi.org/10.18239/jornadas_2021.34.53

Quantum and post-quantum cryptography and cybersecurity: A systematic mapping

Beatriz García Markaida*, Xabier Larrucea†, Manuel Graña Romay‡
*ORCID: 0000-0002-7943-0323
Department of Computer Sciences and Artificial Intelligence, University of the Basque Country (UPV/EHU)
Pº Manuel Lardizabal, 1, 20018 Donostia-San Sebastián, Gipuzkoa, Spain
Email: b.garcia.markaida@gmail.com
† ORCID: 0000-0002-6402-922X
TECNALIA, Basque Research and Technology Alliance (BRTA)
Astondo bidea, 48160 Derio, Bizkaia, Spain
Email: xabier.larrucea@tecnalia.com
‡ORCID: 0000-0001-7373-4097
Department of Computer Sciences and Artificial Intelligence, University of the Basque Country (UPV/EHU)
Pº Manuel Lardizabal, 1, 20018 Donostia-San Sebastián, Gipuzkoa, Spain
Email: manuel.grana@ehu.eus

***Abstract*—Despite large-scale fault-tolerant quantum computers not yet being a reality, their eventual advent poses a menace for current cybersecurity systems. Post-quantum cryptography examines cryptographic solutions in an era where large-scale quantum computers will be available to organisms that may have an interest in breaking our current cybersecurity systems. Quantum cryptography aims to use quantum technologies to create new cryptographic solutions. In this systematic mapping, we study the current state of the art of quantum and post-quantum cryptography through a systematic search in chosen databases, and the impact this post-quantum paradigm will have on present cybersecurity.**

***Index Terms*—Cybersecurity, Post-quantum, Quantum computing, Quantum cryptography, Post-quantum cryptography, Systematic mapping**

Type of contribution: *Original research*

I. Introduction

Since its onset in the 1980s, quantum computing has been widely researched. While theoretical advancements have existed for several decades [1]–[5], it has not been until recently that technological advancements have reached the necessary level for practical implementations [6]–[10]. Even if still in their early stages, quantum computers are being refined and improved every year, and are now even available for commercial use [11]. Despite large-scale fault-tolerant quantum computers not yet being a reality, their eventual advent poses a menace for current cybersecurity systems based on hard mathematical problems such as integer factorization or the discrete logarithm problem [12], for which there will be a exponential speed-up once quantum computers reach full maturity [5]. This opens up the question of whether we should already be prepared and *how* we should be prepared for this breakthrough in quantum technologies.

In our data-driven society, securing communications is more important than ever. Current cryptosystems do not only protect our own personal privacy, but are also key in economic transactions and the proper functioning of most modern technologies [13]. As a response to the danger that quantum computing may pose to the privacy and security of our information, two main approaches have been taken. On the one hand, post-quantum cryptography examines cryptographic solutions in an era where large-scale quantum computers are available to organisms that may have an interest in breaking our current cybersecurity systems [14]. Quantum cryptography, on the other hand, aims to utilize quantum technologies to create new cryptosystems [15], [16].

With this systematic mapping, we aim to set the stage of the current state of the art of quantum and post-quantum cryptography, and the impact this post-quantum paradigm will have on present cybersecurity. To assess the urgence of the groundwork in quantum-resistant cybersecurity, we intend to examine current literature and find a sensible time-scale for the point where quantum computing would disrupt present cybersecurity. Our main interest lays in the recent advancements in the field of cryptography, whether these are feasible through classical technology or require quantum aspects.

We have structured our paper as follows: we first detail our systematic search methodology, presenting our research questions and delineating the search process and quality assessment. We then move on to present and classify our results, extracting the relevant information we have found throughout our search. We conclude our process by giving answers to our research questions in the discussion.

II. Systematic Search Methodology

To give a well-founded picture of the current state-of-the-art of quantum and post-quantum cryptography and cybersecurity, we have conducted a systematic search across several databases. We have structured this search by establishing relevant research questions and designing a search string to be used in all considered databases, following the process delineated in [17].

A. Research questions

When conducting our systematic search and further assessment of the materials found, we have considered the following

research questions:

1) By what year can we expect quantum computing to become a threat for cybersecurity?
2) Which strategies do we currently have at hand in the event that quantum computing becomes a menace for cybersecurity?
3) Which cybersecurity protocols are currently available that rely on quantum computing or quantum technologies?

B. Information sources and search process

We have conducted our search in four databases: ACM digital library, IEEE Explore, ScienceDirect and SpringerLink. Different search strings were analyzed manually in each of these data sources along November 2020. A diagram depiction our workflow can be found in Fig. 1.

Before designing a more specific search string, we first conducted a pre-identification phase; a generalistic search of the broader concepts involved in our research, namely "cryptography", "cybersecurity" OR "cyber-security" and "quantum computing". We then conducted a search that combined these three concepts, using the search string "quantum computing" AND "cryptography" AND ("cybersecurity" OR "cyber-security"). In this pre-identification stage, we found $n = 280$ articles. To the search string of the pre-identification stage, we added the term "post-quantum", since our research interest lies in an era where large-scale quantum computing is available. This is the so-called "post-quantum" era. This new and final search string, "quantum computing" AND "cryptography" AND ("cybersecurity" OR "cyber-security") AND "post-quantum" leaves us with a total of $n = 69$ search results.

C. Quality assessment and exclusion criteria

Search results were compiled into a BibTex file and imported to Mendeley Desktop. We then proceeded to check individual titles and abstracts for each article. In this process, we found that two of the search results were titles of conferences on the topic of post-quantum cybersecurity, but made reference to no specific article. Furthermore, another three entries corresponded to a chapter in a book we had no access to. We have thus discarded these $n = 5$ search results due to their lack in availability.

An initial quality assessment was performed by checking titles, abstracts and keywords of the remaining $n = 66$ search results. We fixed an exclusion criterion that at least one of these should mention either quantum computing or the post-quantum paradigm of cyptography and cybersecurity. We found that this criterion alone sufficed to discard articles that were irrelevant to our research topic, since quantum computing is often mentioned as a vague area to consider for future research even when the article at hand does not delve into it nor its consequences. This way, we discarded a further $n = 12$ search results. The remaining $n = 52$ search results were individually read for data extraction.

D. Reference follow-up and addition of relevant articles

While reading the articles selected from our systematic search, we found references to $n = 2$ items that had not appeared among our search results in any of the considered databases, but were considered relevant due to their experimental contribution [18] or extension on topics of considerable interest that had only been superficially mentioned in other results [19]. We included these publications in our study, reaching a final $n = 54$ articles to be reviewed.

III. Systematic Search Results

We proceed to analyze the results of our systematic literature search. We first classify our results and then extract relevant information from them.

A. Classification of search results

Before reading the selected publications, we compiled data on the number of results we obtained by year of publication, as seen in Fig. 2. Other than a small decrease from 2016 to 2017, we can observe a clear increase in popularity on our research topic, illustrated by the growing number of publications on the matter since 2017. It must be noted that our search was conducted in November 2020, an thus, the number of publications in 2021 correspong to online first publications.

We classified the selected publications into one of five distinct categories: communications, new cryptosystem proposals, evaluations of known cryptosystmens, experiments and reviews. This classification can be seen in Tab. I. It is worth noting that, from the 31 cryptosystem proposals, 6 of them corresponded to protocols using quantum algorithms or other quantum technologies, while the rest of the contributions, 26, did not inlcude any quantum technology in their functioning. As for the evaluations, 3 of them were presented in the form of attacks to known cryptosystems, of which 2 made use of a quantum algorithm. We also classified our search results by main topic in Tab II, by examining the abstract, title and keywords of each article.

B. The impact of future Quantum Computing on present cybersecurity

The impact of fully operational large-scale quantum computers is based on a quantum speed-up of the resolution of classically hard mathematical problems. Shor's algorithm speeds up the resolution of the discrete logarithm and integer factorization problems from exponential time in classical computing to polynomial time. The computational complexity of these problems are the basis of several encryption systems, such as RSA encryption, Diffie-Hellman, digital signature algorithms (DSA), and elliptic-curve encryption [19], [34]–[36], [39]–[41]. Grover's algorithm provides a quadratic speed-up for unstructured search problems, demanding key lengths of symmetric-key cryptosystems to be doubled in order to maintain their current security levels [66], [67]. Simon's and Bernstein-Verazani algorithms are also menaces to current cybersecurity [24], [66], [67].

This jeopardy of the currently used cryptographic primitives does not only menace our day-to-day privacy, but also government secrets [34] and large-scale technologies. Blockchain has gained significant popularity due to its transparency, redundancy and accountability. These characteristics are obtained through public-key cryptography and hash functions, and are now menaced by the eventual advent of large-scale quantum computers capable of implementing Grover's and Shor's algorithm [23], [50]. Alongside blockchain technologies, bitcoin

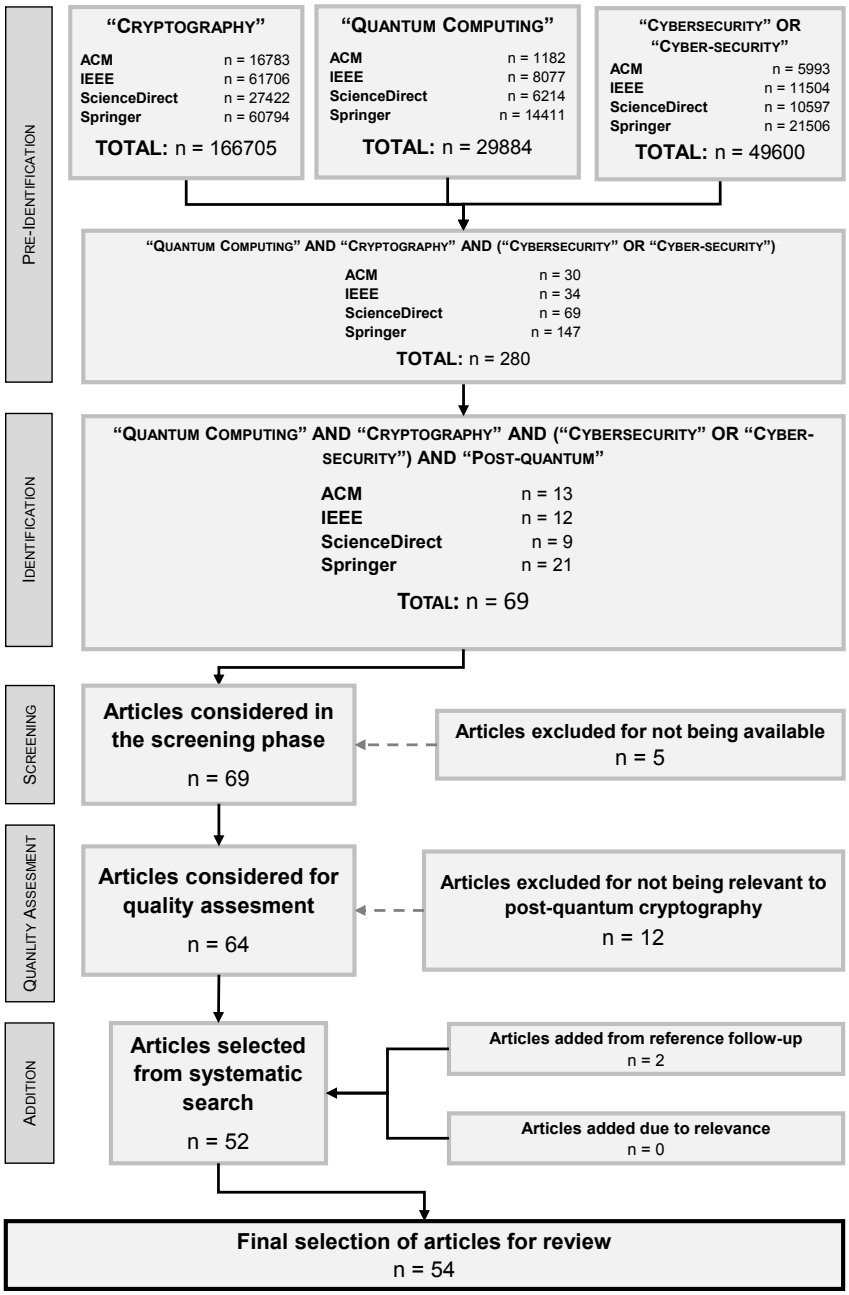


Figure 1. Diagram showing the workflow of the systematic search and refinement of article selection.

mining [22] and the internet of things (IoT) [51] are also at danger. While traditional information technologies may easily be patched in the coming years, doing so for the large scale networks of IoT devices can become a big ordeal, thus making an original quantum-resistan design of these networks a pressing matter [49].

The prospective arrival of large-scale quantum computers is presented as a menace in all the articles we reviewed. This event is estimated to arrive in the coming years, with predictions placing it as early as 2026 [69]. The most conservative estimate we have found is 2050 [39]. In all, there seems to be a consensus on establishing the time-scale for development of fully operational large-scale quantum computers capable of disrupting our current cryptosystems between five to fifteen years from now [22], [34], [36], [40], [50].

In response to the foreseeable impact of quantum computing, the National Institute of Standards and Technology (NIST) opened a call for new, quantum-resistant cryptographic primitives in 2016 [34], [36], [39], [71]. With this surge in post-quantum cryptography proposals comes the need of hardware benchmarking, which brings callenges of its own, due to the complexity of these proposals [44]. Besides NIST, other

Table I
NUMBER OF SEARCH RESULTS CLASSIFIED BY CONTRIBUTION TYPE.

Database	Communication	Cryptosystem proposal	Evaluation	Experiment	Review
ACM	1	3	3	0	2
IEEE	0	5	5	0	2
ScienceDirect	1	5	1	0	2
SpringerLink	1	21	5	0	8
Discards	0	4	3	0	5
Additions	0	0	0	1	1
Total	3	30	11	1	9

Table II
SEARCH RESULTS CLASSIFIED BY MAIN TOPIC OF THE PUBLICATION.

Topic	Communication	Cryptosystem proposal	Evaluation	Experiment	Review
Blockchain		[20]–[22]			[23], [24]
Code-based		[25]–[29]			[30]
Elliptic curves		[31], [32]	[33]		
Generalistic	[34]–[36]		[37]		[19], [38]–[41]
Group-based			[42], [43]		
Hardware			[44]		
Hash-based		[45], [46]	[47]		
Hybrid systems		[48]			
Internet of Things (IoT)			[49]		[50], [51]
Lattice-based		[52]–[61]	[62]		
Mersenne numbers			[63]		
Multivariate systems		[64], [65]			
Symmetric ciphers			[66], [67]		
Quantum bit commitment		[68]			
Quantum Key Distribution (QKD)		[69]		[18]	
Quantum money		[70]			
Quantum one-way function		[71]			

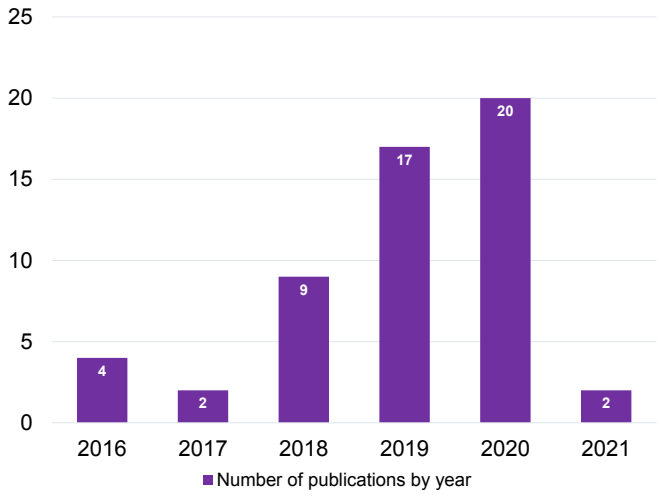


Figure 2. Number of publications by year, as of November 6th, 2020. The publications of 2021 correspond to those submitted in 2020 and accepted for publishing in 2021.

agencies and corporations have shown their interest in these aspects, such as the US National Security Agency, with their objective to transition into quantum-resistant cybersecurity systems [39] and Volkswagen, who is interested in developing quantum technological industrial applications themselves, in collaboration with Google and D-Wave systems [35].

C. Quantum-resistant classical cryptography

Classically implementable quantum-resistant cryptography proposals are mainly classified into lattice-based cryptography, solutions basen on error-correcting codes, hash-based signature schemes, multivariate systems of quadratic equations, and isogenies on elliptic curves [19], [34], [36], [40], [41], [71]. However, the computational complexity of the problems behind most of these schemes has not been mathematically proven to be NP-complete [19], [36].

1) Lattice-based cryptography: The most promising primitives among lattice-based cryptography, and the strongest candidates in the NIST competition, are NTRU and LWE [41], provinding the foundations for several protocols [52]–[54], [57]–[61]. Some of these lattice-based protocols have been proposed to be implemented in IoT settings [49], [54], provide data integrity for cloud-storage applications [61], to make multi-group signature schemes quantum-resistant [56], and to simplify certification management [57]. They also are the most promising candidates to construct post-quantum Blockchain and cryptocurrency schemes [20], [21], [23], [50]. Other contributions in this field include a method for verification of matrix-vector products, applicable in Diagonal Reduction Signature (DRS) scheme [55], password-authenticated key exchange (PAKE) protocols [59], a dual-receiver encryption scheme (DRE) [53]. However, the security estimations of [62] suggests that different NTRU and LWE schemes may be claiming the same security category while fulfilling disparate requirements.

2) Code-based cryptography: Code-based cryptosystems rank second in number of sumbitted projects and number of schemes that were selected for the second round in the NIST competition [29], [39], and yield the best performance results in the analysis by [37]. This kind of cryptosystems uses error correction codes to generate public keys from private matrices, and their low complexity for encryption and decryption makes them a fast solution with befitting security

features [39]. The proposals we have analyzed include features such as providing quantum-resistant pseudo-random number generators that also contribute to higher speed of cryptocurrencies [29] or elimination of known quasi-cyclic attacks, [26]. A significant downsize to code-based cryptography is the large size of keys, making reduction of key size a goal for many proposals in this category [25]–[28], [30]. Other lines of future research include reduction of complexity generating signatures for code-based signature systems [28], exploiting density of generator matrices of monomial codes [30], increasing performance of pseudo-random number generators [29] and construction of more rank-metric based primitives, such as proxy re-encryption and attribute-based encryption [27].

3) Hash-based cryptography: Compared with other quantum-resistant classical cryptography proposals, hash-based cryptography has the advantage of higher speed. This is achieved by basing their security on the inversibility of a one-way function, rather than on the resolution of complex mathematical problems [19], [39]. The strongest hash-based candidates build from the Merkle signature scheme [45]–[47].

4) Multivariate systems: Cryptosystems based on multivariate systems of equations produce significantly shorter signatures than those obtained in RSA, and thus offer fast signature generation and verification. They have also been classified as NP-complete [49], and are thus regarded to be quantum-resistant. However, key sizes of multivariate signature schemes are still large [19], [23], and they may not be const-friendly for practical applications [64]. Nevertheless, some multivariate systems have been optimized for uses in blockchain technologies [23], and methods for generic multivariate systems on graphic processing units that optimizes computation cost for post-quantum systems have been proposed [64].

5) Isogenies on elliptic curves: Among the proposals sumbitted to NIST, elliptic curve schemes based on isogenies show the smallest key sizes at practical security level [37]. Algorithms have been proposed that, besides displaying this characteristic shortness of keys, also improve supersingular isogeny Diffie-Hellman schemes with higher speeds, running in constant time [31] and benefiting from proposed hardware architectures [32]. Creating provably-secure isogeny-based password-authenticated key establishment protocols (PAKEs) is still a challenge; while in the classical setting this has been achieved in a number of ways, translating these protocols is vulnerable to known man-in-the-middle and offline dictionary attacks [33].

6) Other classical cryptographic systems: Research on other miscellaneous cryptosystems has been found in the literature:

- Mersenne number based schemes typically admit a low probability of decryption failure that can be exploited to gain information about secret keys [63].
- Group-based ciphers have been proposed, however, large expansion of encrypted data [42] and attacks on promising group-based NIST candidates like WalnutDSA [43] make group based knapsack ciphers unsuitable post-quantum cryptographic candidates.

D. Quantum cryptography

While classically implementable quantum-resistant solutions exist and are being analyzed, the development of quantum technologies presents us with the idea of using the very quantum mechanical properties that jeopardize our current cybersecurity paradigm to enable safe communication. Among the cryptographic tasks that quantum technologies bring to the table are quantum bit commitment, oblivious transfer, secure multiparty computation, Quantum Key Distribution, quantum fingerprints, quantum random number generators, quantum digital signatures, and quantum money [38]–[40]. Short term, with the current state of quantum technologies, we may aim for quantumly *enhanced* cybersecurity, relying on small quantum devices and establishing short-scale quantum communication channels. The arrival of large-scale quantum computers would allow us to achieve quantumly *enabled* cybersecurity, which would require verifiable blind quantum computing [40].

1) Quantum Key Distribution: The most widely researched area in quantum cryptography is Quantum Key Distribution (QKD) [38]. This key distribution protocol gives unlimited amounts of secure symmetric keys for use in one-time pad applications [39]. Through exchange quantum bits −or *qubits*− through a quantum channel, two parties are able to create and distribute a secure key [38]. This may be done through discrete qubits, where each bit of is encoded into the discrete degrees of freedom of an optical signal, or continuous variable qubits using coherent communication techniques [69]. This protocol also relies on an authenticated classical channel [38].

QKD has already been physically implemented in several contexts, such as securing communications during a 2007 election in Switzerland [34], [39], Secure communication between Austria and China [18], [40], and other working QKD network implementations in the US, Europe, Japan and China [69], as well as commercially available systems [34], [39]. QKD has even been proposed as a countermeasure to possible quantum attacks in drone and hand-based IoT devices [69] or to construct confidential channels and post-quantum cryptography for the deployment of authenticated blockchain channels [24]. Classical-quantum hybrid approaches have also been proposed, such as an Hybrid Authentication Key Scheme (HAKE) that uses symmetric keys of QKD to provide message authentication without resorting to costly post-quantum signature schemes, though it does not integrate QKD security proofs and considerations into their otherwise benchmarked protocol [48].

Advancements notwithstanding, QKD implementations are still far from ideal and have many challenges to face. Imperfections in infrastructure including optical fibers, photon sources and detectors [34] can result in side-channels that can be used by an eavesdropper, compromising data security [49], and quantum bits are also vulnerable to photon number and beam-splitting attacks [40]. Long-distance application of QKD is still a challenge [69], even though communications between partys as far apart as 7600 km have been achieved through relying on a trusted satellite [18]. This need for a trusted satellite may be overcome with entanglement-based QKD protocols, and multiparty connections from satellites to various ground stations and between large ground networks

are left as interesting projects to be implemented in the future [18].

2) Other quantum technologies: Other cybersecurity proposals have been found in our systematic search:

- An introduction of game-theoretic security for quantum bit commitment, with a test experiment on IBM's cloud quantum computers [68].
- A proposal of a semi-quantum money scheme that requires quantum technologies just from the user, instead of both user and bank [70]. In this protocol, the key generation stage is the only quantum part, which presents an interesting open question: Can other multiparty protocols be carried out where at least one party does not have access to a quantum computer, or if they have access to an untrusted quantum computer alongside a trusted classical system?
- A quantum one-way function [71]. Quantum computers can obtain properties of chemical systems from a given Hamiltonian, but the opposite process is not possible. This protocol proposes to explot this fact to build a one-way function than cannot be broken trough quantum computers. Whether classical computers have the ability of to perform this task is unknown.

IV. Discussion

With the knowledge obtained from our results, we may find answers to the research questions presented in section II-A.

A. By what year can we expect quantum computing to become a menace for cybersecurity?

We have found different estimations for the arrival of large-scale quantum computers that would be able to disrupt current cryptography. From earliest to latest, the estimates are 2026 [69], 2027 with a 1 in 6 probability and 2031 [34] with a 1 in 2 probability, 2050 [39], and wider time-intervals as "over the next 5 or 10 years" in a 2019 publication [40] and "10 or 15 years in the future" in a 2020 publication [36]. In conclusion, the advent of fault-tolerant large-scale quantum computers is being placed between 2024 and 2050, with a stronger tendency towards the late 2020s or early 2030s.

B. Which strategies do we currently have at hand in the event that quantum computing becomes a menace for cybersecurity?

Several promising quantum-resistant classical proposals have been submitted to the NIST call of 2016. Among them, the most auspicious cryptographic primitives are lattice-based cryptography (especially, NTRU and LWE schemes) [20], [21], [23], [49], [50], [52]–[61] and code-based cryptography [25]–[30]. Draft standards are expected between 2021 and 2022 [71]. When it comes to strategies relying on quantum technologies, Quantum Key Distribution is the most promising [18], [24], [39], [48], [69], and will be addressed in the next questions.

C. Which are the currently used cybersecurity protocols that rely on quantum computing or quantum technologies?

Most quantum cryptography proposals have not yet been implemented. However, despite challenges in practical implementation, several Quantum Key Distribution applications have been achieved, namely, securing communications during elections in Switzerland in 2007 [34], [39] or establishing long-distance communication between China and Austria in 2018 [18], as well as several commercial systems being available [34], [39]. Despite these advancements, this field still has many challenges to overcome, such as implementation and improvement of infrastructure and further enabling long-distance communication.

V. Threats to validity

A. Construct validity

Our research questions may not be able to completely cover all the relevant issues around the impact of quantum computing on cryptography and cybersecurity. An inaccurate use of keywords in the publications we mapped may also threaten the validity of the study.

B. Internal validity

In spite of the broad nature of our search string, the number of search results after combination of all four platforms is relatively small, and thus all relevant areas in the topic might not have been covered. All published results presented quantum computing as a menace to cybersecurity, which might have been influenced by a sort of publication bias; quantum-resistant cryptographic protocols do not need to be published if quantum computing is not a menace at all. On another note, this was a partial study carried out by a small group of researchers, so they might have bypassed important information that would have otherwise been detected by a larger group of authors.

C. External validity

Some of the authors could not access several search results and could thus make no judgements of the contents that were presented in them. This may have brought the disregard of relevant information.

D. Conclusion validity

Following the steps in our search and mapping process should lead to the same conclusions we found. The process has been transparent and can be replicated at ease. We do not anticipate new results that may arise while replicating our process in the future.

VI. Conclusion

This paper carried out a systematic search and mapping of existing literature on quantum computing and its impact on current cryptography and cybersecurity. We examined 54 research publications and were able to identify the repercussions of emerging quantum technologies, the present strategies for making our cybersecurity systems quantum-resistant and some applications of these new technologies. We found strong evidence that the consensus about the eventual advent of fault-tolerant large-scale quantum computers is that it is bound to happen in the next decade. Should this be considered a "short-term" menace? Given the interest of organizations like the US National Security Agency [39] and the National Institute of Standards and Technology (NIST) [34], [36], [39], [71], it does seem that this is short-term *enough* to take action now. Testing new cryptographic protocols, hardware benchmarking and implementation into systems and networks takes work and

time [44], and we should not wait until fault-tolerant large-scale quantum computers are a reality to search for strong quantum-resistant cybersecurity candidates. NIST's call for post-quantum cryptographic solutions reached its third round in July 2020 [72], with a variety of classical schemes that will concievably grant this protection, such as lattice-, code- and hash-based cryptography, multivariate systems of non-linear equations and elliptic-curve cryptography. Quantum technologies are also being proposed and applied in cybersecurity, Quantum Key Distribution being the strongest candidate in this realm. However, further research and development is needed in this area.

Appendix: Papers from the systematic mapping

All the papers used in our study correspond to the references in Tab. II.

Aknowledgements

This work has been partially supported by the Basque Governement (SPRI) project called Trustind - Creating Trust In The Industrial Digital Transformation (KK-2020/00054) and in collaboration with the UPV/EHU.

References

[1] R. P. Feynman, "Simulating physics with computers," *International Journal of Theoretical Physics*, vol. 21, no. 6-7, pp. 467–488, Jun. 1982.

[2] A. Peres, "Reversible logic and quantum computers," *Phys. Rev. A*, vol. 32, pp. 3266–3276, Dec 1985.

[3] I. L. Chuang and Y. Yamamoto, "Simple quantum computer," *Phys. Rev. A*, vol. 52, pp. 3489–3496, Nov 1995.

[4] W. G. Unruh, "Maintaining coherence in quantum computers," *Phys. Rev. A*, vol. 51, pp. 992–997, Feb 1995.

[5] P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," *SIAM Review*, vol. 41, no. 2, pp. 303–332, Jan. 1999.

[6] J. Emerson, R. Alicki, and K. Życzkowski, "Scalable noise estimation with random unitary operators," *Journal of Optics B: Quantum and Semiclassical Optics*, vol. 7, no. 10, pp. S347–S352, Sep. 2005.

[7] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, "Randomized benchmarking of quantum gates," *Phys. Rev. A*, vol. 77, p. 012307, Jan 2008.

[8] E. Kapit, "Hardware-efficient and fully autonomous quantum error correction in superconducting circuits," *Phys. Rev. Lett.*, vol. 116, p. 150501, Apr 2016.

[9] J. J. Wallman and J. Emerson, "Noise tailoring for scalable quantum computation via randomized compiling," *Phys. Rev. A*, vol. 94, p. 052325, Nov 2016.

[10] B. Pokharel, N. Anand, B. Fortman, and D. A. Lidar, "Demonstration of fidelity improvement using dynamical decoupling with superconducting qubits," *Phys. Rev. Lett.*, vol. 121, p. 220502, Nov 2018.

[11] System one - IBM quantum. [Online]. Available: https://www.research.ibm.com/quantum-computing/system-one/

[12] T. Elgamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *IEEE Transactions on Information Theory*, vol. 31, no. 4, pp. 469–472, Jul. 1985.

[13] M. Amara and A. Siad, "Elliptic curve cryptography and its applications," in *International Workshop on Systems, Signal Processing and their Applications, WOSSPA*. IEEE, May 2011.

[14] D. J. Bernstein, "Introduction to post-quantum cryptography," in *Post-Quantum Cryptography*. Springer Berlin Heidelberg, pp. 1–14.

[15] C. H. Bennett, G. Brassard, S. Popescu, B. Schumacher, J. A. Smolin, and W. K. Wootters, "Purification of noisy entanglement and faithful teleportation via noisy channels," *Phys. Rev. Lett.*, vol. 76, pp. 722–725, Jan 1996.

[16] F. Cavaliere, J. Mattsson, and B. Smeets, "The security implications of quantum cryptography and quantum computing," *Network Security*, vol. 2020, no. 9, pp. 9–15, Sep. 2020.

[17] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – a systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, Jan. 2009.

[18] S.-K. Liao, W.-Q. Cai, J. Handsteiner, B. Liu, J. Yin, L. Zhang, D. Rauch, M. Fink, J.-G. Ren, W.-Y. Liu, Y. Li, Q. Shen, Y. Cao, F.-Z. Li, J.-F. Wang, Y.-M. Huang, L. Deng, T. Xi, L. Ma, T. Hu, L. Li, N.-L. Liu, F. Koidl, P. Wang, Y.-A. Chen, X.-B. Wang, M. Steindorfer, G. Kirchner, C.-Y. Lu, R. Shu, R. Ursin, T. Scheidl, C.-Z. Peng, J.-Y. Wang, A. Zeilinger, and J.-W. Pan, "Satellite-Relayed Intercontinental Quantum Network," *Physical Review Letters*, vol. 120, no. 3, p. 030501, jan 2018.

[19] J. A. Buchmann, D. Butin, F. Göpfert, and A. Petzoldt, "Post-quantum cryptography: State of the art," in *The New Codebreakers*. Springer Berlin Heidelberg, 2016, pp. 88–108.

[20] Y. Gao, X. Chen, Y. Chen, Y. Sun, X. Niu, and Y. Yang, "A Secure Cryptocurrency Scheme Based on Post-Quantum Blockchain," *IEEE Access*, vol. 6, pp. 27 205–27 213, 2018.

[21] N. Alkeilani Alkadri, P. Das, A. Erwig, S. Faust, J. Krämer, S. Riahi, and P. Struck, "Deterministic Wallets in a Quantum World," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1017–1031.

[22] O. Sattath, "On the insecurity of quantum Bitcoin mining," *International Journal of Information Security*, vol. 19, no. 3, pp. 291–302, Jun 2020.

[23] T. M. Fernández-Caramès and P. Fraga-Lamas, "Towards Post-Quantum Blockchain: A Review on Blockchain Cryptography Resistant to Quantum Computing Attacks," *IEEE Access*, vol. 8, pp. 21 091–21 116, 2020.

[24] V. Fernandez, A. B. Orue, and D. Arroyo, "Securing Blockchain with Quantum Safe Cryptography: When and How?" in *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020)*, Á. Herrero, C. Cambra, D. Urda, J. Sedano, H. Quintián, and E. Corchado, Eds. Cham: Springer International Publishing, 2021, pp. 371–379.

[25] K. N. Rzaev, "Mathematical Models of Modified Crypto-Code Means of Information Protection Based on Coding Theory Schemes," *Automation and Remote Control*, vol. 80, no. 7, pp. 1304–1316, Jul 2019.

[26] L.-P. Wang and J. Hu, "Two new module-code-based KEMs with rank metric," in *Information Security and Privacy*. Springer International Publishing, 2019, pp. 176–191.

[27] P. Zeng, S. Chen, and K.-K. R. Choo, "An IND-CCA2 secure post-quantum encryption scheme and a secure cloud storage use case," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, p. 32, Dec 2019.

[28] A. Kuznetsov, A. Kiian, V. Babenko, I. Perevozova, I. Chepurko, and O. Smirnov, "New Approach to the Implementation of Post-Quantum Digital Signature Scheme," in *2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, May 2020, pp. 166–171.

[29] A. Kuznetsov, A. Kiian, O. Smirnov, A. Cherep, M. Kanabekova, and I. Chepurko, "Testing of Code-Based Pseudorandom Number Generators for Post-Quantum Application," in *2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, May 2020, pp. 172–177.

[30] M. Baldi, P. Santini, and G. Cancellieri, "Post-quantum cryptography based on codes: State of the art and open challenges," in *2017 AEIT International Annual Conference*, 2017, pp. 1–6.

[31] C. Costello, P. Longa, and M. Naehrig, "Efficient algorithms for supersingular isogeny diffie-hellman," in *Advances in Cryptology – CRYPTO 2016*. Springer Berlin Heidelberg, 2016, pp. 572–601.

[32] C. Liu, J. Ni, W. Liu, Z. Liu, and M. O'Neill, "Design and Optimization of Modular Multiplication for SIDH," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5.

[33] R. Azarderakhsh, D. Jao, B. Koziel, J. T. LeGrow, V. Soukharev, and O. Taraskin, "How not to create an isogeny-based PAKE," in *Applied Cryptography and Network Security*. Springer International Publishing, 2020, pp. 169–186.

[34] M. Brooks, "S.O.S. (Save Our Secrets)," *New Scientist*, vol. 237, no. 3167, pp. 40–43, 2018.

[35] S. Yarkoni, M. Leib, A. Skolik, M. Streif, F. Neukart, and D. von Dollen, "Volkswagen and quantum computing: An industrial perspective," *Digitale Welt*, vol. 3, no. 2, pp. 34–37, Apr 2019.

[36] G. Mone, "The Quantum Threat," *Commun. ACM*, vol. 63, no. 7, pp. 12–14, 2020.

[37] F. Borges, P. R. Reis, and D. Pereira, "A Comparison of Security and its Performance for Key Agreements in Post-Quantum Cryptography," *IEEE Access*, vol. 8, pp. 142 413–142 422, 2020.

[38] A. Broadbent and C. Schaffner, "Quantum cryptography beyond quantum key distribution," *Designs, Codes and Cryptography*, vol. 78, no. 1, pp. 351–382, jan 2016.

[39] L. O. Mailloux, C. D. Lewis II, C. Riggs, and M. R. Grimaila, "Post-Quantum Cryptography: What Advancements in Quantum Computing Mean for IT Professionals," *IT Professional*, vol. 18, no. 5, pp. 42–47, 2016.

[40] P. Wallden and E. Kashefi, "Cyber security in the quantum era," *Communications of the ACM*, vol. 62, no. 4, pp. 120–120, Mar. 2019.

[41] S. Agrawal, "Post-quantum cryptography: An introduction," in *Cyber Security in India*. Springer Singapore, 2020, pp. 103–108.

[42] A. Vambol, "The prospects for group-based knapsack ciphers in the post-quantum era," in *2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, May 2018, pp. 271–275.

[43] S.-P. Merz and C. Petit, "Factoring products of braids via garside normal form," in *Public-Key Cryptography – PKC 2019*. Springer International Publishing, 2019, pp. 646–678.

[44] K. Gaj, "Challenges and Rewards of Implementing and Benchmarking Post-Quantum Cryptography in Hardware," in *Proceedings of the 2018 on Great Lakes Symposium on VLSI*, ser. GLSVLSI '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 359–364.

[45] V. B. Kumar, N. Gupta, A. Chattopadhyay, M. Kasper, C. Kraus, and R. Niederhagen, "Post-Quantum Secure Boot," in *Proceedings of the 2020 Design, Automation and Test in Europe Conference and Exhibition, DATE 2020*, ser. DATE '20. San Jose, CA, USA: EDA Consortium, 2020, pp. 1582–1585.

[46] M. Iavich, S. Gnatyuk, A. Arakelian, G. Iashvili, Y. Polishchuk, and D. Prysiazhnyy, "Improved post-quantum merkle algorithm based on threads," in *Advances in Computer Science for Engineering and Education III*. Springer International Publishing, Aug. 2021, pp. 454–464.

[47] M. D. Noel, O. V. Waziri, M. S. Abdulhamid, and A. J. Ojeniyi, "Stateful Hash-based Digital Signature Schemes for Bitcoin Cryptocurrency," in *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*, 2019, pp. 1–6.

[48] B. Dowling, T. B. Hansen, and K. G. Paterson, "Many a mickle makes a muckle: A framework for provably quantum-secure hybrid key exchange," in *Post-Quantum Cryptography*. Springer International Publishing, 2020, pp. 483–502.

[49] O. S. Althobaiti and M. Dohler, "Cybersecurity Challenges Associated With the Internet of Things in a Post-Quantum World," *IEEE Access*, vol. 8, pp. 157 356–157 381, 2020.

[50] Q. Zhu, S. W. Loke, R. Trujillo-Rasua, F. Jiang, and Y. Xiang, "Applications of Distributed Ledger Technologies to the Internet of Things: A Survey," *ACM Comput. Surv.*, vol. 52, no. 6, Nov 2019.

[51] A. Jurcut, T. Niculcea, P. Ranaweera, and N.-A. Le-Khac, "Security Considerations for Internet of Things: A Survey," *SN Computer Science*, vol. 1, no. 4, p. 193, Jul 2020.

[52] C. Borcea, A. B. D. Gupta, Y. Polyakov, K. Rohloff, and G. Ryan, "PICADOR: End-to-end encrypted Publish–Subscribe information distribution with proxy re-encryption," *Future Generation Computer Systems*, vol. 71, pp. 177–191, 2017.

[53] D. Zhang, K. Zhang, B. Li, X. Lu, H. Xue, and J. Li, "Lattice-based dual receiver encryption and more," in *Information Security and Privacy*. Springer International Publishing, 2018, pp. 520–538.

[54] B. Mi, D. Huang, S. Wan, Y. Hu, and K.-K. R. K. R. Choo, "A post-quantum light weight 1-out-n oblivious transfer protocol," *Computers and Electrical Engineering*, vol. 75, pp. 90–100, 2019.

[55] A. Sipasseuth, T. Plantard, and W. Susilo, "Using freivalds' algorithm to accelerate lattice-based signature verifications," in *Information Security Practice and Experience*. Springer International Publishing, 2019, pp. 401–412.

[56] T. Qiu, L. Hou, and D. Lin, "A multi-group signature scheme from lattices," in *Information and Communications Security*. Springer International Publishing, 2020, pp. 359–377.

[57] Y. Shi, S. Qiu, and J. Liu, "An efficient lattice-based IBE scheme using combined public key," in *Communications in Computer and Information Science*. Springer Singapore, 2020, pp. 3–16.

[58] W. Tan, B. M. Case, G. Hu, S. Gao, and Y. Lao, "An Ultra-Highly Parallel Polynomial Multiplier for the Bootstrapping Algorithm in a Fully Homomorphic Encryption Scheme," *Journal of Signal Processing Systems*, Oct 2020.

[59] Y. Yang, X. Gu, B. Wang, and T. Xu, "Efficient password-authenticated key exchange from RLWE based on asymmetric key consensus," in *Information Security and Cryptology*. Springer International Publishing, 2020, pp. 31–49.

[60] X. Zhang, Y. Tang, H. Wang, C. Xu, Y. Miao, and H. Cheng, "Lattice-based proxy-oriented identity-based encryption with keyword search for cloud storage," *Information Sciences*, vol. 494, pp. 193–207, 2019.

[61] X. Zhang, C. Huang, Y. Zhang, J. Zhang, and J. Gong, "LDVAS: Lattice-Based Designated Verifier Auditing Scheme for Electronic Medical Data in Cloud-Assisted WBANs," *IEEE Access*, vol. 8, pp. 54 402–54 414, 2020.

[62] M. R. Albrecht, B. R. Curtis, A. Deo, A. Davidson, R. Player, E. W. Postlethwaite, F. Virdia, and T. Wunderer, "Estimate all the {LWE, NTRU} schemes!" in *Lecture Notes in Computer Science*. Springer International Publishing, 2018, pp. 351–367.

[63] M. Tiepelt and J.-P. D'Anvers, "Exploiting Decryption Failures in Mersenne Number Cryptosystems," in *Proceedings of the 7th ACM Workshop on ASIA Public-Key Cryptography*, ser. APKC '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 45–54.

[64] G. Liao, Z. Gong, Z. Huang, and W. Qiu, "A generic optimization method of multivariate systems on graphic processing units," *Soft Computing*, vol. 22, no. 23, pp. 7857–7864, Dec 2018.

[65] A. Andrushkevych, Y. Gorbenko, O. Kuznetsov, R. Oliynykov, and M. Rodinko, "A prospective lightweight block cipher for green IT engineering," in *Green IT Engineering: Social, Business and Industrial Applications*. Springer International Publishing, Sep. 2019, pp. 95–112.

[66] H. Xie and L. Yang, "Using Bernstein–Vazirani algorithm to attack block ciphers," *Designs, Codes and Cryptography*, vol. 87, no. 5, pp. 1161–1182, May 2019.

[67] ——, "A quantum related-key attack based on the Bernstein–Vazirani algorithm," *Quantum Information Processing*, vol. 19, no. 8, p. 240, Aug 2020.

[68] L. Zhou, X. Sun, C. Su, Z. Liu, and K.-K. K. Raymond Choo, "Game theoretic security of quantum bit commitment," *Information Sciences*, vol. 479, pp. 503–514, 2019.

[69] S. Arnon and J. Kupferman, "Effects of Weather on Drone to IoT QKD," in *Cyber Security Cryptography and Machine Learning*, S. Dolev, D. Hendler, S. Lodha, and M. Yung, Eds. Cham: Springer International Publishing, 2019, pp. 67–74.

[70] R. Radian and O. Sattath, "Semi-quantum money," in *AFT 2019 - Proceedings of the 1st ACM Conference on Advances in Financial Technologies*, ser. AFT '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 132–146.

[71] M. Paeschke, W. Fumy, and A. Wilke, "Ensuring Security and Trust in a Post-Quantum Environment," *Datenschutz und Datensicherheit - DuD*, vol. 43, no. 7, pp. 440–443, Jul 2019.

[72] Post-quantum cryptography — CSRC. [Online]. Available: https://csrc.nist.gov/projects/post-quantum-cryptography/round-3-submissions

Sesión de Investigación B4: Criptografía
http://doi.org/10.18239/jornadas_2021.34.54

Homomorphic SVM Inference for Fraud Detection

A. Vázquez-Saavedra, G. Jiménez-Balsa, J. Loureiro-Acuña
Gradiant
Carretera de Vilar, 56, Vigo, Pontevedra, 36214, Spain
{avsaavedra, gjimenez, jloureiro}@gradiant.org

M. Fernández-Veiga, A. Pedrouzo-Ulloa
atlanTTic Research Center, Universidade de Vigo
E.E. Telecomunicación, Vigo 36310, Spain
mveiga@det.uvigo.es, apedrouzo@gts.uvigo.es

***Abstract*—Nowadays, cloud computing has become a very promising solution for almost all companies, as it offers the possibility of saving costs by outsourcing computation on-demand. However, some companies deal with private information, which must be protected before outsourcing. Banks, whose financial information is highly sensitive, are one remarkable example of this problem. Their typical processes must be run on their systems for security and regulation reasons, which impedes to take advantage of the scalability and flexibility benefits introduced by the cloud. A relevant example on which we focus in this work is the case of fraud detection systems, for which we propose the use of modern lattice-based homomorphic encryption for its secure execution. To this end, we implement and validate the performance of a homomorphic SVM (Support Vector Machine) classifier for secure fraud detection, showing the feasibility of securely outsourcing fraud detection inference.**

***Index Terms*—Support Vector Machines, Fraud Detection, Homomorphic Encryption, Lattice-based Cryptography**

Type of contribution: *Ongoing research*

I. Introduction

The impact of cloud computing on industry and also end users is difficult to estimate: many aspects of everyday life have been transformed by the omnipresence of software running on cloud networks. On the one hand, cloud computing allows companies to optimise costs and increase their offerings, without having the need of purchasing and managing all the hardware and software. This allows them to launch globally-available apps and online services without having to spend resources on the platform on which they will run. On the other hand, end users can access these applications immediately and without any specific requirements, as all these services are easily available on the Internet. Unfortunately, not all companies can benefit from the cloud computing approach.

A major disadvantage which is present in cloud-based applications is their underlying security. Actually, the use of cloud-based services always leads to the storage of information in third party systems, which could be often considered as untrusted environments. Although, in general, this is not a problem for many applications, the situation is considerably aggravated in those use cases which deal with especially privacy-sensitive information, on which security is a priority and must be maximised.

The list of related use cases encompasses companies that must handle sensitive information, such as financial and medical data, religious information, etc. This type of data is protected by law and, as companies could be exposed to different sanctions, they are forced to guarantee a certain level of protection. Even so, although conventional cryptographic techniques allow for secure storage in the cloud, data protection is not so easy if some sort of computation has to be applied on the data. Thus, in view of all these shortcomings, companies which handle sensitive data should seemingly avoid cloud-based solutions.

Precisely, the field of privacy-preserving machine learning (PPML) [1] deals with the different security and privacy threats appearing in machine learning (ML). Actually, paying attention to the scenario of cloud computing, and among the broad set of available tools inside PPML, homomorphic encryption techniques, which enable secure processing on encrypted data, seem to be a perfect fit for secure outsourcing.

A. Our Contributions

Our scenario is set inside a banking fraud context [2], on which banks make use of fraud detection systems that require a large amount of resources to operate. Bank transactions do not follow a uniform time distribution, so a good solution for banks is to outsource their fraud detection systems, which allows them to take advantage of the flexibility and scalability provided by the cloud.

Our main contribution is to *implement and showcase the feasibility of a solution based on homomorphic cryptography* (see Section II), which *enables to securely outsource the fraud detection systems*. In particular, we have implemented a SVM (Support Vector Machine) inference algorithm in the encrypted domain (see Sections III and IV); being this classifier tailored for bank fraud detection. *Our homomorphic SVM classifier allows a bank to make secure encrypted predictions in an untrusted environment*.

In order to test its feasibility: (1) we have designed a proof of concept based on a client-server model (see Section IV) and, (2) we have evaluated the runtime and classification performance of the system (see Section V).

Notation and Structure: We represent vectors and matrices by boldface lowercase and uppercase letters, respectively. Polynomials are denoted with regular lowercase letters, ignoring the polynomial variable (e.g., a instead of $a(z)$). Finally, the Hadamard product of two vectors is $\boldsymbol{a} \circ \boldsymbol{b}$.

The rest of the paper is organized as follows: Section II briefly reviews the used homomorphic encryption scheme. Section III details the bank fraud detection scenario, the dataset and the SVM classifier. Section IV introduces the design of our system and the homomorphic classifier. Finally,

This work was funded by the Ayudas Cervera para Centros Tecnológicos, grant of the Spanish Centre for the Development of Industrial Technology (CDTI) under the project EGIDA (CER-20191012). Also funded by the Agencia Estatal de Investigación (Spain) and the European Regional Development Fund (ERDF) under project RODIN (PID2019-105717RB-C21), and by the Xunta de Galicia and ERDF under project ED431G2019/08.

TABLE I
HIGH LEVEL DESCRIPTION OF THE CKKS SCHEME [9]

Parameters: Let $R_q[z]$ be the quotient polynomial ring $\mathbb{Z}_q[z]/(1+z^n)$. Then, ciphertexts from the $\mathsf{E}_{\mathsf{CKKS}}$ scheme are composed of two polynomial elements from $R_q[z]$, and plaintexts belong to the set $\mathbb{C}^P$ such that $P = n/2$. Finally, q and n (for simplicity we omit some internal parameters) are chosen in terms of the security parameter λ.	
$\mathsf{E}_{\mathsf{CKKS}}$.SecKeyGen	**Input**: Security parameter λ **Output**: Secret key sk
$\mathsf{E}_{\mathsf{CKKS}}$.PubKeyGen	**Input**: Secret key sk **Output**: Public key pk, evaluation key evk and rotation key rtk
$\mathsf{E}_{\mathsf{CKKS}}$.Enc	**Input**: Plaintext $\boldsymbol{m} \in \mathbb{C}^P$ **Output**: A ciphertext $c = (c_0, c_1) \in R_q^2[z]$
$\mathsf{E}_{\mathsf{CKKS}}$.Dec	**Input**: A ciphertext $c' = (c'_0, c'_1) \in R_q^2[z]$ encrypting a plaintext $\boldsymbol{m}' \in \mathbb{C}^P$ **Output**: A plaintext $\boldsymbol{m}' \in \mathbb{C}^P$
$\mathsf{E}_{\mathsf{CKKS}}$.Add	**Input**: Ciphertexts c and c' encrypting, respectively, $\boldsymbol{m}$ and $\boldsymbol{m}'$ **Output**: A ciphertext $c'' = (c''_0, c''_1) \in R_q^2[z]$ encrypting $\boldsymbol{m} + \boldsymbol{m}' = \boldsymbol{m}'' \in \mathbb{C}^P$
$\mathsf{E}_{\mathsf{CKKS}}$.LinHadMult	**Input**: A plaintext $\boldsymbol{m} \in \mathbb{C}^P$ and a ciphertext c' encrypting $\boldsymbol{m}'$ **Output**: A ciphertext $c'' = (c''_0, c''_1) \in R_q^2[z]$ encrypting $\boldsymbol{m} \circ \boldsymbol{m}' = \boldsymbol{m}'' \in \mathbb{C}^P$
$\mathsf{E}_{\mathsf{CKKS}}$.HadMult	**Input**: Evaluation key evk, and ciphertexts c and c' encrypting, respectively, $\boldsymbol{m}$ and $\boldsymbol{m}'$ **Output**: A ciphertext $c'' = (c''_0, c''_1) \in R_q^2[z]$ encrypting $\boldsymbol{m} \circ \boldsymbol{m}' = \boldsymbol{m}'' \in \mathbb{C}^P$
$\mathsf{E}_{\mathsf{CKKS}}$.Rot	**Input**: A ciphertext c encrypting $\boldsymbol{m} \in \mathbb{C}^P$ **Output**: A ciphertext $c' = (c'_0, c'_1) \in R_q^2[z]$ encrypting $\boldsymbol{m}' \in \mathbb{C}^P$, which is the result of applying a rotation over the components of $\boldsymbol{m}$

Section V evaluates the performance of our system, and Section VI discusses some future work lines.

II. PRELIMINARIES: HOMOMORPHIC ENCRYPTION

Homomorphic encryption appears as a very promising tool for secure processing [3]. One relevant example is the Paillier cryptosystem [4], which has been used in a broad range of different applications [5]. It presents a group homomorphism which enables additions between two encrypted values, and multiplications between a ciphertext and a plaintext.

Consequently, Paillier could be a perfect candidate to implement our proposed encrypted detector for bank fraud detection. However, modern lattice-based cryptosystems outperform Paillier in several aspects: (1) *more complex applications are possible* as they present a ring homomorphism which enables multiplication between two ciphertexts (e.g., the training of a SVM [6]), and (2) *better performance*, as Paillier is slower even when considering only linear operations [7], [8].

In view of the above points, we make use of the CKKS scheme [9], which is a lattice-based cryptosystem especially adapted to work with approximate arithmetic operations.

A. A concrete homomorphic encryption scheme

We give a brief description of the CKKS scheme $\mathsf{E}_{\mathsf{CKKS}}$ in Table I. It is defined as a conventional public-key encryption scheme $\mathsf{E} = \{\mathsf{SecKeyGen}, \mathsf{PubKeyGen}, \mathsf{Enc}, \mathsf{Dec}\}$, but extended with Add, HadMult, LinHadMult and Rot procedures. For further details of the scheme we refer the reader to [9].

III. USE CASE

Banks use fraud detection systems that require a large amount of computational resources. This results in significant costs for banks to keep their systems up and running. One possible approach for banks is to outsource these systems to the cloud where, in addition to saving costs, they would obtain more flexible systems.

However, banking information is sensitive and traditional techniques only allow to work with data in clear, what can be a security problem. Consequently, we propose to implement the evaluation phase of an important machine learning algorithm such as Support Vector Machines (SVM) in the encrypted domain. Therefore, by directly working with encrypted data, banks could outsource this information securely and take advantage of the benefits provided by the cloud.

A. Dataset

The dataset used for this paper is called IEEE-CIS Fraud Detection [10]. It is composed of real-world e-commerce transactions provided by Vesta, a leader payment service company. The dataset is divided into 4 files train_transaction.csv, train_identity.csv, test_transaction.csv and test_identity.csv. However, we only use the files train_transaction.csv and train_identity.csv, as the other two are unlabeled and we could not use the samples they contain to test the results of the inference. Consequently, these two training files are combined, finally obtaining a dataset with 590540 rows and 434 features.

B. Dataset pre-processing

Given the large dimensionality of the training dataset, an adequate pre-processing is required before training. As our main objective in this work focuses on implementing the SVM inference function in the encrypted domain, we have followed the guidelines for data pre-processing and feature selection proposed in [10]. According to their recommendations:

- The number of features is reduced from 434 to 115.
- The columns with missing values are filled with the mean, mode or the most frequent category among the present values in the corresponding column.

C. SVM linear kernel

We have chosen a Support Vector Machine (SVM) classifier due to its good performance for the fraud detection scenario [10]. SVMs correspond to a type of binary classifiers which, given a set of training samples, maximize the gap between the two considered groups (e.g., 1 fraud vs -1 non-fraud in our scenario).

Specifically, in this work we are interested in the expression of the SVM classifier:

$$\mathsf{sign}\left(\underbrace{\sum_{i=1}^{L} y_i \alpha_i \mathsf{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b}_{\mathsf{score}}\right), \tag{1}$$

where $\boldsymbol{x}$ is the input sample, and the rest of parameters are obtained through the SVM training: b is the bias, α_i are the dual coefficients, and $\boldsymbol{x}_i$ and y_i are, respectively, the support vectors together with their associated label (1 or -1). We refer to [11] for more details on the training of SVMs.

However, due to the limitations of homomorphic encryption schemes, two additional points must be considered:

- The $\mathsf{sign}(\cdot)$ function can be very costly to be homomorphically computed. As it does not introduce a relevant overhead, we have moved it to the client side, which calculates it after decryption.
- Among the possible choices for the kernel $\mathsf{K}(\cdot,\cdot)$ function, we make use of a linear kernel $\mathsf{K}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u} \cdot \boldsymbol{v}$; mainly due to its simplicity and good behaviour for the fraud detection problem [10]. Additionally, it brings about an important efficiency advantage thanks to the fact that the inner product is distributive over vector addition, which enables to considerably simplify the classifier.

Consequently, taking into account the above changes, our homomorphic classifier turns out to be:

$$\mathsf{score} = \underbrace{\left(\sum_i y_i \alpha_i \boldsymbol{x}_i\right)}_{\text{linear SVM model}} \cdot \underbrace{\boldsymbol{x}}_{\text{input sample}} + b. \qquad (2)$$

For more details on its homomorphic computation and the different packing methods, we refer the reader to Section IV-B.

IV. SVM Homomorphic Inference

So as to showcase the potential and practicability of homomorphic encryption for PPML, and more specifically, for the encrypted execution of SVM inference, we have implemented a client-server prototype. Starting from the use case, bank fraud detection with a SVM, the needed homomorphic operations for both encryption/decryption and for the encrypted inference using a linear kernel have been implemented.

A. Design

The design of the prototype follows a client-server approach to emulate the client and cloud side for an scenario of outsourced computation in an untrusted environment. The client is responsible of encryption and decryption, so the input data is protected from the moment it leaves the client. The server is responsible of processing the encrypted data, and of finally sending back to the client the encrypted result. Figure 1 depicts the main components of the secure SVM prototype.

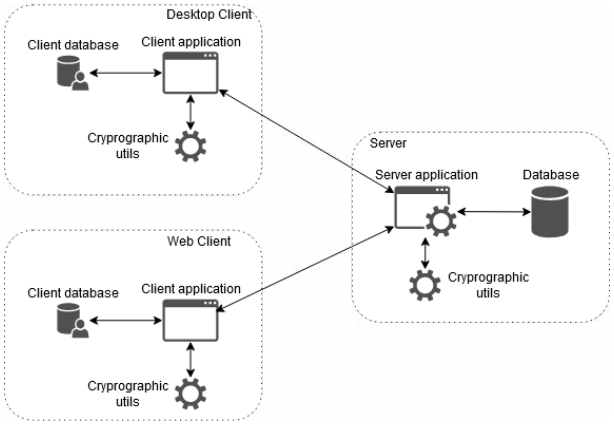


Fig. 1. System architecture

As shown in Figure 1, the prototype is composed of two different clients (one is web based and the other one command line based) and one server. All 3 components make use of a database and of a cryptographic library that implements the needed cryptographic primitives for each component. The clients database stores the user homomorphic cryptographic keys and the transactions data samples that are classified with the Homomorphic SVM Inference. The server database stores users authentication details, users public homomorphic cryptographic keys and parameters, and the SVM model. The clients cryptographic library implements the generation of cryptographic keys and parameters, and also the data packing/unpacking and encryption/decryption primitives, while the server cryptographic library implements the encrypted kernel evaluation primitives. In order to implement the primitives for encrypted processing, the cryptographic libraries make use of Microsoft Seal [12] and Lattigo [13] libraries. For running the homomorphic inference, the following steps must be followed:

1) The *Client* authenticates to the server
2) The *Client* creates a key pair and sends its public part to the server
3) The *Client* packs and encrypts the data samples that wants to infer and sends them to the server
4) The *Server* retrieves the SVM model and public keys from database
5) The *Server* does the Homomorphic SVM Inference and sends the result to the client
6) The *Client* decrypts the result and shows it to the user

B. Data packing

As the CKKS scheme [9] allows for homomorphic approximate additions and Hadamard products between encrypted vectors (see Section II), finding the best way of packing the data before encryption is fundamental to optimize the performance of our encrypted SVM classifier.

With this aim, we have explored two different packing methods [2], which we denote in this work as *column packing* and *flattened packing* respectively. Let $\boldsymbol{X}$ be a matrix of size $N \times M$ where N represents the number of input samples we want to query for detection, and M corresponds to the number of features for each input sample. We also assume that the instantiated CKKS scheme has a packing capacity of P complex slots. We briefly describe these two methods next:

- Column packing: It encodes the columns of $\boldsymbol{X}$ separately in different ciphertexts. This packing naturally fits into the SIMD (Single Instruction, Multiple Data) paradigm, which makes it convenient for those cases where the number of input samples N in the client query is large.
- Flattened packing: We represent the matrix $\boldsymbol{X}$ as a reshape on which all the rows are concatenated into a "flattened" vector of length $N \cdot M$ such as $[\mathsf{row}_1(\boldsymbol{X}), \ldots, \mathsf{row}_N(\boldsymbol{X})]$. Then, we encrypt all the row_i blocks using the minimal possible number of ciphertexts, i.e., each ciphertext has $\lfloor \frac{P}{M} \rfloor$ different rows from $\boldsymbol{X}$.

Contrarily to the former, the use of a *flattened packing* optimizes the storage in those cases on which N is small in relation to the ciphertext packing capacity P.

Homomorphic inference: Regarding the concrete algorithms in the encrypted domain, there are also some important differences for each packing:

- Inference with column packing: It follows the same structure as its counterpart in clear. We homomorphically compute Eq. (2) by means of both $\mathsf{E}_{\mathsf{CKKS}}$.LinHadMult and $\mathsf{E}_{\mathsf{CKKS}}$.Add primitives. It is worth noting that, as a different score is calculated by default in parallel for P complex slots, we directly obtain the inference for P different input samples in each output ciphertext.
- Inference with flattened packing: It requires a more cumbersome algorithm on which not only $\mathsf{E}_{\mathsf{CKKS}}$.LinHadMult and $\mathsf{E}_{\mathsf{CKKS}}$.Add primitives are needed, but also $\mathsf{E}_{\mathsf{CKKS}}$.Rot must be applied to relocate the partial results of the inner product. In this case each output ciphertext contains $\lfloor \frac{P}{M} \rfloor$ different score values.

Tables II and III include a summary of the existing trade-offs between both packing methods. Note that the flattened

TABLE II
COMPUTATIONAL COST FOR EACH PACKING METHOD

Method	Ciphertext operations
Column packing	$\lceil \frac{N}{P} \rceil (M \text{ mult.} + M \text{ add.})$
Flattened packing	$\lceil \frac{N}{\lfloor \frac{P}{M} \rfloor} \rceil (1 \text{ mult.} + \log_2 \lceil M \rceil \text{ add.} + \log_2 \lceil M \rceil \text{ rot.})$

TABLE III
STORAGE REQUIREMENTS FOR EACH PACKING METHOD

Method	# of ciphertexts
Column packing	$M \lceil \frac{N}{P} \rceil$
Flattened packing	$\lceil \frac{N}{\lfloor \frac{P}{M} \rfloor} \rceil$

packing also requires to generate the rotation key matrices rtk which are used for the homomorphic slot rotations.

V. IMPLEMENTATION RESULTS

In this section we show the results obtained in our implementation. All the experiments were conducted using a testbed composed of a physical device with a CPU i7-4710HQ, 12GB of RAM and Ubuntu Desktop 20.04.

In the first place, we evaluate the quality of the inference of our implementation against the inference in the clear. These results are shown in Table IV. In order to test the classification performance, we have considered three different metrics:

- Accuracy: % of correct predictions.
- Precision: % of correctly detected positives among all detected positives.
- Recall: % of correctly detected positives among all positives.

TABLE IV
INFERENCE RESULTS: MACHINE LEARNING CONTEXT

Library	Packing	Accuracy	Precision	Recall
Plain	-	96.6%	5.44%	46.8%
SEAL	Flattened	96.4%	4.39%	32.5%
SEAL	Column	96.6%	3.72%	70.4%
Lattigo	Flattened	96.6%	3.72%	70.4%
Lattigo	Column	96.6%	3.72%	70.4%

TABLE V
INFERENCE RESULTS: PERFORMANCE CONTEXT
($N = 147635, M = 114, P = 4096$)

Library	Packing	$\mathcal{P}_\mathsf{S}$ time (ms)	$\mathcal{P}_\mathsf{C}$ time (ms)	Size (KB)
SEAL	Flattened	4.03	3.66	31.3
SEAL	Column	0.12	1.69	14.0
Lattigo	Flattened	2.34	16.8	41.9
Lattigo	Column	0.03	8.29	18.7

Another important aspect that we have assessed is the runtime performance. In this case, Table V includes:

- Server ($\mathcal{P}_\mathsf{S}$) runtime: the required time to perform the inference function in the server.
- Client ($\mathcal{P}_\mathsf{C}$) runtime: total time taken by the client to encode/decode + encrypt/decrypt the different samples.
- Size: size of the information sent from client to server.

Note that our results consider the execution of several inferences in parallel, so *the included runtimes are normalized to measure the estimated runtime for a single inference of each type of packing*. As the transmission times have not been taken into account, we include *the size of transmitted information per inference*.

VI. CONCLUSIONS AND FUTURE WORK

This work shows that homomorphic encryption is a key enabler technology for migrating highly sensitive process to the cloud. Our proof of concept prototype demonstrates the feasibility of running the fraud detection inference in the untrusted domain, while maintaining the prediction performance and also with a feasible execution performance. As future work, we envision several extensions for our prototype:

- A more complete comparison considering other alternative cryptographic schemes/techniques as Paillier [4], BFV [14], and functional encryption [15]
- Deploy the server in a real cloud environment in order to measure the instantiation costs in a realistic scenario.
- Protect the model in the execution environment. This corresponds to the case on which the server infrastructure provider and the model owner are not the same entity.
- Extend the prototype considering more complex kernels; e.g., polynomial and RBF (Radial Basis Function).
- The implementation in the untrusted domain of the $\mathsf{sign}(\cdot)$ function. Currently, there exist mechanisms for its homomorphic approximation with CKKS [16]. Note this adds an additional degree of protection to the model, as it is harder for the client reverse engineering the model.

REFERENCES

[1] M. Al-Rubaie and J. M. Chang, "Privacy-preserving machine learning: Threats and solutions," *IEEE Secur. Priv.*, vol. 17, no. 2, pp. 49–58, 2019.

[2] A. Vázquez-Saavedra, "Study and applications of homomorphic encryption algorithms to privacy preserving svm inference for a bank fraud detection context," Master's thesis, Telecommunication Engineering School, University of Vigo, 2021. [Online]. Available: http://castor.det.uvigo.es:8080/xmlui/handle/123456789/533

[3] Homomorphic encryption standardization. [Online]. Available: https://homomorphicencryption.org/

[4] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *EUROCRYPT*, ser. LNCS, vol. 1592. Springer, 1999, pp. 223–238.

[5] R. L. Lagendijk, Z. Erkin, and M. Barni, "Encrypted signal processing for privacy protection: Conveying the utility of homomorphic encryption and multiparty computation," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 82–105, 2013.

[6] S. Park, J. Byun, J. Lee, J. H. Cheon, and J. Lee, "He-friendly algorithm for privacy-preserving SVM training," *IEEE Access*, vol. 8, pp. 57 414–57 425, 2020.

[7] A. Pedrouzo-Ulloa, J. R. Troncoso-Pastoriza, and F. Pérez-González, "Number theoretic transforms for secure signal processing," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 5, pp. 1125–1140, 2017.

[8] J. R. Troncoso-Pastoriza, A. Pedrouzo-Ulloa, and F. Pérez-González, "Secure genomic susceptibility testing based on lattice encryption," in *IEEE ICASSP*. IEEE, 2017, pp. 2067–2071.

[9] J. H. Cheon, A. Kim, M. Kim, and Y. S. Song, "Homomorphic encryption for arithmetic of approximate numbers," in *ASIACRYPT*, ser. LNCS, vol. 10624. Springer, 2017, pp. 409–437.

[10] Z. Wu. (2020) Kaggle — IEEE Fraud Detection (EDA). [Online]. Available: https://blog.csdn.net/Tinky2013/article/details/104215953

[11] I. Steinwart and A. Christmann, "Support vector machines," in *Information science and statistics*, 2008.

[12] "Microsoft SEAL (release 3.0)," http://sealcrypto.org, Oct. 2018, microsoft Research, Redmond, WA.

[13] C. Mouchet, J.-P. Bossuat, J. Troncoso-Pastoriza, and J. Hubaux, "Lattigo: a multiparty homomorphic encryption library in go," 2020.

[14] J. Fan and F. Vercauteren, "Somewhat practical fully homomorphic encryption," *IACR Cryptol. ePrint Arch.*, vol. 2012, p. 144, 2012.

[15] T. Marc, M. Stopar, J. Hartman, M. Bizjak, and J. Modic, "Privacy-enhanced machine learning with functional encryption," in *ESORICS*, ser. LNCS, vol. 11735. Springer, 2019, pp. 3–21.

[16] E. Lee, J. Lee, J. No, and Y. Kim, "Minimax approximation of sign function by composite polynomial for homomorphic comparison," *IACR Cryptol. ePrint Arch.*, vol. 2020, p. 834, 2020.

http://doi.org/10.18239/jornadas_2021.34.55

Intercambio de clave multiusuario en anillos de grupo

María Dolores Gómez Olvera
Universidad de Almería
Crta. Sacramento S/N, Cañada de San Urbano (Almería)
gomezolvera@ual.es
0000-0001-9855-6936

Juan Antonio López Ramos
Universidad de Almería
Crta. Sacramento S/N, Cañada de San Urbano (Almería)
jlopez@ual.es
0000-0002-2263-2178

Blas Torrecillas Jover
Universidad de Almería
Crta. Sacramento S/N, Cañada de San Urbano (Almería)
btorreci@ual.es
0000-0002-0992-9911

***Resumen*—La Criptografía que utilizamos actualmente para asegurar nuestras comunicaciones, podría verse comprometida en los próximos años. Se están produciendo cambios en los estándares de los procotolos de clave pública, ante la mejora de los métodos de criptoanálisis, y el posible advenimiento de ordenadores cuánticos suficientemente potentes para implementar el algoritmo de Shor o sus variantes. En este trabajo, que constituye un resumen de [3], proponemos un ambiente alternativo para ofrecer seguridad en un contexto post-cuántico, el álgebra no conmutativa. En particular, proponemos un anillo de grupo torcido mediante un cociclo, y protocolos de acuerdos de clave para dos, y también de varios usuarios; estos últimos se enfrentan a problemas específicos que tenemos en cuenta a la hora de proponer un protocolo post-cuántico para ellos.**

***Index Terms*—Álgebra no conmutativa, Intercambio de clave, Anillo de grupo, Criptografía de clave pública, Multiusuario**

Tipo de contribución: *Investigación ya publicada (límite 2 páginas)*

I. Introducción

El problema del intercambio de clave es fundamental en el ámbito de la Seguridad de la Información. Los protocolos que utilizamos actualmente podrían verse comprometidos en un tiempo, y por ello se buscan activamente alternativas a los mismos, ante el posible advenimiento de la computación cuántica.

El primer intercambio de clave, propuesto por W. Diffie y M. Hellman en 1976, ha sido y continua siendo central en nuestras comunicaciones diarias. Este intercambio, que permite el acuerdo de una clave común entre dos usuarios, marcó el inicio de la Criptografía de Clave Pública. Y pronto se hizo relevante la cuestión de cómo sería posible generalizar esta idea con el objetivo de asegurar las comunicaciones entre varios usuarios. En 1994 hubo una propuesta de Burmester y Desmedt [1]; finalmente en 1996, M. Steiner, G. Tsudik y M. Waidner dieron una solución firme a esta cuestión [4].

El acuerdo de una clave grupal se enfrenta a cuestiones que no cabe plantearse en un intercambio entre 2 usuarios, entre los que podemos citar: posible pérdida de mayor cantidad de información, ya que necesariamente se envían más mensajes a través de la red (medio inseguro) para conseguir esta clave secreta; el impedimento a usuarios previos del grupo al acceso a información nueva compartida en el mismo una vez ellos han salido; y también a los usuarios nuevos, el acceso a la información compartida previamente a su incorporación al grupo.

La solución para estos dos últimos problemas podría ser un refresco de clave cada vez que hay una modificación en el grupo; también es deseable realizarlo de manera periódica, aunque los usuarios se mantengan estables. Pero este refresco de clave también supone, de nuevo, un reto. Tanto en el establecimiento de clave inicial como en el refresco es deseable la mayor eficiencia posible. En el caso de n usuarios, se torna más complicado poder mantener un protocolo eficiente. En este sentido, el protocolo de [4] mejoraba la eficiencia del de Burmester y Desmedt, ya que su refresco de clave es mucho más breve y eficiente, manteniendo la seguridad.

Volviendo a los problemas que encontramos actualmente en la Criptografía de Clave Pública, el principal de ellos es la posible implementación del algoritmo de Shor o sus variaciones, que pondría en jaque la seguridad que usamos hoy en día.

Actualmente se están proponiendo ambientes alternativos en los que basar la seguridad de nuestras comunicaciones. Uno de ellos es el álgebra no conmutativa. Se ha venido proponiendo que los anillos de grupo pueden ser una buena base para realizar un intercambio de clave [2], [5], [6], [7], si bien es cierto que algunos han presentado problemas.

Este trabajo [3] va en esta dirección, proponiendo algunas variaciones para evitar las dificultades a las que se encuentran algunos protocolos citados recientemente, y añadiendo además la propuesta de un intercambio de clave en grupo. Además, demostramos que su seguridad es computacionalmente equivalente al intercambio entre dos usuarios.

II. Anillos de grupo torcidos

El intercambio que proponemos utiliza una variación del problema de la descomposición (DP), y un anillo de grupo, pero en este caso, torcido mediante un cociclo. Describimos en primer lugar la estructura.

Sea G un grupo multiplicativo, K un anillo conmutativo unitario y α un 2-cociclo.

$$\alpha : G \times G \longrightarrow K$$

El **anillo de grupo torcido** $K^{\alpha}G$ está definido como

$$\sum_{g_i \in G} a_i g_i$$

con $a_i \in K$, $a_i = 0$ para todo i salvo un número finito.

La suma de los elementos en $K^{\alpha}G$ se define como:

$$\Big(\sum_{g_i \in G} a_i g_i\Big) + \Big(\sum_{g_i \in G} b_i g_i\Big) = \sum_{g_i \in G} (a_i + b_i) g_i$$

Y la multiplicación torcida mediante un cociclo se define de la siguiente forma:

$$\Big(\sum_{g_i \in G} a_i g_i\Big) \cdot \Big(\sum_{g_i \in G} b_i g_i\Big) = \sum_{g_i \in G} \Big(\sum_{g_j g_k = g_i} a_j b_k \ \alpha(g_j, g_k)\Big) g_i$$

III. Intercambio de clave

El intercambio entre dos usuarios es el siguiente: Sea h un elemento cualquiera público de un anillo de grupo R, que escinda como $R = R_1 \oplus R_2$, con R_1 y R_2 tales que los elementos de ambos 'conmuten' a través de una aplicación *. El intercambio de clave entre Alicia y Bruno es el siguiente:

1. Alicia elige un par de elementos $s_A = (g_1, k_1)$, con $g_1 \in R_1, k_1 \in A_2 \subseteq R_2$.
2. Bruno elige un par de elementos $s_B = (g_2, k_2)$, con $g_2 \in R_1, k_2 \in A_2 \subseteq R_2$.
3. Alicia envía a Bruno $p_A = g_1 h k_1$, y Bruno envía a Alicia $p_B = g_2 h k_2$.
4. Alicia calcula $K_A = g_1 p_B k_1^*$, y Bruno calcula $K_B = g_2 p_A k_2^*$, siendo esta su clave compartida.

$$K = K_A = K_B$$

donde k_1^*, k_2^* son elementos que dependen de k_1, k_2 y el cociclo α; y donde A_2 es un subconjunto de R_2.

IV. Intercambio de clave multiusuario

El intercambio de clave para n usuarios es el siguiente. Sea $h \in R$. Para $i = 1, ..., n$, el usuario U_i tiene un par secreto $s_i = (g_i, k_i)$, donde $g_i \in R_1$ and $k_i \in A_2 \subseteq R_2$. Sea $\phi(s_i, h) = g_i h k_i$, la multiplicación por ambos lados. Y sea $s_i^* = (g_i, k_i^*)$.

1. Para $i = 1, ..., n$, el usuario U_i enviar al usuario U_{i+1} el mensaje
$$\{C_i^1, C_i^2, ..., C_i^{i+1}\},$$
donde $C_1^1 = h$, $C_1^2 = g_1 h k_1$ y
   - para $i > 1$ par, $C_i^j = \phi(s_i, C_{i-1}^j)$, cuando $j < i$, $C_i^i = C_{i-1}^i$, $C_i^{i+1} = \phi(s_i^*, C_{i-1}^i)$,
   - para $i > 1$ impar, $C_i^j = \phi(s_i^*, C_{i-1}^j)$, cuando $j < i$, $C_i^i = C_{i-1}^i$, $C_i^{i+1} = \phi(s_i, C_{i-1}^i)$.
2. El usuario U_n calcula $\phi(s_n, C_{n-1}^n)$ si n es impar y $\phi(s_n^*, C_{n-1}^n)$ si n es par.
3. El usuario U_n comparte
$$\{C_n^1, C_n^2, ..., C_n^n\}.$$
4. Cada usuario U_i calcula $\phi(s_i, C_n^i)$ si n es impar, o $\phi(s_i^*, C_n^i)$ si n es par, y obtiene la clave secreta compartida.

Puede comprobarse en que para $n = 2$, este protocolo coincide con el anteriormente propuesto para dos usuarios, y que por tanto se trata de una generalización natural del mismo.

En el caso del refresco de clave que proponemos, no es necesario hacer el mismo número de rondas que en el intercambio de clave inicial para n usuarios que acabamos de proponer. Es suficiente con una única ronda, y demostramos que su seguridad se sigue manteniendo, siendo también computacionalmente equivalente a la de dos usuarios. Con ello, mejora la propuesta de Burmester y Desmedt, siguiendo en la línea de [4], pero trasladado al contexto del álgebra no conmutativa, que en este momento ofrece más garantías.

V. Conclusiones

En este trabajo se proponen nuevos protocolos de intercambio de clave para 2 y para n usuarios, teniendo en cuenta las particularidades de cada caso, y con nuevas características con respecto a otros protocolos existentes, que suponen una ventaja con respecto a posibles ataques.

Como ejemplo más específico de estructura, proponemos el anillo de grupo $R = GF(2^n)^{\alpha} D_{2m}$, siendo $GF(2^n)$ el cuerpo finito de 2^n elementos, D_{2m} el grupo diédrico de $2m$ elementos, y α el 2-cociclo

$$\alpha : D_m \times D_m \longrightarrow GF(2^n)^*$$

con $\alpha(x^i, x^j y^k) = 1$ y $\alpha(x^i y, x^j y^k) = \epsilon^j$, para todo k; con ϵ una raíz primitiva de la unidad que genera $GF(2^n)$. Además, hemos implementado parcialmente los protocolos mencionados, y comprobado que ciertos ataques conocidos no suponen una amenaza para los mismos.

Referencias

[1] M. Burmester, Y. Desmedt, "A Secure and Efficient Conference Key Distribution System", en *Advances in Cryptology - EUROCRYPT '94 Lecture Notes in Computer Science*, 1994.

[2] M. Eftekhari, "A Diffie-Hellman key exchange protocol using matrices over group rings", en *Groups Complex. Cryptol.*, vol. 4, n. 1, pp. 167-176, 2012.

[3] M. D. Gómez Olvera, J. A. López Ramos, B. Torrecillas Jover, "Public Key Protocols over Dihedral Group Rings", en *Symmetry*, vol. 11, n. 8, 1019, 2019.

[4] M. Steiner, G. Tsudik, M. Waidner, "Diffie-Hellman key distribution extended to group communication", en *SIGSAC Proceedings of the 3rd ACM Conference on Computer and Communications Security*, pp. 31-37.

[5] I. Gupta, A. Pandey, M. Kant Dubey, "A Key Exchange Protocol using Matrices over Group Rings", en *Asian-European Journal of Mathematics*, https://doi.org/10.1142/S179355711950075X, 2018.

[6] M. Habeeb, D. Kahrobaei, C. Koupparis, C. Shpilrain, "Public key exchange using semidirect product of (semi)groups", en *LNCS 2013, Lecture Notes Comp. Sc.*, vol. 7954, pp. 475-486, 2013.

[7] D. Kahrobaei, C. Koupparis, V. Shpilrain "Public key exchange using matrices over group rings", en *Groups Complex. Cryptol.*, vol. 5, n. 1, pp. 97-115, 2013.

Sesión de Investigación B4: Criptografía
http://doi.org/10.18239/jornadas_2021.34.56

Implementación de un protocolo postcuántico de intercambio de clave en grupo seguro en el *Quantum Random Oracle Model basado en Kyber*

José Ignacio Escribano Pablos*†
*BBVA Next Technologies, Av. de Manoteras 44, 28050 Madrid
†MACIMTE, Universidad Rey Juan Carlos, 28933 Móstoles, Madrid
https://orcid.org/0000-0002-0079-642X

***Resumen*—La computación cuántica plantea fascinantes retos para la criptografía actual, al amenazar la seguridad de muchos esquemas y protocolos ampliamente utilizados hoy en día. Para adaptarse a esta nueva realidad es necesaria la estandarización de distintas construcciones criptográficas susceptibles de sustituir a las más utilizadas en la actualidad. En esta línea, el mayor esfuerzo para estandarizar los nuevos esquemas criptográficos resistentes a adversarios cuánticos lo está realizando el NIST en un concurso específico en el que se evalúan este tipo de herramientas. En este artículo, implementamos un protocolo postcuántico de intercambio de clave en grupo seguro en el llamado *Quantum Random Oracle Model*. Nuestro diseño está basado en Kyber, uno de los finalistas del concurso del NIST. Se demuestra experimentalmente que este protocolo es más rápido que otras alternativas para realizar la misma tarea cuya seguridad se demuestra además en modelos menos ambiciosos, en los que, por ejemplo, la interacción con las funciones *hash* del sistema se supone exclusivamente clásica.**

***Index Terms*—Criptografía Postcuántica, Intercambio de Clave, Intercambio de Clave en Grupo, Kyber, Compilador, Quantum Random Oracle Model**

Tipo de contribución: *Investigación original*

I. Introducción

La computación cuántica es una apasionante área de investigación que pone su foco en el diseño y desarrollo de arquitecturas que usan tecnología cuántica, así como en el diseño de algoritmos que puedan ser ejecutados en ellas. Los ordenadores cuánticos tendrán la capacidad de resolver eficientemente problemas que los ordenadores clásicos tardan demasiado tiempo en abordar y esto tiene graves implicaciones en criptografía. Por ejemplo, el algoritmo de Grover es un algoritmo cuántico que permite buscar en una secuencia no ordenada de n datos en tiempo $\mathcal{O}(\sqrt{n})$, haciendo que sea necesario duplicar la longitud de clave para alcanzar la seguridad actual cuando se usan ciertas herramientas de criptografía simétrica (esquemas de cifrado y funciones *hash*). Por otro lado, el algoritmo de Shor es un algoritmo cuántico para descomponer un número n en sus factores primos en un tiempo $\mathcal{O}(\log^3 n)$. Esto hace que sea posible factorizar un número en tiempo polinomial (en el número de bits necesario para representarlo). Esquemas como el cifrado RSA fundamentan su seguridad en la dificultad de este problema, por lo que no deberán ser utilizados en presencia de adversarios con acceso a ordenadores cuánticos. De hecho, el alcance del algoritmo de Shor es muy amplio y de hecho deja en evidencia la seguridad de numerosas construcciones ampliamente utilizadas más allá de RSA, como aquellas que utilizan el problema del logaritmo discreto en grupos asociados a cuerpos finitos o curvas elípticas. En este sentido es indudable que, en gran medida, la criptografía asimétrica actual no será segura frente a adversarios cuánticos.

Nadie conoce con exactitud cuándo se conseguirá desarrollar ordenadores cuánticos para romper de facto los esquemas criptográficos mencionados. Michele Mosca estimó en 2017 que hay una posibilidad entre 6 de que RSA-2048 sea roto en una década y un 50 % de posibilidades de que sea roto en los próximos 15 años [1]. En cualquier caso, la comunidad criptográfica internacional asume que es imprescindible disponer a medio plazo de algoritmos y protocolos que sean resistentes frente a ataques implementados en arquitecturas cuánticas.

El objetivo de la llamada *criptografía postcuántica* es diseñar e implementar esquemas y protocolos resistentes a ataques que utilicen computación cuántica. Estas nuevas construcciones basan su seguridad en problemas matemáticos que se cree que no pueden ser resueltos por ordenadores cuánticos (ni clásicos) de manera eficiente. Hay distintas aproximaciones a la criptografía postcuántica, cada una con sus ventajas y desventajas. Habitualmente se establece una clasificación en cinco áreas: criptografía basada en retículos, multivariante, basada en *hashes*, basada en códigos y basada en grupos de isogenias de curvas elípticas supersingulares.

Como ya hemos mencionado, un gran impulso al desarrollo e implementación de estas nuevas construcciones viene marcado por el NIST a través de su concurso de estandarización [2], que comenzó en 2017. En el momento de escribir este artículo ya se conocen los finalistas y las alternativas en el caso de que los finalistas sufran ataques graves. Los finalistas en la categoría de cifrado de clave pública y establecimiento de clave son Classic McEliece [3] (códigos), CRYSTALS-Kyber [4] (retículos), NTRU [5] (retículos) y SABER [6] (retículos). En lo que respecta a firma digital, se han seleccionado las propuestas CRYSTALS-Dilithium [7] (retículos), FALCON [8] (retículos) y Rainbow [9] (multivariante).

Al margen del concurso del NIST, se han propuesto alternativas postcuánticas para esquemas y protocolos que no están cubiertos por el mismo (que aborda esencialmente cifrado, firma e intercambio de clave en el escenario 2-parte). Uno de ellos, es el intercambio autenticado de clave en grupo (GAKE, por sus siglas en inglés), que consiste en una generalización del intercambio de clave entre dos partes a un grupo de $n > 2$ usuarios. En este sentido, se han propuesto distintos protocolos de GAKE postcuántico basados principalmente en

retículos e isogenias [10], [11]. Desafortunadamente, la gran mayoría de estas construcciones se plantean con un enfoque teórico, sin explorar de manera rigurosa su viabilidad práctica.

Contribución. En este artículo, analizamos la viabilidad de una propuesta reciente para GAKE post-cuántico desarrollada en [12], describiendo una implementación que hemos realizado y analizando distintos resultados experimentales obtenidos de su ejecución. En concreto, partimos del trabajo de [12], que implementa un GAKE seguro en el *Quantum Random Oracle Random Model* (QROM), y se construye usando como pieza esencial uno de los esquemas los finalistas del concurso del NIST, Kyber [4]. Esta construcción es, de hecho, el primer protocolo demostrado seguro en dicho modelo que se apoya en un diseño dentro de los finalistas. El objetivo del artículo es comparar si las modificaciones necesarias para hacer seguro un protocolo en el QROM introducen un empeoramiento del rendimiento, comparadas con sus alternativas en el *Random Oracle Model* (ROM).

Organización. En la Sección II se describe el protocolo postcuántico de intercambio de clave en grupo basado en Kyber, detallando las primitivas de las que depende. En la Sección III, se describen los detalles de implementación del protocolo, así como los resultados experimentales y cuál es la caída de rendimiento frente al mismo protocolo seguro en el *Random Oracle Model*. Por último, en la Sección IV se describen las conclusiones y trabajo futuro.

I-A. ROM vs. QROM

En esta sección, describimos brevemente los modelos de demostración ROM y QROM. El *Random Oracle Model* (ROM) es un modelo para construir demostraciones de seguridad. En el ROM, todas las partes tienen acceso a un oráculo aleatorio $\mathcal{O}$. Cualquier parte puede hacer una consulta $x \in \{0,1\}^*$ al oráculo $\mathcal{O}$, que devolverá $\mathcal{O}(x) \in \{0,1\}^*$, siendo cada bit de la respuesta elegido independiente e uniformemente al azar. Si x ya ha sido consultado con anterioridad, el oráculo aleatorio devolverá el mismo resultado que en la consulta previa. Los protocolos que se demuestran seguros en el ROM sustituyen el oráculo aleatorio por una función *hash*.

El *Quantum Random Oracle* (QROM) es una generalización del ROM, en la que se asume que el adversario tiene acceso a un oráculo aleatorio cuántico en el que puede hacer una petición y obtener una superposición cuántica de estados. Esto permite a un adversario cuántico tener una ventaja considerable sobre un adversario clásico. Las partes honestas se suponen que se mantienen *clásicas*.

Por último, las demostraciones en el QROM son también válidas en el ROM, pero el recíproco no es cierto [13].

II. Protocolo postcuántico

Un protocolo autenticado de intercambio de claves en grupo (GAKE) tiene por objetivo permitir a los miembros de un grupo (que se reconocen a través de claves de firma o contraseñas) acordar una clave secreta común para posteriormente cifrar las comunicaciones con ella. Esta recibe el nombre de clave de sesión y suele ser efímera, por lo que hemos de asumir que las claves acordadas deben ser renegociadas cada cierto período de tiempo. En este artículo, se supone que la autenticación de los miembros legítimos del grupo se autentican a través de un esquema de firma digital cuyas claves están distribuidas y certificadas a priori.

El protocolo GAKE basa su funcionamiento en dos primitivas básicas: Kyber, un Key Encapsulation Mechanism (KEM) IND-CCA finalista del concurso de estandarización del NIST y el compilador de Abdalla et al. [14], que permite transformar un protocolo autenticado de intercambio de claves (AKE, por sus siglas en inglés) entre dos partes en un GAKE.

Tabla I
Resumen de los algoritmos empleados en cada una de las variantes del protocolo.

Algoritmo	QROM	ROM
PKE IND-CCA	`Kyber.CPA'` = (KeyGen, Enc, Dec)	
KEM IND-CCA	$\texttt{Kyber}^{\not\perp}$ (*Algs.* 1 y 2)	`Kyber` (KeyGen, Encaps, Decaps)
2-AKE	Init (*Alg.* 6)	initA (*Alg.* 3)
	Der_{resp} (*Alg.* 7)	sharedB (*Alg.* 4)
	Der_{init} (*Alg.* 8)	sharedA (*Alg.* 5)
PKE IND-CCA	*Algs.* 6, 7 y 8 de [15]. `Kyber.PKE` = (KeyGen, Enc, Dec)	
Esquema de compromiso	Aplicación directa del PKE IND-CCA	

II-A. Kyber

Kyber [4] es un KEM IND-CCA finalista del concurso de estandarización del concurso del NIST, que pertenece, junto a Dilithium, a la *suite* criptográfica CRYSTALS. Kyber basa su seguridad en problemas sobre retículos, en concreto, en la hipótesis de que es difícil resolver el problema de Module Learning With Errors (MLWE). Kyber tiene un espacio de mensajes $\mathcal{M} = \{0,1\}^{256}$. Kyber nace como un esquema de cifrado de clave pública (PKE, por sus siglas en inglés) IND-CPA denominado `Kyber.CPA'`, que es transformado en un KEM IND-CCA a través de una transformación similar a la transformación de Fujisaki-Okamoto [16], pero adaptada para los casos en los que se producen errores de descifrado. Esta versión recibe el nombre de `Kyber` y los algoritmos de generación de claves, encapsulación y decapsulación se pueden encontrar en los algoritmos 1, 4 y 5 de [4].

Los autores de Kyber demuestran que tanto el esquema `Kyber.CPA'` como `Kyber` son seguros en el ROM y sólo dan unas cotas de seguridad en el QROM. En [12], partiendo del esquema `Kyber.CPA'`, aplican una transformación de [17], llamada $FO_m^{\not\perp}$, para obtener un KEM similar a `Kyber` que es seguro en el QROM, denominado $\texttt{Kyber}^{\not\perp}$. Los *Algs.* 1 y 2 muestran cómo se realiza la encapsulación y decapsulación. La generación de claves es similar a la de `Kyber.CPA'`, que se puede encontrar en [4].

Algoritmo 1: $\texttt{Kyber}^{\not\perp}\texttt{.Encaps}(pk)$

$m \xleftarrow{\$} \mathcal{M}$
$c := \texttt{Kyber.CPA}'\texttt{.Enc}(pk, m; \mathsf{G}(m))$
$\mathsf{k} := \mathsf{H}(\mathrm{m})$
devolver (k, c)

Kyber destaca por su versatilidad, y es posible construir a partir de `Kyber` un esquema PKE IND-CCA y un AKE 2-parte de sólo dos mensajes. El PKE, que se denomina `Kyber.PKE`, se obtiene a partir de una aproximación

Algoritmo 2: $\texttt{Kyber}^{\not\perp}\texttt{.Decaps}(sk, c)$

$m' := \texttt{Kyber.CPA}'\texttt{.Dec}(sk, c)$
si $m' = \perp$ **o** $\texttt{Kyber.CPA}'\texttt{.Enc}(pk, m'; \mathsf{G}(m')) \neq c$
entonces
  devolver $\mathsf{k} := \mathsf{H}_r(c)$
en otro caso
  devolver $\mathsf{k} := \mathsf{H}(m')$

KEM/DEM [18]. Los autores de Kyber proponen emplear como DEM AES-OCB, AES-GCM o ChaCha20-Poly1305 [15]. Los algoritmos concretos se pueden consultar en el Anexo A de [15].

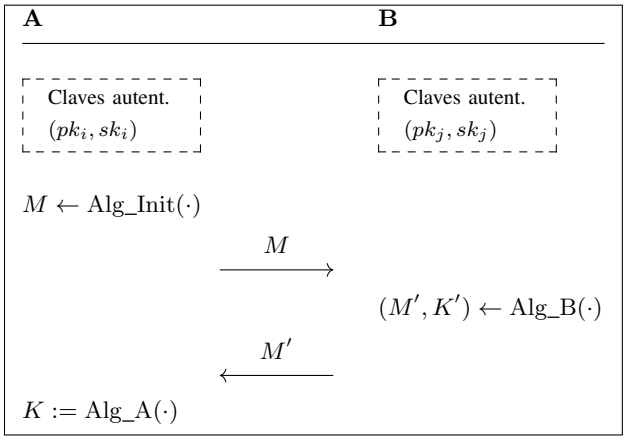


Figura 1. Estructura de un AKE 2-parte de dos mensajes.

Algoritmo 3: $\text{initA}(pk_j)$

$(pk, sk) \leftarrow \texttt{Kyber.KeyGen}()$
$(c_j, K_j) \leftarrow \texttt{Kyber.Encaps}(pk_j)$
devolver $(M := (pk, c_j), sk, K_j)$

Algoritmo 4: $\text{sharedB}(pk, c_j, pk_i, sk_j)$

$(C, K) \leftarrow \texttt{Kyber.Encaps}(pk)$
$(c_i, K_i) \leftarrow \texttt{Kyber.Encaps}(pk_i)$
$K'_j := \texttt{Kyber.Decaps}(sk_j, c_j)$
$K := \mathsf{H}(K, K_i, K'_j)$
devolver $(K, M' := (c, c_i))$

En [4], se propone un AKE 2-parte de sólo dos mensajes, aplicando `Kyber`. La estructura del protocolo se puede ver en la *Fig.* 1. Las 2 partes tienen claves público/privadas de autenticación y una de las partes inicia el intercambio de clave, enviando un mensaje M a la otra parte, que calcula la clave secreta K' y un mensaje M' que le devuelve a la parte iniciadora para que obtenga la clave K, siendo $K = K'$. Los algoritmos Alg_Init, Alg_B y Alg_A se corresponden con los algoritmos initA (*Alg.* 3), sharedB (*Alg.* 4) y sharedA (*Alg.* 5), respectivamente. Como en el caso anterior, el AKE sólo es demostrado seguro en el ROM.

En [12] se aplica la transformación FO_{AKE} de [17] para convertir el esquema $\texttt{Kyber.CPA}'$ en un AKE 2-parte, seguro en el QROM. El AKE resultante también es de dos

Algoritmo 5: $\text{sharedA}(sk, sk_i, c, c_i, K_j)$

$K' := \texttt{Kyber.Decaps}(sk, c)$
$K'_i := \texttt{Kyber.Decaps}(sk_i, c_i)$
$K := \mathsf{H}(K', K'_i, K_j)$
devolver K

mensajes con la estructura de la *Fig.* 1, pero los algoritmos son distintos. En este caso, los algoritmos Alg_Init, Alg_B y Alg_A se corresponden con los algoritmos Init (*Alg.* 6), Der_{resp} (*Alg.* 7) y Der_{init} (*Alg.* 8), respectivamente. En estos dos últimos algoritmos se emplean funciones *hash* (modeladas como oráculos aleatorios) como H'_R, H'_{L2} o H'_{L2}. En [17] apuntan que estas funciones se pueden implementar usando una función pseudoaletoria.

Notar, que aunque a primera vista, puedan parecer AKE totalmente distintos, los mensajes M y M' tienen el mismo tamaño. En el caso de M, se el tamaño es de una clave pública y un texto cifrado y en M' es la longitud de 2 textos cifrados.

Algoritmo 6: $\text{Init}(sk_i, pk_j)$

$m_j \xleftarrow{\$} \mathcal{M}$
$c_j := \texttt{Kyber.CPA}'\texttt{.Enc}(pk_j, m_j; \mathsf{G}(m_j))$
$(\tilde{sk}, \tilde{pk}) \leftarrow \texttt{Kyber.CPA}'\texttt{.KeyGen}()$
$M := (\tilde{pk}, c_j)$
$st := (\tilde{sk}, m_j, M)$
devolver (M, st)

Algoritmo 7: $\text{Der}_{\text{resp}}(sk_j, pk_i, M)$

$(\tilde{pk}, c_j) := M$
$m_i, \widetilde{m} \xleftarrow{\$} \mathcal{M}$
$c_i := \texttt{Kyber.CPA}'\texttt{.Enc}(pk_i, m_i; \mathsf{G}(m_i))$
$\widetilde{c} := \texttt{Kyber.CPA}'\texttt{.Enc}(\tilde{pk}, \widetilde{m}; \mathsf{G}(\widetilde{m}))$
$M' := (c_i, \widetilde{c})$
$m'_j := \texttt{Kyber.CPA}'\texttt{.Dec}(sk_j, c_j)$
si $m'_j = \perp$ **o** $c_j \neq \texttt{Kyber.CPA}'\texttt{.Enc}(pk_j, m'_j; \mathsf{G}(m'_j))$
entonces
  $K' := \mathsf{H}'_R(m_i, c_j, \widetilde{m}, i, j, M, M')$
en otro caso
  $K' := \mathsf{H}(m_i, m'_j, \widetilde{m}, i, j, M, M')$
devolver (M', K')

Los diferentes algoritmos de los esquemas QROM y ROM se pueden encontrar en la Tabla I.

Los autores de Kyber han instanciado el KEM con tres niveles de seguridad. Estas versiones de Kyber reciben el nombre de `KYBER512` (menor seguridad), `KYBER768` (versión recomendada) y `KYBER1024` (mayor seguridad). Nos referiremos a ellas como `512`, `768` y `1024` por brevedad a lo largo del resto del artículo. Los detalles de cómo se obtienen los parámetros de cada versión se pueden encontrar en [4], [19].

II-B. Compilador de Abdalla et al.

El compilador de Abdalla [14] et al. es la otra construcción requerida del protocolo. El compilador convierte un AKE 2-parte seguro en un GAKE seguro, sin necesidad de hipótesis adicionales. Su funcionamiento se basa en disponer a los

Algoritmo 8: $\mathrm{Der}_{\mathrm{init}}(sk_i, pk_j, M', st)$

$(c_i, \widetilde{c}) := M'$
$(\widetilde{sk}, m_j, M) := st$
$m'_i := \texttt{Kyber.CPA}'\texttt{.Dec}(sk_i, c_i)$
$\widetilde{m}'_i := \texttt{Kyber.CPA}'\texttt{.Dec}(\widetilde{sk}, \widetilde{c})$
si $m'_i = \perp$ **o** $c_i \neq \mathit{Kyber.CPA}'\mathit{.Enc}(pk_i, m'_i; \mathsf{G}(m'_i))$
 entonces
  si $\widetilde{m}' = \perp$ **entonces**
   $K := \mathsf{H}'_{L1}(c_i, m_j, \widetilde{c}, i, j, M, M')$
  en otro caso
   $K := \mathsf{H}'_{L2}(c_i, m_j, \widetilde{m}', i, j, M, M')$
si no, si $\widetilde{m}' = \perp$ **entonces**
 $K := \mathsf{H}'_{L3}(m'_i, m_j, \widetilde{c}, i, j, M, M')$
en otro caso
 $K := \mathsf{H}(m'_i, m_j, \widetilde{m}', i, j, M, M')$
devolver K

participantes en el protocolo en forma de anillo, de tal forma que un participante sólo se comunica con los participantes a su izquierda y derecha. El compilador sólo añade 2 rondas de comunicación adicionales al AKE 2-parte, independientemente del número de participantes. El compilador requiere de un esquema de compromiso que sea no maleable y no interactivo (*perfectly binding* para múltiples compromisos).

El modelo adversarial considera a un adversario que tiene control total sobre la red de comunicaciones que le permite borrar, escuchar, modificar y retrasar mensajes a su elección. Un usuario puede ejecutar un número polinomial de instancias del compilador, que son modeladas mediante una serie de variables. Entre ellas, destacan `sk` que designa a la clave de sesión (por defecto, con valor distinguido `NULL`), `sid` que describe un identificador público de la clave de sesión `sk`, `term` que indica si la instancia ha terminado. Por otro lado `acc` que indica si la instancia ha aceptado el resultado de la ejecución y `pid` que guarda los participantes en esa instancia del compilador.

II-C. Descripción del protocolo

El protocolo propuesto en [12], construye un esquema de compromiso y un AKE 2-parte seguros en el QROM para obtener un protocolo de intercambio de clave en grupo de 4 rondas (2 rondas del AKE 2-parte y 2 rondas introducidas por el compilador de Abdalla et al.).

El esquema de compromiso se puede obtener a través de cualquier esquema PKE IND-CCA, como se indica en [14]. En concreto se emplea el esquema `Kyber.PKE`. El AKE 2-parte (denominado `Kyber.2AKE`) se obtiene de los algoritmos de la Sección II-A.

El protocolo completo se puede encontrar en la *Fig.* 2. A lo largo de una ejecución exitosa con n participantes se retransmiten $2n$ mensajes al resto de participantes y $2n$ mensajes se envían punto a punto, es decir, entre los participantes que se tienen a la izquierda y a la derecha del anillo.

III. Implementación del protocolo

Los autores de Kyber proporcionan código de su esquema en GitHub[1] escrito en C99 [20]. En el repositorio, Kyber ha sido instanciado con funciones *hash* derivadas de Keccak

[1]https://github.com/pq-crystals/kyber

- **Ronda 1-2**.
  - Se ejecuta el intercambio de clave 2-parte $\texttt{Kyber.2AKE}(U_i, U_{i+1})$ obteniendo cada participante U_i dos claves $\overrightarrow{K_i}$ y $\overleftarrow{K_i}$ compartidas con U_{i+1} y U_{i-1}.
- **Ronda 3**.
  - Cada U_i calcula $X_i := \overrightarrow{K_i} \oplus \overleftarrow{K_i}$ y elige un valor al azar r_i para calcular el compromiso
    $$C_i = \texttt{Kyber.PKE}(i, X_i; r_i).$$
  - Cada U_i transmite el valor $M_i^1 := (U_i, C_i)$.
- **Ronda 4**.
  - Cada U_i transmite $M_i^2 := (U_i, X_i, r_i)$.
  - Cada U_i comprueba que $\bigoplus_{i=1}^{n} X_i = 0$ y la corrección de los compromisos. Si al menos una de las comprobaciones falla, poner $\texttt{acc}_i := \texttt{false}$ y terminar la ejecución del protocolo.
  - Cada U_i define $K_i := \overleftarrow{K_i}$ y calcula los valores
    $$K_{i-j} := \overleftarrow{K_i} \oplus X_{i-1} \oplus \cdots \oplus X_{i-j}$$
    para $j = 1, \ldots, n-1$, que definen la clave maestra
    $$K := (K_1, \ldots, K_n, \texttt{pid}_i)$$
    y asigna
    $$\texttt{sk}_i := H(K), \texttt{sid}_i := F(K) \text{ y } \texttt{acc}_i := \texttt{true}$$
    donde $F, H\colon \{0,1\}^* \to \{0,1\}^\ell$ son funciones *hash* modeladas como oráculos aleatorios.

Figura 2. Protocolo postcuántico de intercambio de clave en grupo seguro en el QROM.

estandarizadas en FIPS 202 [21]. En concreto, Kyber emplea las funciones `SHAKE-128` y `SHAKE-256`; y `SHA3-256` y `SHA3-512`. Con estos parámetros, el tamaño de clave privada y publica y, el texto cifrado se pueden ver la Tabla II.

Tabla II
Longitud (en bytes) del texto cifrado, de la clave pública y privada según el nivel de seguridad. La longitud no varía dependiendo de si la implementación es QROM o ROM.

	512	**768**	**1024**
Longitud texto cifrado	736	1088	1568
Longitud clave pública	800	1184	1568
Longitud clave privada	1632	2400	3168

Nuestra implementación hereda las mismas funciones *hash* y los mismos tamaños en el protocolo autenticado de intercambio de clave en grupo. Lo mismo sucede en el AKE 2-parte con los mensajes enviados, que los parámetros quedan determinados por el tamaño del texto cifrado y la clave pública del KEM (ver Tabla III).

El repositorio contiene dos implementaciones: la de referencia, que sirve para comprobar que Kyber funciona correctamente y la implementación optimizada, que emplea instrucciones AVX2 para paralelizar las funciones de Kec-

Tabla III
LONGITUD (EN BYTES) LOS MENSAJES ENVIADOS POR LAS PARTES DEL AKE 2-PARTE Y LONGITUD DE LA CLAVE OBTENIDA SEGÚN EL NIVEL DE SEGURIDAD.

	512	**768**	**1024**
Longitud de los mensajes enviados por A	1536	2272	3136
Longitud de los mensajes enviados por B	1472	4448	3136
Longitud total de los mensajes enviados	3008	4448	6272
Longitud de la clave		32	

cak [4]. Esta última implementación sólo funciona cuando las instrucciones AVX2 están disponibles en el procesador. A lo largo del artículo nos referiremos a la implementación de referencia como `ref` y a la implementación optimizada con AVX2 como `avx2`.

El protocolo GAKE se ha instanciado empleando las funciones `SHA3-256` y `SHA3-512` para los oráculos aleatorios tanto del compilador como para el AKE 2-parte (ver funciones hash H y G de los *Algs.* 1, 2, 7 y 8). El esquema de compromiso se consigue a través de una aproximación KEM/DEM, donde el KEM es `Kyber` en la versión ROM o $\texttt{Kyber}^{\not\perp}$ en la versión QROM y, el DEM es `AES256-GCM` empleando la librería OpenSSL [22] en su versión `1.1.1f`. El tamaño del compromiso viene dado el tamaño del KEM (Tabla II) más el tamaño del DEM. Este último, está formado por el texto cifrado y el *tag* de `AES256-GCM`. 36 bytes son necesarios para cifrar la concatenación de i y X_i y 16 bytes para el *tag*. El tamaño completo para cada nivel de seguridad se puede ver en la Tabla IV.

Tabla IV
LONGITUD (EN BYTES) DEL COMPROMISO EN FUNCIÓN DEL NIVEL DE SEGURIDAD.

	512	**768**	**1024**
Longitud del compromiso	788	1140	1620

El tamaño del mensaje M_i^1 viene dado por el tamaño del compromiso (Tabla IV) y el tamaño de U_i. Este último lo hemos fijado en 20 bytes, pero podría tener la longitud requerida por la aplicación del protocolo. El tamaño del mensaje M_i^2 viene dado por el tamaño de U_i (20 bytes), más el tamaño de X_i que viene fijado por del tamaño de las claves del AKE (Tabla III) y el tamaño de las aleatoriedades del esquema de compromiso. Las aleatoriedades están formadas por la aleatoriedad del KEM y el DEM. Para el KEM, es de 32 bytes (fijado por los autores de Kyber) y para el DEM es de 12 bytes (fijado para `AES`), que se corresponde con el vector de inicialización del mismo. La Tabla V muestra el tamaño de cada mensaje.

Tabla V
LONGITUD (EN BYTES) DE LOS MENSAJES TRANSMITIDOS DURANTE EL PROTOCOLO.

	512	**768**	**1024**
Longitud del mensaje M_i^1	808	1160	1640
Longitud del mensaje M_i^2		96	

La longitud total de los mensajes enviados por ronda y totales del protocolo en función de n participantes se puede ver en la Tabla VI.

Tabla VI
LONGITUD (EN BYTES) DE LOS MENSAJES ENVIADOS POR RONDA EN FUNCIÓN DEL NIVEL DE SEGURIDAD Y EL NÚMERO DE PARTICIPANTES n EN LA EJECUCIÓN COMPLETA DEL PROTOCOLO GAKE.

	512	**768**	**1024**
Ronda 1-2	$2 \cdot 3008 \cdot n$	$2 \cdot 4448 \cdot n$	$2 \cdot 6272 \cdot n$
Ronda 3	$808 \cdot n$	$1160 \cdot n$	$1640 \cdot n$
Ronda 4		$96 \cdot n$	
Total	$6920 \cdot n$	$10\,152 \cdot n$	$14\,280 \cdot n$

III-A. Resultados experimentales

Toda la implementación y los datos obtenidos se pueden consultar en GitHub[2]. Todos los experimentos se han ejecutado en una máquina virtualizada con las características de la Tabla VII.

Tabla VII
CARACTERÍSTICAS HARDWARE DE LA MÁQUINA QUE HA EJECUTADO LOS EXPERIMENTOS.

Característica	**Valor**
Sistema operativo	Ubuntu 20.04.1 LTS
CPU	i7-6700HQ@2.60 GHz
Memoria RAM	8GB
AVX2 habilitado	Sí

En las siguientes secciones mostramos los resultados de comparar la implementación `ref` con la implementación `avx2`, el rendimiento de las funciones "atómicas" del KEM, del esquema de compromiso y el AKE para las variantes QROM y ROM y por último, el rendimiento del GAKE en conjunto.

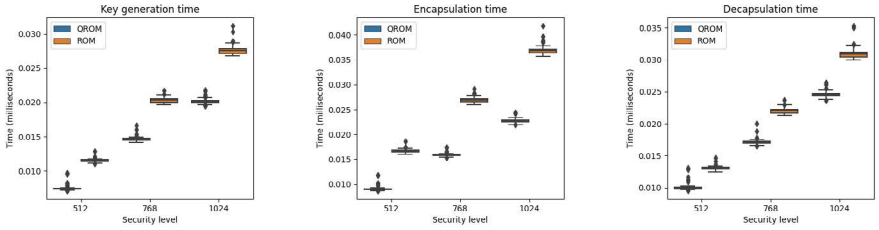
III-B. Implementación `avx2` vs. `ref`

La Tabla VIII muestra una comparativa de cuántas veces es más rápida la implementación `avx2` con respecto a `ref` en distintas operaciones, tanto en la variante QROM como ROM. Los datos son una media de 10 000 tiempos de cada operación. Las mayores diferencias se llevan a cabo en la variante QROM, llegando a ser más de 16 veces más rápida en ciertas operaciones. Las operaciones que más diferencia presentan son la decapsulación del KEM y los algoritmos Der_{init}/sharedA del AKE 2-parte. Por contra, donde se produce la menor diferencia es en las operaciones relativas al compromiso en la variante del ROM.

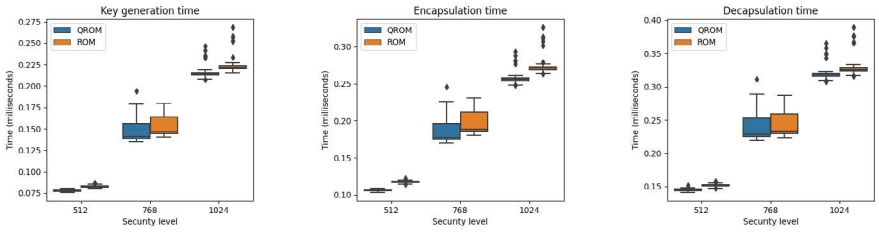
III-C. Operaciones atómicas

En esta sección comparamos el rendimiento de las variantes QROM y ROM en distintas operaciones que se dan en el protocolo GAKE. Los resultados se muestran en las *Figs.* 3, 4 y 5. Los datos son 10 ejecuciones de la media de 10 000 ejecuciones de cada una de las operaciones. Se puede observar que en cada una de las operaciones del KEM, esquema de compromiso y AKE 2-parte la variante QROM es más rápida que la versión ROM, independientemente del nivel de seguridad elegido.

[2]https://github.com/jiep/kyber-gake

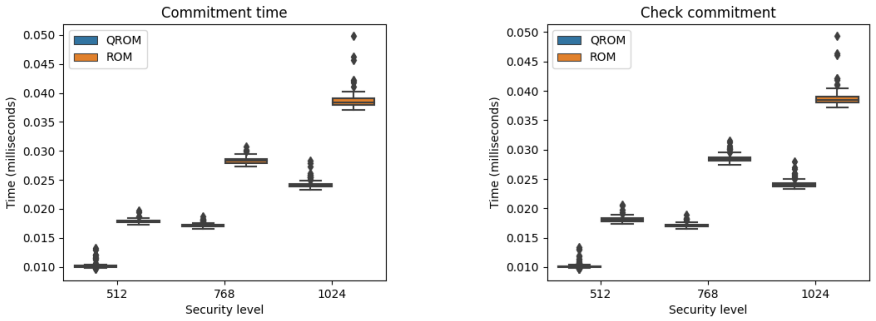


(a) Implementación `avx2`.

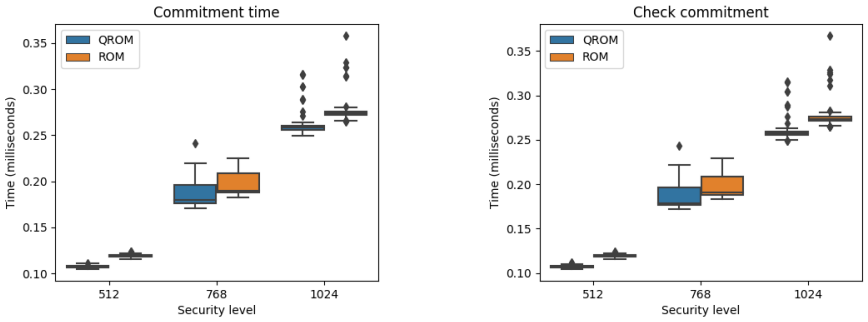


(b) Implementación `ref`.

Figura 3. Tiempo de las operaciones del KEM en función del nivel de seguridad y las primitivas empleadas (QROM o ROM).



(a) Implementación `avx2`.



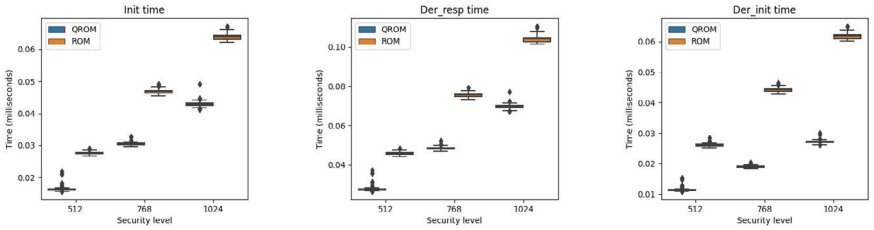
(b) Implementación `ref`.

Figura 4. Tiempo de las operaciones del esquema de compromiso en función del nivel de seguridad y las primitivas empleadas (QROM o ROM).

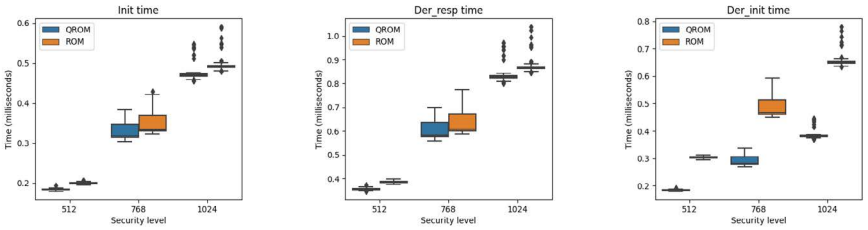
III-D. GAKE

Por último, medimos el rendimiento del tiempo de ejecución del protocolo en función del nivel de seguridad y el número de participantes del protocolo. La *Fig.* 6 muestra los resultados. Los datos mostrados son 10 ejecuciones del protocolo por nivel de seguridad, número de participantes y variante del protocolo (QROM o ROM). El eje de ordenadas muestra el tiempo total de ejecución de cada uno de los participantes y no el tiempo ejecución real que tendría cada una de las partes por separado. Además se asume un tiempo de comunicación de 0. Los resultados vuelven a mostrar que la variante QROM es más rápida que la ROM, independientemente del número de participantes del protocolos y la seguridad.

En cuanto al tiempo total por ronda (*Fig.* 7), se muestra que las rondas que más tiempo tardan en ejecutarse son la ronda 1-2 (ejecución del protocolo AKE 2-parte) y la ronda 4 (comprobación de los compromisos, generación de la clave maestra y derivación de la clave de sesión). Esta última ronda llega a ocupar casi la totalidad del tiempo total cuando el

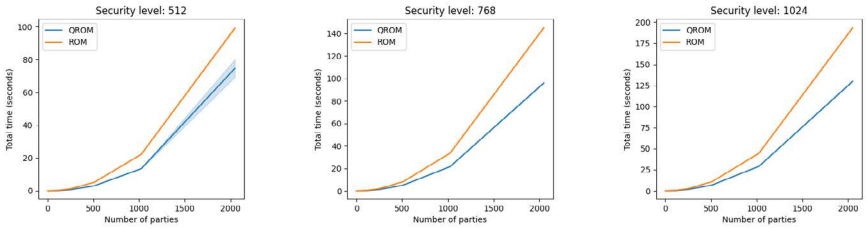


(a) Implementación `avx2`.

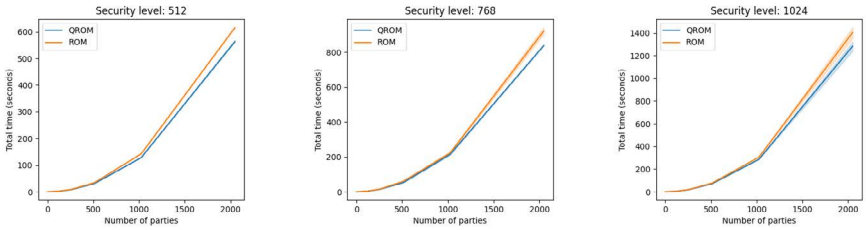


(b) Implementación `ref`.

Figura 5. Tiempo de las operaciones del protocolo de intercambio de clave 2-parte en función del nivel de seguridad y las primitivas empleadas (QROM o ROM).



(a) Implementación `avx2`.



(b) Implementación `ref`.

Figura 6. Tiempo total del protocolo en función del número de participantes, del nivel de seguridad y las primitivas empleadas (QROM o ROM).

número de participantes en el protocolo aumenta. Esto se debe a que se están comprobando los n^2 (n compromisos de n participantes) compromisos del protocolo y no los n que le correspondería a cada una de los n participantes del protocolo. Esto hace que se muestre un comportamiento cuadrático en la *Fig.* 6.

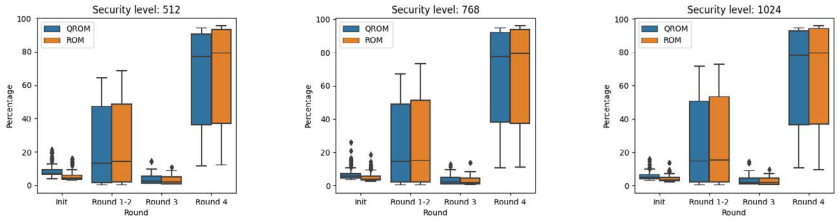
IV. Conclusiones y Trabajo futuro

En este artículo, hemos implementado el protocolo postcuántico de intercambio clave autenticado en grupo seguro en el *Quantum Random Oracle Model*, propuesto en [12]. Su funcionamiento se basa en Kyber, finalista del concurso de estandarización de diseños postcuánticos del NIST y el compilador de Abdalla et. al. Se ha demostrado de forma experimental que tanto las operaciones *atómicas* de las primitivas utilizadas como el protocolo autenticado de intercambio de clave en grupo son más rápidos en el *Quantum Random Oracle Model* que sus alternativas seguras *Random Oracle Model* propuestas en el artículo original de Kyber. Esto pone de manifiesto que un modelo de seguridad más fuerte no implica necesariamente una pérdida de rendimiento.

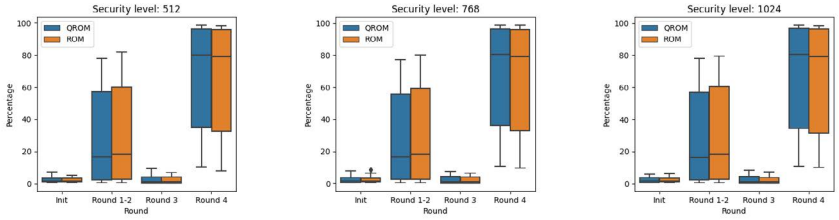
Como trabajo futuro, se proyecta implementar el protocolo con otros KEM finalistas del NIST distintos a Kyber y probar el rendimiento del protocolo con otros esquemas de compromiso. También, sería interesante comparar el rendimiento en otras arquitecturas con recursos limitados como ARM y ejecutar el protocolo en una red de comunicaciones real.

Agradecimientos

Agradecer a María Isabel González Vasco la revisión y los cambios sugeridos para facilitar la comprensión del artículo.



(a) Implementación `avx2`.



(b) Implementación `ref`.

Figura 7. Porcentaje del tiempo total del protocolo que se emplea en cada una de las rondas en función del nivel de seguridad y las primitivas empleadas (QROM o ROM).

Tabla VIII
COMPARATIVA DE LA VELOCIDAD DE DISTINTAS OPERACIONES ENTRE LAS IMPLEMENTACIONES, SEGÚN EL NIVEL DE SEGURIDAD. SE MUESTRA CUÁNTAS VECES ES MÁS RÁPIDA LA IMPLEMENTACIÓN `AVX2` CON RESPECTO A `REF`.

Operación	**512**		**768**		**1024**	
	QROM	ROM	QROM	ROM	QROM	ROM
KEM						
KeyGen	10.42x	7.07x	9.15x	6.87x	10.82x	7.79x
Encaps	11.56x	6.91x	10.64x	6.75x	11.46x	7.16x
Decaps	14.32x	11.23x	12.87x	9.93x	13.04x	10.20x
Compromiso						
Generar compromiso	10.56x	6.53x	10.57x	6.47x	10.96x	6.99x
Comprobar compromiso	10.66x	6.58x	10.57x	6.46x	10.93x	6.96x
2-AKE						
InitA/Init	11.34x	7.09x	10.21x	6.77x	10.91x	7.67x
$\mathrm{Der}_{\mathrm{resp}}$/ sharedB	12.79x	8.23x	11.51x	7.62x	11.80x	8.24x
$\mathrm{Der}_{\mathrm{init}}$/ sharedA	16.48x	11.37x	14.32x	10.01x	13.80x	10.43x

REFERENCIAS

[1] M. Mosca, "Cybersecurity in an Era with Quantum Computers: Will We Be Ready?" *IEEE Secur. Priv.*, vol. 16, no. 5, pp. 38–41, 2018. [Online]. Available: https://doi.org/10.1109/MSP.2018.3761723

[2] NIST. Post-Quantum Cryptography Standardization. [Online]. Available: https://csrc.nist.gov/projects/post-quantum-cryptography/post-quantum-cryptography-standardization

[3] Classic McEliece. Classic McEliece. [Online]. Available: https://classic.mceliece.org

[4] J. W. Bos, L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, J. M. Schanck, P. Schwabe, G. Seiler, and D. Stehlé, "CRYSTALS - kyber: A CCA-Secure Module-Lattice-Based KEM," in *EuroS&P*. IEEE, 2018, pp. 353–367.

[5] NTRU. A submission to the NIST post-quantum standardization effort. [Online]. Available: https://ntru.org

[6] SABER. MLWR-based KEM. [Online]. Available: https://www.esat.kuleuven.be/cosic/pqcrypto/saber

[7] CRYSTALS. Dilithium. [Online]. Available: https://pq-crystals.org/dilithium/index.shtml

[8] FALCON. Fast-Fourier Lattice-based Compact Signatures over NTRU. [Online]. Available: https://falcon-sign.info

[9] PQCRainbow. Rainbow Signature. [Online]. Available: https://www.pqcrainbow.org

[10] D. Apon, D. Dachman-Soled, H. Gong, and J. Katz, "Constant-round group key exchange from the ring-lwe assumption," in *PQCrypto*, ser. Lecture Notes in Computer Science, vol. 11505. Springer, 2019, pp. 189–205.

[11] A. Fujioka, K. Takashima, and K. Yoneyama, "One-round authenticated group key exchange from isogenies," in *ProvSec*, ser. Lecture Notes in Computer Science, vol. 11821. Springer, 2019, pp. 330–338.

[12] J. I. Escribano Pablos, M. I. González Vasco, M. E. Marriaga, and Á. L. Pérez del Pozo, "Compiled Constructions towards Post-Quantum Group Key Exchange: A Design from Kyber," *Mathematics*, vol. 8, no. 10, p. 1853, Oct 2020. [Online]. Available: https://dx.doi.org/10.3390/math8101853

[13] E. Blum, M. Castillo-Martin, and M. Rosenberg, "Survey on the Security of the Quantum ROM," 2019, https://www.cs.umd.edu/class/fall2019/cmsc657/projects/group_2.pdf.

[14] M. Abdalla, J. Bohli, M. I. G. Vasco, and R. Steinwandt, "(Password) Authenticated Key Establishment: From 2-Party to Group," in *TCC*, ser. Lecture Notes in Computer Science, vol. 4392. Springer, 2007, pp. 499–514.

[15] J. Bos, L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, J. M. Schanck, P. Schwabe, G. Seiler, and D. Stehlé, "CRYSTALS – Kyber: a CCA-secure module-lattice-based KEM (version 1)," Cryptology ePrint Archive, Report 2017/634, 2017, https://eprint.iacr.org/eprint-bin/getfile.pl?entry=2017/634&version=20170627:201157&file=634.pdf.

[16] E. Fujisaki and T. Okamoto, "Secure integration of asymmetric and symmetric encryption schemes," in *CRYPTO'99*, ser. Lecture Notes in Computer Science, vol. 1666. Springer, 1999, pp. 537–554.

[17] K. Hövelmanns, E. Kiltz, S. Schäge, and D. Unruh, "Generic authenticated key exchange in the quantum random oracle model," *IACR Cryptology ePrint Archive*, vol. 2018, p. 928, 2018. [Online]. Available: https://eprint.iacr.org/2018/928

[18] R. Cramer and V. Shoup, "Design and Analysis of Practical Public-Key Encryption Schemes Secure against Adaptive Chosen Ciphertext Attack," *SIAM J. Comput.*, vol. 33, no. 1, pp. 167–226, 2003.

[19] R. Avanzi, J. Bos, L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, J. M. Schanck, P. Schwabe, G. Seiler, and D. Stehlé. CRYSTALS-Kyber Algorithm Specifications And Supporting Documentation (version 3.0). [Online]. Available: https://pq-crystals.org/kyber/data/kyber-specification-round3.pdf

[20] ISO, "ISO C Standard 1999," ISO, Tech. Rep., 1999, iSO/IEC 9899:1999 draft. [Online]. Available: http://www.open-std.org/jtc1/sc22/wg14/www/docs/n1124.pdf

[21] M. J. Dworkin, "SHA-3 standard: Permutation-Based Hash and Extendable-Output Functions," National Institute of Standards and Technology, Tech. Rep., Jul. 2015. [Online]. Available: https://doi.org/10.6028/nist.fips.202

[22] The OpenSSL Project, "OpenSSL: The Open Source toolkit for SSL/TLS," April 2003, www.openssl.org.

http://doi.org/10.18239/jornadas_2021.34.57

Análisis y evaluación experimental del circuito generador de números aleatorios *Lampert Circuit*

Alejandro Rodríguez[1], Javier Matanza[2], Gregorio López[2], Carlos Rodríguez-Morcillo[2], Álvaro López[2], Julio Hernández-Castro[3]
[1]ICAI, Universidad Pontificia Comillas, Madrid, España
[2]Instituto de Investigación Tecnológica, ICAI, Universidad Pontificia Comillas, Madrid, España
[3]*University of Kent*, Canterbury, Reino Unido
0000-0001-6021-4957, 0000-0002-0391-133, 0000-0001-9954-3504, 0000-0002-8540-7664, 0000-0001-9879-5603, 0000-0002-6432-5328

Resumen- **La generación de números aleatorios es crítica para la criptografía actual. En el marco del proyecto "*Secure Internet of Things*", liderado por la Universidad de Stanford, se desarrolló un circuito generador de números aleatorios denominado *Lampert Circuit*. Este circuito se basa en electrónica sencilla y destaca por producir números aleatorios con alta entropía (0.98 bits/muestra) y por ser barato, pequeño y auditable, ideal, por tanto, para sistemas IoT. El objetivo de este trabajo de investigación es precisamente auditar el funcionamiento de dicho circuito experimentalmente. Para ello se pretenden replicar parte de los tests descritos en el artículo "*Robust, low-cost, auditable random number generation for embedded system security*", publicado en *ACM SenSys'16*, sobre 100 circuitos fabricados por 4 fabricantes distintos para comparar los resultados obtenidos con los del artículo original. Este artículo presenta la metodología que se pretende seguir en este trabajo y los recursos de los que se dispone para llevarlo a cabo, los problemas encontrados y resultados obtenidos hasta la fecha, y las líneas de trabajo futuras.**

***Index Terms*- Entropía, Generación de Números Aleatorios, Ruido de avalancha**

Tipo de contribución: *Investigación en desarrollo*

I. Motivación y Objetivos

La generación de números aleatorios es crítica para la criptografía actual. En [1] se presenta un circuito generador de números aleatorios denominado *Lampert circuit*, en honor de su inventor, Ben Lampert. Este circuito, desarrollado en el marco del proyecto "*Secure Internet of Things*", liderado por la Universidad de Stanford, se basa en electrónica sencilla (dos fuentes de ruido basadas en diodos Zener polarizados en inversa que sirven de entrada a un amplificador operacional en lazo abierto que funciona como comparador) y, por tanto, los autores del artículo destacan que es transparente y auditable. Además, los autores también defienden que: al basarse en ruido de avalancha el circuito es inherentemente probabilístico y resistente al control de un adversario; al comparar dos fuentes de ruido de avalancha el circuito es robusto frente a perturbaciones de radiofrecuencia o rizado en la alimentación; y es barato, pequeño, puede deshabilitarse para ahorrar energía y produce una entropía de hasta 0.98 bits/muestra, lo que lo convierte en un generador de números aleatorios ideal para sistemas IoT (*Internet of Things*) [1].

El objetivo de este trabajo de investigación es precisamente auditar el funcionamiento de dicho circuito experimentalmente. Concretamente, se pretende responder a las siguientes preguntas:

- ¿Puede replicarse el diseño y las pruebas descritas en [1]?
- ¿Coinciden los resultados obtenidos con los reportados en [1]?
- ¿El rendimiento del generador de números aleatorios depende del circuito fabricado?
- ¿Es posible atacar al generador de números aleatorios propuesto?
- ¿Cómo podría mejorarse dicho generador?

Los resultados de este estudio pueden resultar valiosos para determinar la utilidad real de este generador de números aleatorios para sistemas IoT.

El resto del artículo se organiza de la siguiente manera. La sección II presenta la metodología del trabajo de investigación y los recursos disponibles para llevarlo a cabo. La sección III describe el estado en el que se encuentra actualmente el proyecto, destacando los problemas encontrados y los resultados obtenidos hasta la fecha. Por último, la sección IV resume las principales conclusiones y futuras líneas de trabajo.

II. Metodología y Recursos

Para poder evaluar si el rendimiento del generador de números aleatorios depende del circuito utilizado se han fabricado 100 unidades del *Lampert circuit*. Para poder evaluar además si el fabricante del circuito influye en su rendimiento, se han encargado 25 circuitos a 4 fabricantes distintos: *European circuits* (Reino Unido), *Wurth* (España), *Micron20* (Bulgaria) y *ShenZhenU2* (China). La Fig. 1 muestra los circuitos fabricados.

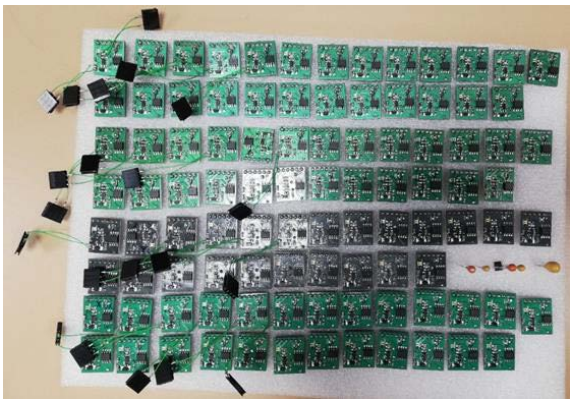


Fig. 1. Circuitos fabricados para la realización del estudio

Las pruebas descritas en [1] se repetirán para cada uno de estos circuitos. Concretamente, se comenzará con los ensayos para obtener la entropía y la correlación serie del generador de números aleatorios en función de la frecuencia de

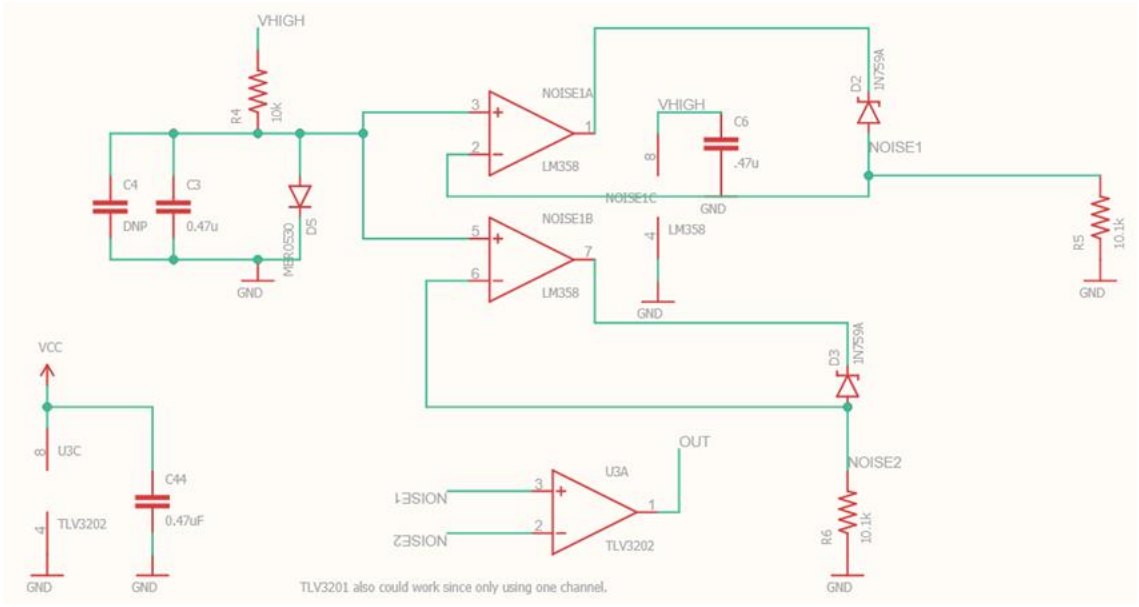


Fig. 3. Esquemáticos y componentes del circuito utilizado en las pruebas [2]

muestreo. Una vez replicados estos experimentos y en base a los resultados obtenidos, se decidirá si se repiten las pruebas descritas en [1] para obtener la evolución de la entropía y la correlación serie con el tiempo y con la temperatura.

La Fig. 2 muestra el laboratorio y el instrumental del que se dispone para llevar a cabo las pruebas, que se resume brevemente a continuación:

- Osciloscopio Tektronix TDS5104 (rodeado en rojo en la Fig. 2), con un ancho de banda máximo de 1GHz, 4 canales y una tasa de muestreo de hasta 5Gs/s. Este osciloscopio está conectado a la red LAN del laboratorio para poder procesar las medidas realizadas con MATLAB.
- Generador de señales RIGOL DG1000Z (rodeado en verde en la Fig. 2).
- Fuente de alimentación Siglent SPD3303C (rodeado en azul en la Fig. 2).

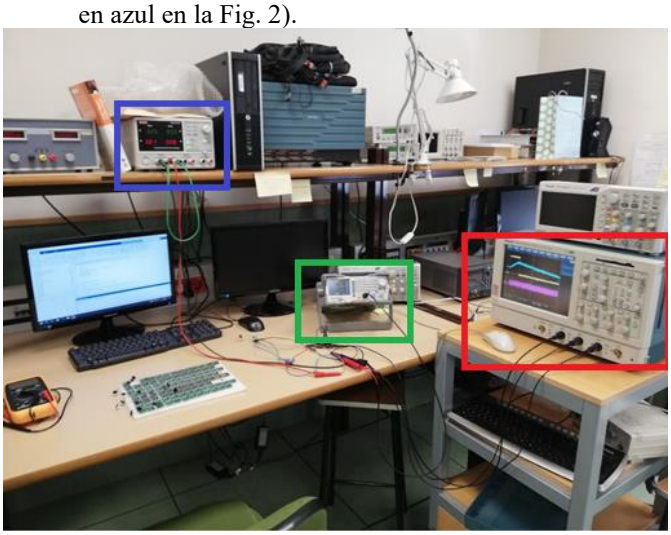


Fig. 2. Laboratorio e instrumental disponible para llevar a cabo las pruebas

III. Descripción del Trabajo Realizado y del Estado Actual del Estudio

El primer problema con el que nos encontramos de cara a replicar los experimentos descritos en el artículo es que había discrepancias entre los esquemáticos y la descripción de componentes incluida en el artículo y la versión disponible en el repositorio de *GitHub* [2], que era la que se decía que se había utilizado en las pruebas del artículo. Después de intercambiar varios correos con los autores, actualizaron la información relativa al circuito con el que se llevaron a cabo las pruebas descritas en el artículo en el repositorio de *GitHub*. La Fig. 3 muestra el esquemático y los valores de los componentes utilizados.

La Fig. 3 sirve para entender el funcionamiento del generador de números aleatorios. El circuito formado por R4, C3 y D5 sirve para fijar la tensión de referencia en la pata positiva de los operacionales LM358. Las fuentes de ruido forman parte de la red de realimentación de dichos operacionales. El diodo usado para para la generación de ruido es el *Zener* 1N759A, que tiene una tensión *Zener* nominal de 12V con una tolerancia del 5%. Las dos fuentes de ruido van al comparador TLV3202 y la salida de dicho comparador (*OUT*) es la salida del generador de números aleatorios. Por tanto, si *NOISE1* > *NOISE2*, el comparador saturará a *Vcc* y si *NOISE1* < *NOISE2*, saturará a GND.

Para que el circuito funcione como se acaba de describir, es necesario polarizar el diodo *Zener* en inversa, por lo *VHIGH* debe ser superior a 12 V. Sin embargo, la alimentación de la mayoría de circuitos empotrados varía entre 1.8 y 3.3 V. Para conseguir tensiones superiores a 12 V partiendo de 3.3 V se utiliza el circuito denominado *boost converter*, que se muestra en la Fig. 4. Como puede verse en la Fig. 4, también hay discrepancias entre el circuito *boost converter* que aparece en el artículo y el utilizado en las pruebas.

Este circuito se controla con la señal de *enable*. Cuando se activa la señal de *enable*, el condensador conectado a *VHIGH*/*Vout* se carga hasta llegar a los 18 V. Cuando *VHIGH* supera los 12 V, el generador de números aleatorios comienza a funcionar. Cuando se desactiva la señal de *enable,* el condensador se descarga. Cuando *VHIGH* cae por debajo de 12 V, el generador de números aleatorios deja de funcionar.

La Fig. 5 (a) muestra el funcionamiento indicado en [1], la Fig. 5(b) muestra el funcionamiento medido en nuestro laboratorio y la Fig. 5(c) muestra el funcionamiento indicado

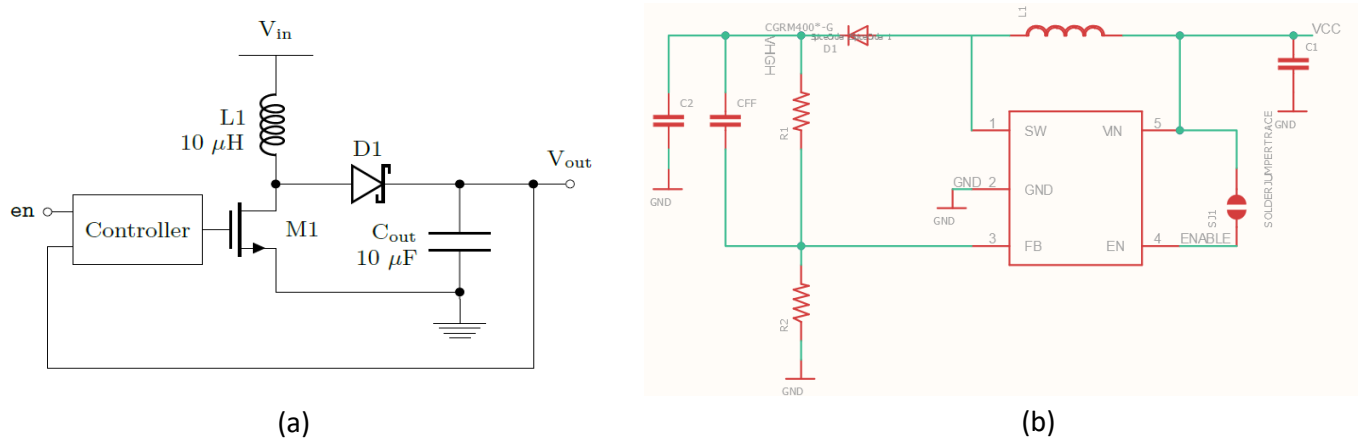


Fig. 4. (a) *Boost converter* que aparece en [1]; (b) *Boost converter* disponible en [2], utilizado para las pruebas de [1]

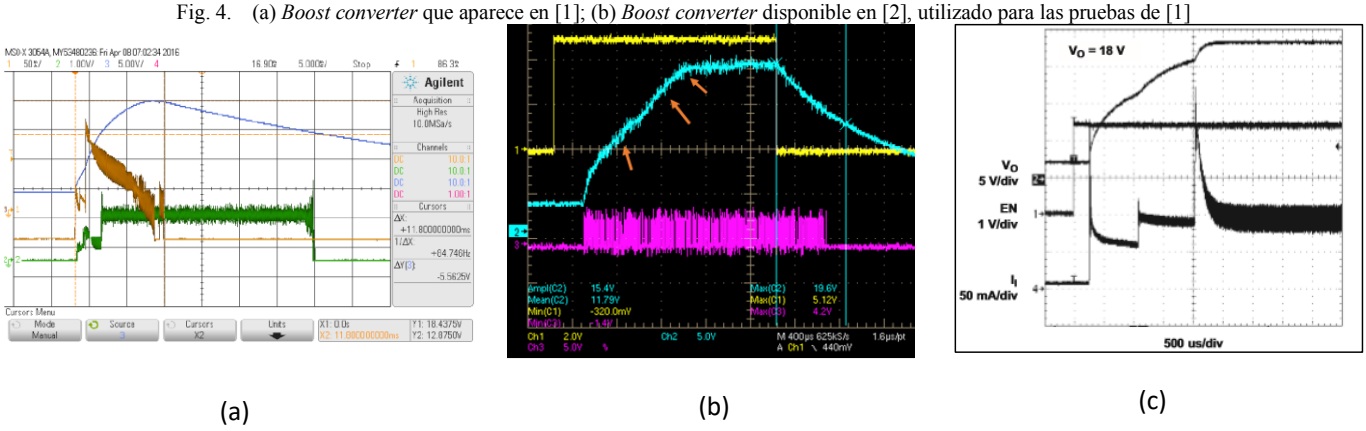


Fig. 5. (a) Funcionamiento del *boost converter* indicado en [1]; (b) Funcionamiento del *boost converter* medido en nuestro laboratorio; (c) Funcionamiento del *boost converter* según el *datasheet* del circuito TPS61041

en el *datasheet* del circuito TPS61041, utilizado como controlador en el *boost converter* usado en las pruebas descritas en el artículo (y en nuestros experimentos). Si se compara la carga mostrada en la Fig.5(a) con la mostrada en las Fig.5(b) y (c), se observa como en estas últimas la carga parece realizarse en varias etapas. Esto se debe a que el controlador controla, valga la redundancia, la corriente de carga del condensador para evitar que se alcancen corrientes elevadas que podrían causar daños en el circuito. Además, si comparamos la descarga de la Fig. 5(a) con la de la Fig. 5(b), puede observarse que en el caso de la Fig. 5(b) la descarga es mucho más rápida que en la Fig. 5(a). En la Fig.5(b) se observa también que en las medidas tomadas en el laboratorio el generador de números aleatorios deja de funcionar antes de que *VHIGH* llegue a 12 V.

La Tabla I muestra el tiempo de descarga medido para 5 placas de cada uno de los fabricantes. Analizando el tiempo medio de descarga medido se observa que no hay diferencias significativas entre fabricantes, salvo para el caso de los circuitos del fabricante *ShenZhen2U*. Para *European Circuits*, *Wurth* y *Micron20*, el tiempo de descarga medido varía entre los 0,6 y 0,8 ms, lo que está muy por debajo de los 20 ms que tarda el *boost converter* en pasar de 18V a 12V en el artículo. Para aumentar el tiempo de descarga, se aumentó la capacidad del condensador C2 hasta 10 uF utilizando el montaje mostrado en la Fig. 6, obteniéndose tiempos de descarga de alrededor de 30 ms.

Tabla I. Tiempo de descarga del *boost converter* para 5 circuitos de cada uno de los fabricantes

Tiempo de descarga desde la desactivación del *enable* hasta alcanzar los 12V (ms)						
Fabricante\ID circuito	**001**	**002**	**003**	**004**	**005**	**Tiempo medio**
ShenZhen2U	1,68	1,71	1,70	1,76	1,56	1,68
European Circuits	0,624	0,752	0,760	0,768	0,656	0,712
Wurth	0,680	0,602	0,723	0,776	0,816	0,719
Micron20	0,812	0,640	0,616	0,784	0,624[1]	0,695

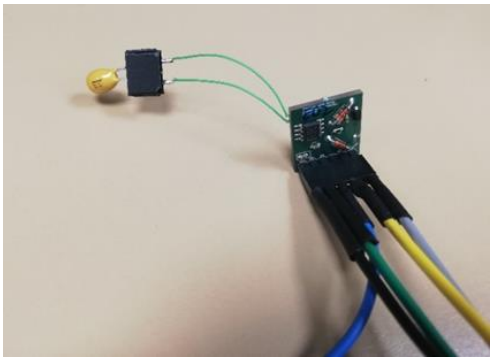


Fig. 6. Montaje realizado para aumentar la capacidad de C2

Una vez solventadas las discrepancias encontradas en el funcionamiento del *boost converter*, actualmente se van a

[1] Dado que el circuito 005 de Micron20 y el 001 de Shenzhen2U no funcionan se ha muestreado el 006 y el 002 en su lugar respectivamente.

comenzar a realizar medidas de la entropía del generador de números aleatorios utilizando la frecuencia de muestreo recomendada en el artículo (128 kHz) y midiendo más de 768 bits, para tener al menos 690 bits de entropía [1]. Cuando se se corrobore que los resultados obtenidos para esta frecuencia de muestreo concuerdan con los obtenidos en el artículo, se procederá a variar este parámetro.

IV. Conclusiones y Trabajos Futuros

Actualmente aún no hemos podido dar respuesta a las preguntas de investigación que nos planteamos en este estudio. La principal conclusión del trabajo llevado a cabo hasta la fecha es que, a pesar de que en teoría cualquier artículo científico debería ser fácilmente replicable, en este caso no está siendo fácil reproducir en nuestro laboratorio las mismas condiciones descritas en el artículo. Esta conclusión podría generalizarse a que es más desafiante replicar un trabajo de investigación *hardware* que uno *software*.

En cuanto a los próximos pasos, como se acaba de comentar, primero procederemos a medir la entropía para la frecuencia de muestreo recomendada (128 kHz) para comprobar que encaja con los resultados del artículo. Una vez hecho esto, variaremos este parámetro para comprobar si en nuestro laboratorio medimos la misma relación entre la entropía/correlación serie y la frecuencia de muestreo que se reporta en el artículo. Si los resultados son positivos, evaluaremos si ocurre lo mismo con la relación de la entropía y la correlación serie con el tiempo y la temperatura.

Si todo fuera bien y obtuviéramos resultados coherentes con el artículo al medir la entropía para la frecuencia de muestreo recomendada, uno de los principales retos para llevar a cabo las medidas descritas anteriormente será la automatización de las mismas.

Una vez hayamos finalizado con las medidas descritas en el párrafo anterior, el objetivo es evaluar cómo se podría atacar el circuito, así como proponer mejoras para el mismo.

Teniendo en cuenta el auge del IoT y que cada vez va a ser más necesario proporcionar seguridad en este tipo de entornos ajustándose a las restricciones que presentan, un circuito generador de números aleatorios como éste creemos que puede tener mucho potencial, siempre y cuando su rendimiento sea fiable y se ajuste a lo que se destaca de él en [1].

Referencias

[1] B. Lampert, R.S. Wahby, S. Leonard, P. Levis, “Robust, low-cost, auditable random number generation for embedded system security,” *SenSys '16: Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems*. November 2016, Pages 16–27, doi: https://doi.org/10.1145/2994551.2994568

[2] GitHub Lampert Circuit: https://github.com/lampertb/LampertCircuitRNG

Sesión de Investigación B5:
Privacidad

Sesión de Investigación B5: Privacidad
http://doi.org/10.18239/jornadas_2021.34.58

A privacy preserving approach for avoiding database recovery attacks in software developments

Xabier Larrucea
TECNALIA, Basque Research and Technology Alliance (BRTA)
Astondo bidea, 48160 Derio, Bizkaia, Spain
xabier.larrucea@tecnalia.com
ORCID: 0000-0002-6402-922X

Izaskun Santamaría
TECNALIA, Basque Research and Technology Alliance (BRTA)
Astondo bidea, 48160 Derio, Bizkaia, Spain
Izaskun.santamaria@tecnalia.com
ORCID: 0000-0003-2135-4644

***Abstract*— Privacy is being considered an asset in several business domains where cybersecurity plays a key role. This is especially relevant in contexts where Personally Identifiable Information is exchanged, managed or processed, e.g. the health domain. GDPR and other regulations are stressing the privacy relevance, and the need to protect it with advanced techniques such as differential privacy. Health systems are allowing the exchange of sensitive information and, therefore, software companies building solutions including sensitive information exchange must adopt approaches to enhance privacy preservation. In order to assist these organizations in these matters, authors propose the integration of a set of steps for enhancing privacy and shifting left privacy concerns within the software implementation activities of the ISO/IEC 29110 basic profile In addition, this method is illustrated with a use case.**

Keywords*— *Security, Privacy, Information sharing

I. Introduction

It is widely known that cybercriminals can sneak through loopholes of any software system. These weaknesses are being exploited and some of them are related to software defects [1]. Therefore, in order to reduce the existing vulnerabilities, some approaches are focused on tackling the problems from early stages of the software developments [2]. In addition, privacy is being considered a critical issue during the development of a system [1]. This aspect is especially relevant since the release of the General Data Protection Regulation (GDPR) [3]. Software systems managing Personally Identifiable Information (PII) [4] must consider this aspect as a critical factor, especially when systems must be interconnected, or they need to exchange this information. This is the case for healthcare management systems (HMS) where medical records and personal data are being processed. Systems must guarantee some degree of protection to stakeholders and technologies should be used in this sense. Therefore, information exchange must be considered during software development, especially when dealing with PII.

As stated by ENISA, an "*Information sharing efforts must respect privacy and civil liberties and should be designed with the aim of protecting these to the highest degree*" [5]. Traditionally Critical Information Infrastructure Protection (CIIP) operators and users are interested in enhancing privacy, but there are some hurdles related to information sharing such as lack of trust.

Software development companies must consider privacy as a critical factor in every project. This is applicable to those companies related to the medical or healthcare sector, but also for companies developing software where PII is exchanged, managed or processed. This is especially relevant for small companies developing software. According to the ISO/IEC 29110 [6], Very Small Entities (VSE) are those teams with less than 25 teammates. There are several studies on ISO/IEC 29110 where it is highlighted the increasing interest in this standard [7]. In fact, some research works are focused on how to increase security in these settings [8], and on the integration of security practices in the standard [9].

As stated before, HMS are using PII, and there is an increasing number of solutions providing new services to stakeholders. The medical sector is moving to an integrated set of digitalized healthcare products and digitalized healthcare services. There is a huge number of software development companies developing software for this sector. The integration of different technologies with healthcare systems is representing a major challenge especially when dealing with patients health records [10], and its compliance to GDPR rules. Medical Data management including transfers of personal data to third countries or international organizations are major subjects within this European law, and mechanisms must be set up for assuring security and privacy, especially when managing patient's health records [11].

However, it is a double-edged sword because these new services must be integrated with the existing systems, and potentially they are compromising the whole system, especially with respect to privacy issues.

Based on this context, we have identified the following research question (RQ):

RQ1: *How can be integrated a privacy preserving method within the software implementation activities of VSEs?* This research question is focused on providing a "shift-left" strategy for preserving privacy along the software development process of a VSE. In fact, ISO/IEC 29110 provides the means for defining a profile for a specific

purpose. However, instead of defining a new profile, our approach is to define some steps to be performed on every activity of the basic profile.

RQ2: *Is there any possibility to identify Personally Identifiable Information within an existing system exchanging medical records?* This research question is focused on not assuming that existing software solutions are considering privacy in a right way. Therefore, we need to include a data analysis method for analyzing the dataset managed by the resulting solution.

RQ3: *What recommendations can be provided from a real use case?* This research question is focused on providing an experience about a VSE developing a software integrating two different OpenNCP. This VSE developed a solution and we integrated a DP approach into the ISO/IEC 29110 software implementation's activities.

This paper is structured as follows. Firstly, authors provide a background on the main topics of this paper. Secondly, we define a privacy preserving software development method aligning CRISP-DM, the ISO/IEC 29110 software implementation's activities from its basic profile, and a differential privacy technique for keeping private some sensitive data. Thirdly, authors describe a use case related to the medical sector where sensitive information is being processed, and where our method is tested. Fourthly, we provide some lessons learned from this experience. Finally, we present conclusions on the work performed.

II. Background

A. ISO/IEC 29110

Since the first release of the ISO/IEC 29110 series back in 2011, several research papers have been published revealing new contributions, studies, and experiences carried out by VSEs. In fact, this standard helps this kind of organizations to adopt or to embark on software process improvement initiatives from different perspectives. Literature reflects interest on this standard during these recent years [7]. The number and types of contributions are quite diverse ranging from improvement experiences [12], [13] to project management practices [14], and it is not always evident where the focus is from these main contributions and which improvements are proposing from these articles and papers. Some studies are focused on survivability analysis within the context of VSEs [12].

However, there are too few proposals focused on how to improve privacy preserving techniques within the umbrella of VSEs. To date, we have been proposing how to increase security within their software developments [8], and on to integrate them in small settings [9], but we have left aside how to include and enhance privacy within their developments.

B. PERSONALLY IDENTIFIABLE INFORMATION

There are several definitions and references for what constitutes PII. Organizations such as NIST provide their own PII definition [4], and they use to refer the OMB Memorandum M-07-1616 [15] which defines PII as the information that can be used to distinguish or trace an individual's identity, either alone or when combined with other personal or identifying information that is linked or linkable to a specific individual [16]. Other definitions are generalizing the concept as any information (a) that identifies or can be used to identify, contact, or locate the person to whom such information pertains, (b) from which identification or contact information of an individual person can be derived, or (c) that allows linking particular personal characteristics or preferences to an identifiable person [17]. This PII concept dates back in a U.S. Privacy Act 1974 regulating the collection of personal information by government agencies [18].

The protection of any piece of PII is considered the cornerstone for any system processing personal data. Several techniques have been defined for protecting this kind of data. In fact, these systems are relying on privacy-enhancing technologies (PET) [17]. PI data disclosure is a risk to be managed appropriately. Other authors are focusing their research at the network traffic level, and they are proposing a Software Defined Networking (SDN) / Network Function Virtualization (NFV)-enabled architecture for improving the efficiency of leak detection systems [19]. Similar approaches are used for identifying PII in internet traffic [20].

C. PRIVACY PRESERVING TECHNIQUES

There are several regulations focused on protecting privacy and firms are worried about which technique must be used and how. Some regulations such as GDPR use the term 'pseudonymisation' where personal data can no longer identify a specific subject. This kind of regulations promotes the use of advanced techniques for ensuring some degree of privacy. In fact, there is a bunch of different techniques based on statistical approaches with the same purpose. It is widely known that the emergence of new powerful algorithms jeopardizes anonymization techniques [18]. One of these techniques is the K-anonymity [21], and it has been proposed as a mechanism for protecting privacy [22], and there are several variants for enhancing different properties such as multi-record privacy preservation [23] for allowing multi-record datasets anonymity, bidirectional privacy preservation failure, and personalized privacy preservation failure. K-anonymity ensures that there are at least k-1 equivalents for a specific quasi identifier in a dataset. K-anonymity is usually combined with l-diversity [24] and/or with t-closeness [25] which increment privacy by adding more properties to the dataset, but there are several advantages and disadvantages reported[13]. Any anonymization technique should be conducted in a careful manner, such that the published data not only prevents an adversarial attack by inferring sensitive information or by using background information, but it also must remain useful for data analysis.[26]. Anonymization techniques provide a trade-off between the strength of the privacy guarantee and the quality of the anonymized dataset [27].

Differential Privacy (DP) was introduced by Cynthia Dwork [28] and it aims to protect the privacy of statistical databases by adding noise to the original data. The concept of noise is usually misunderstood because some practitioners interpret the noise concept as creating a rubbish database. However, DP preserves the inherent properties of the original dataset in order to preserve sensitive data. One of the fundamental properties of DP is that it guarantees the query output results of a dataset.

This DP concept has been successfully applied to different domains. Different versions of DP have been proposed such as [29] where authors are proposing a new DP scheme base on Recurrent Neural Network (RNN) for ensuring the privacy of the real-time trajectory data. DP has also been combined with microaggregation for preserving utility in differentially private data releases [30]. Other techniques such as data mining has been proposed with differential privacy [31], and DP has been used for publishing set-valued data via differential privacy[32]

The Census Bureau of the United Stated of America has announced that it is adopting a noise-injection mechanism based on differential privacy to provide privacy protection for the underlying microdata collected as part of the 2020 census [33]. Many statistical agencies leave themselves open to the Database Reconstruction Attacks [33], and therefore the use of DP or DP based algorithms are being considered as the solution for publishing data.

III. A PRIVACY PRESERVING SOFTWARE DEVELOPMENT APPROACH

This method is focused on how to improve privacy during the software developments of VSEs. Table I provides an overview of the recommended activities for developing software. Basically, it defines 6 activities which can be applied to any software development.

Table I
ISO/IEC 29110 SOFTWARE IMPLEMENTATION ACTIVITIES

Profile Element2 ID	Profile Element2 Name
SI.1	Software implementation Initiation
SI.2	Software Requirements Analysis
SI.3	Software Architectural and Detailed Design
SI.4	Software Construction
SI.5	Software Integration and Tests
SI.6	Product Delivery

Privacy preserving techniques require a careful analysis of the data. Therefore, our proposal is based on adding a set of steps aligned to the most common process framework for data mining: the Cross Industry Standard Process for Data Mining (CRISP-DM model) [34] which has been widely used in several ways and domains [35]. In fact, this CRISP-DM is an open standard process model, and it is considered as a standard *de facto* [36]. Our approach uses it as the reference model for carrying out activities for data analysis.

Fig. 1 provides an overview of the CRISP-DM phases and ISO/IEC 29110 software implementation activities. As it can be shown in this Fig. 1, there is a light relationship between CRISP-DM phases and the ISO/IEC 29110 software implementation activities. While these implementation activities are focused on developing software including the inception, requirements analysis, design, code, test and delivery, the CRISP-DM phases are focused on analysing the data that the resulting software systems is relying on. This relationship among these two models is defined by aligning them in a single path.

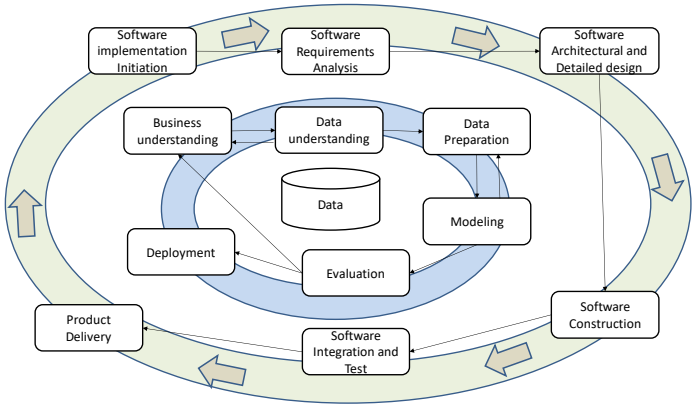


Fig. 1: Cross Industry Standard Process for Data Mining (CRISP-DM model) [34] and the ISO/IEC 29110 software implementation's activities.

For each activity of the ISO/IEC 29110 basic profile we have added a set of recommendations to be performed, and we relate them to the CRISP-DM model's phases:

- **Software Implementation Initiation** which is related to CRISP-DM Business Understanding: This initial phase must include an analysis of the data that it is going to be exchanged among systems. This data is analyzed taking into account privacy preserving techniques such as differential privacy [28].
- **Software Requirements Analysis** which is related to CRISP-DM Data Understanding: in addition to the traditional analysis of the requirements, the data is analyzed taking into account the privacy preserving techniques requirements. In our case, we include DP as our privacy preserving technique. As stated before, if we are going to use k-anonymity we need to ensure that there are at least k-1 equivalents

for a specific quasi identifier in a dataset. This means that individuals cannot be univocally and directly identified, even if we add a generalisation technique. This requirement stemming from the dataset is not always evident, and we need to gain familiarity with the data as much as possible in order to identify data quality problems or to discover new insights. In order to avoid k-anonymity weaknesses, as it has been reported in [33], DP algorithm is adopted in our approach. This concept generally applies to any private data set from which we want to release useful information while keeping private the details. In this context, we consider pairs of databases D_1, D_2 differing in at most one row. In fact, one database is a subset of the other one, and the larger database contains exactly one additional row. Therefore, an algorithm $\mathcal{K}$ satisfies ε - Differential Privacy (ε -DP) if and only if for any two neighboring datasets D_1, D_2, the distributions of $\mathcal{K}\ (D_1)$ and $\mathcal{K}\ (D_2)$ differ at most by a multiplicative factor of ε, and all $S \subseteq Range(\mathcal{K})$

$$Pr\ [K(D\_1\ ) \in S] \leq exp\ (\varepsilon) \times Pr\ [K(D\_2\ ) \in S] \quad ..(1)$$

DP is based on a set of mechanisms and the mechanism used is the Laplace distribution [37] for adding controlled noise, especially for numeric queries where: $f: \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$ which map databases $x \in \mathbb{N}^{|\mathcal{X}|}$ to k real numbers. When defining queries we must to take into account the accuracy of the queries called global sensitivity and which follows:

$$\Delta f \coloneqq \max_{\|x-y\|_1=1} \|f(x) - f(y)\|_1.. \qquad (2)$$

We denote Laplace distribution as Lap(b) with probability density function with scale b and centred at 0. Given any function $f: \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$ the Laplace mechanism is defined as

$$\mathcal{K}_L(x, f(\cdot), \epsilon) \coloneqq f(x) + (Y_1, \ldots, Y_k) \qquad ..(3)$$

Therefore, the Laplace distribution Lap $(\Delta f\ /\ \epsilon)$ is applied to each component of the output. Laplace noise is added to the dataset according to the sensitivity of the function Δf . This is an ϵ differentially private mechanism with $\epsilon = \Delta f\ /\ b$ [38].
As result, we obtain a dataset with noise which pseudonymises the real data, but it keeps its behaviour.

- **Software Architectural and Detailed design** which is related to CRISP-DM Data Preparation: It is tightly related to the previous activity in the sense that we need to ensure the properties of the current dataset are fulfilling with the DP requirements.
- **Software Construction** which is related to CRISP-DM Modeling: we selected the algorithm and we applied it to the dataset. In the CRISP-DM phase, techniques are selected and applied, and their parameters are calibrated to optimal values.
- **Software integration and Test** which is related to CRISP-DM Evaluation: once the system is built, we need to evaluate whether the system compromises the privacy of the data.
- **Product Delivery** which is related to CRISP-DM deployment: this activity includes several checks in order to ensure we are delivering the appropriate product. We need to include a report describing the data nature, and the expected results.

IV. HEALTHCARE USE CASE

As a use case, we describe a VSE developing a software system based on OpenNCP [39] which is a platform connecting National Health Services (NHS) across European countries. Each country is connected throughout a set of services to the rest of national contact points in order to create a network for sharing patient's health records. These records contain PII, and we need to identify them and to apply a privacy preserving technique. GDPR is applicable to this use case.

The connection of different NHS across Europe is a complex scenario requiring a defined and validated framework. Several research efforts have been performed in this sense [40], in order to solve or to reduce the challenges for the healthcare sector [41]. Taking into account this scenario and our recent published works in this sense [42] we apply our method for helping the VSE to include privacy preserving techniques within their developments.

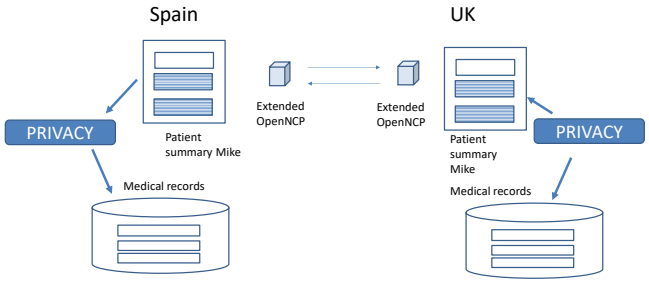


Fig. 2: Use case of the integrated system throughout OpenNCP.

Fig. 2 illustrates two different countries (UK, Spain) connecting their NHS throughout an extended OpenNCP which allows the exchange of patients' medical records[43].

All databases are analysed in order to find out sensitive information. The resulting software system must be able to keep private the communications and any trace related to individuals. In fact, this is the main requirement: **to private any information related to an individual (PII)**.
These NHS are processing health records, and other data, and different events are stored in logs containing this sensitive information. The information exchanged between the two systems follows a JSON format. An excerpt of this JSON file is:

```
"graph": {
    "nodes": [
            {"id": "xmlProcessor",
            "root": true          },
    {"id": "anyProcessor"         },
    {"id": "encryptionProcessor"},
    {"id": "decryptionProcessor"}
    ],
    "edges" : [{"source": "xmlProcessor",
"target": "anyProcessor","selector":
"XPathIdSelector"}
```

This format reveals that there is potential information to be pseudonymised such as “source” and “target”. In addition, HL7 format [44] is the reference for representing patient data (Fig 3). A Patient format contains different attributes that can be used for identifying a subject. For the sake of simplicity, this paper is just focused on “identifier” attribute, but we can extend this approach to the rest of the attributes.

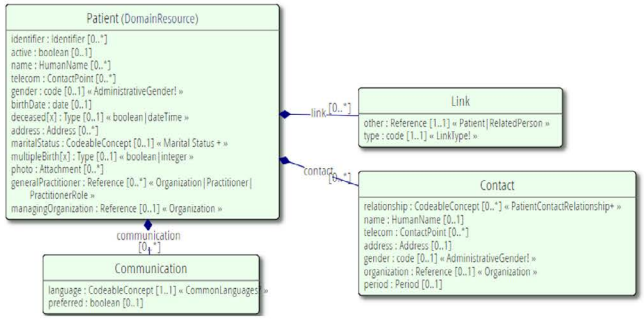


Fig. 3: UML class diagram: Patient data

Our VSE carried out the following ISO/IEC 29110 software implementation activities, and we applied our privacy preserving software development method as follows:

- **Software Implementation Initiation** which is related to CRISP-DM Business Understanding: As initiation step we identify the requirements for connecting NHS through OpenNCP. These requirements are stemming from previous research works such as [42] where GDPR compliance was an explicit requirement which was broken down into a set of activities such as to discover the personal data in the organization datastores, to categorize the data, and finally to apply appropriate methods to protect the data. This is critical use case because the resulting software system manages PII and health data, and GDPR must be ensured. In addition, this use case implies a cross-border processing which is also included within the GDPR definition (23).
- **Software Requirements Analysis** which is related to CRISP-DM Data Understanding: the medical records exchanged within OpenNCP are masked by using data hiding tools. However, patient attributes are not masked, and they can be potentially used for identifying individuals.
- **Software Architectural and Detailed design** which is related to CRISP-DM Data Preparation: a DP algorithm is designed to be applied to patient format.
- **Software Construction** which is related to CRISP-DM Modeling: we select the algorithm and we applied it to the dataset. A DP algorithm is applied to all patient identifiers.
- **Software integration and Test** which is related to CRISP-DM Evaluation: we evaluate the data extracted from applying DP. Fig. 4 describes two histograms. The first histogram (the upper histogram) represents an excerpt of patient identifiers used in the case study. The second histogram (the lower histogram) represents the private data after applying the DP algorithm.

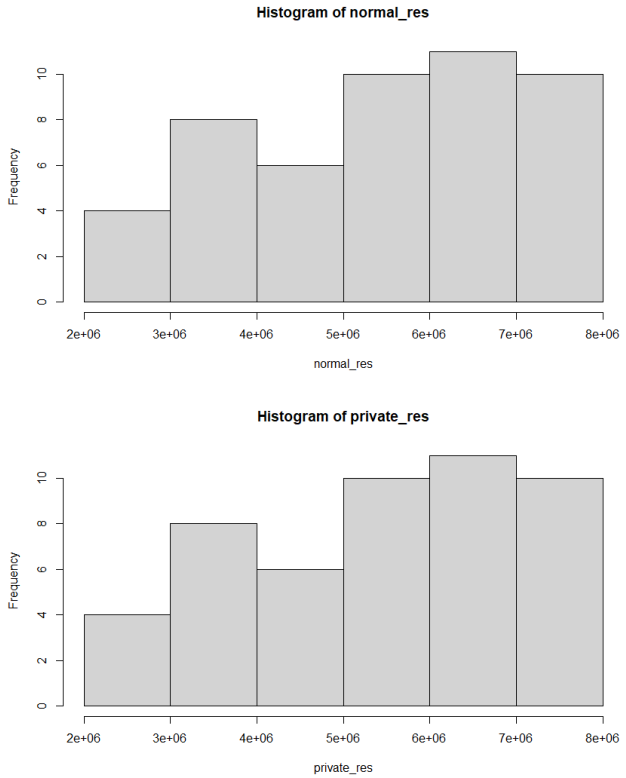


Fig. 4: Histograms with real values (“normal_res”) and with private (DP) results (“private_res”)

Product Delivery which is related to CRISP-DM deployment: the VSE checked the values obtained. In fact, as we can see in Fig. 4 the differences between the two histograms are enough to do not reveal the real values. The second histogram contains the “noise” added by the DP algorithm. In fact, our DP algorithm generates a different answer from the same query. This is a relevant property of any DP approach. Thus, any potential database recovery attack would be frustrated by this property. Each time we query the database we will get a different answer. However, all the answers are similar among them including the original data.

Therefore, the attacker cannot infer anything from these queries. In order to demonstrate this property, we have executed 100 times this algorithm over the same dataset, and as result we get the data with

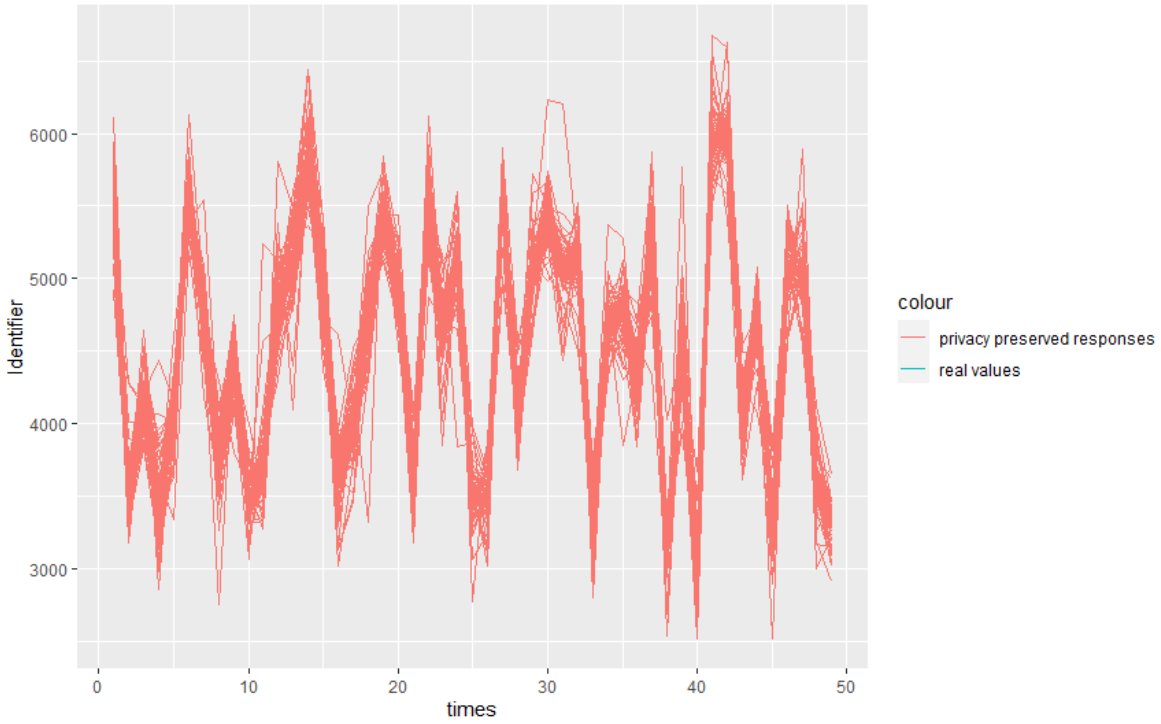


Fig. 5:Results after executing 100 times our DP method to the same dataset.

noise. At first glance we cannot distinguish the differences, but if we overlap them into the same figure, we can identify the noise added by the DP method. Fig. 5 shows this situation and represents these 100 queries. Just note the **Y**-axis represents patient identifiers. The **X**-axis represents the time.

V. Threats to validity

This paper is subjected to threats to construct validity, internal validity, external validity and reliability

A. Construct validity

We have defined a shift left strategy for considering privacy within systems' software developments. We have used a standard software development process as a reference, so any company can apply our approach. In addition, we have applied a differential privacy approach to private any data related to patient identifiers. These elements are just a couple of requirements for this company. Our approach applies this strategy considering requirements since the beginning of any software development. This means that in our case, we apply a DP method to any field related to PII. Obviously, we can apply other anonymisation techniques, but as far as we know, DP is the appropriate approach against database reconstruction attacks as stated previously. It is relevant to stress the fact that the Census Bureau of the United Stated of America has moved to differential privacy as the underlying approach for protecting data. Authors assume a bias during this constructing process.

B. Internal validity

This paper does not gather a wide set of experiments, and it just reports a single use case. However, we can consider this "one shot case study" representative enough as an example where the shift left approach with a DP method is applied. The use case described in this paper reflects a real situation where a company wants to tackle some privacy issues in a software development context. In fact, this company was worried about the logs gathers and exchanged between the NHSs. Nowadays regulations are stressing privacy aspects, and software developments processing PII must consider all these requirements. This paper represents an example, and future case studies will be reported in this sense.

C. External validity

External validity is related to what extend the outcomes of our research can be generalized. In this sense, our focus is based on the ISO/IEC 29110 software implementation activities which is a standard software development process. This standard is not widely adopted by large companies, but it defines the basic phases for developing a software system. This process has been modified in order to analyse the related datasets with a CRISP-DM based method which is an open

standard process model, and it is considered as a standard *de facto*. While analyzing the dataset we identified the sensitive data related to patients, and we applied a DP method which is a method suggested by different frameworks. Fig 4 shows the differences between the normal versus the noisy databases. This resulting noisy database is resistant to database recovery attacks, and the utility of the resulting dataset is higher than other anonymization methods such as k-anonymity because no generalisation techniques are used.

D. Reliability

The use case describes a VSE developing a software connecting two different NHS. The resulting solution is a critical system managing healthcare records, and it must comply with regulations such as GDPR, and privacy must be considered since the inception of a system. This kind of systems manage PII and therefore we need to set up specific methods for enhancing privacy. Data is critical and the CRISP-DM defines a set of activities for analysing data. Our approach integrates the CRISP-DM phase and the ISO/IEC 29110 software implementation activities. In fact, our method stresses the data analysis and the use of privacy enhancing tools. K-anonymity based methods apply generalisation approach, and therefore the utility of the resulting dataset is reduced. In addition, these systems are compromised by *database recovery attacks*, and they are heavily dependent on the dataset. A dataset features a level k of anonymity when the information for each person cannot be distinguished from at least k – 1 individuals included in the dataset. Therefore, a larger dataset with similar entries is required.

In our use case, we use a DP based technique because it is resistant to database recovery attacks, and the utility of the resulting dataset is higher than k-anonymity because no generalisation techniques are used.

In addition, instead of defining a new profile for the ISO/IEC 29110, we identify some steps to be done during the ISO/IEC 29110 software implementation's activities when VSEs must deal with critical systems where privacy must be preserved as much as possible. As described before, VSEs must perform an analysis of the dataset that is managed by the resulting solution. It is not always evident for a VSE to perform these activities at the same time. However, by adding these activities during the software development process for VSEs, we are adding a "shift-left" strategy for including privacy since the inception of the software system. Instead of waiting until the end of the product development or until the product delivery for testing privacy concerns, our proposal includes a set of activities for analysing the dataset and a DP algorithm.

VI. CONCLUSIONS

This paper summarizes a software development process for VSEs with a "shift-left" strategy for including a privacy preserving technique during the whole process. At the beginning of this paper, we defined a set of research questions. The RQ1: *How can be integrated a privacy preserving method within the software implementation activities of VSEs?* This paper summarizes a strategy for adding a privacy preserving approach at the beginning of the development process. Traditionally, this is the so-called "shift-left" strategy. VSEs are special entities because they need to balance the tradeoffs between investments and returns, and they are required to think twice on what activities their resources are going to be dedicated. The ISO/IEC 29110 provides the means for defining a profile, but instead of defining a new profile and overwhelming VSEs with more bureaucracy, we opt for adding specific steps during their developments processes.

The second research question is RQ2: *Is there any possibility to identify Personally Identifiable Information within an existing system exchanging medical records?* Medical records are tightly related to personal information and therefore it is a critical asset to be carefully treated. Therefore, we have added a data analysis method (CRISP-DM) for analyzing the dataset managed by the resulting solution.

The third research question is RQ3: *What recommendations can be provided from a real use case?* This research question describes a VSE experience developing a software solution for integrating two OpenNCP systems. We have integrated a DP approach into the ISO/IEC 29110 software implementation's activities.

As future work, we are investigating on how to fulfill with the whole GDPR requirements.

ACKNOWLEDGEMENT

This work has been partially supported by the Basque Governement (SPRI) project called Trustind - Creating Trust In The Industrial Digital Transformation (KK-2020/00054).

REFERENCES

[1] J. Jang-Jaccard y S. Nepal, «A survey of emerging threats in cybersecurity», *Journal of Computer and System Sciences*, vol. 80, n.º 5, pp. 973-993, ago. 2014, doi: 10.1016/j.jcss.2014.02.005.

[2] J. McManus y D. Mohindra, «The CERT Sun Microsystems Secure Coding Standard for Java», *CERT, USA*, 2009.

[3] THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, «Directive 95/46/EC (General Data Protection Regulation)». Official Journal of the European Union, abr. 27, 2016. Accedido: jun. 25, 2019. [En línea]. Disponible en: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[4] National Institute of Standards and Technology (NIST), «Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)». https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf

[5] ENISA, «Information exchange and communication - What to share». https://www.enisa.europa.eu/topics/national-cyber-security-strategies/information-sharing/isacs-toolkit/tools/build/information-exchange-and-communication/what-to-share (accedido dic. 30, 2020).

[6] ISO/IEC, «ISO/IEC TR 29110-1:2011», ISO/IEC, 2011. [En línea]. Disponible en: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=51150

[7] X. Larrucea y B. Fernandez-Gauna, «A mapping study about the standard ISO/IEC29110», *Computer Standards & Interfaces*, abr. 2019, doi: 10.1016/j.csi.2019.03.005.

[8] X. Larrucea, I. Santamaria, y B. Fernandez-Gauna, «Managing security debt across PLC phases in a VSE context», *Journal of Software: Evolution and Process*, jul. 2019, doi: 10.1002/smr.2214.

[9] M.-L. Sanchez-Gordon, A. de Amescua, R. V. O'Connor, y X. Larrucea, «A standard-based framework to integrate software work in small settings», *Computer Standards & Interfaces*, vol. 54, pp. 162-175, nov. 2017, doi: 10.1016/j.csi.2016.11.009.

[10] M. Elhoseny, A. Abdelaziz, A. S. Salama, A. M. Riad, K. Muhammad, y A. K. Sangaiah, «A hybrid model of Internet of Things and cloud computing to manage big data in health services applications», *Future Generation Computer Systems*, vol. 86, pp. 1383-1394, sep. 2018, doi: 10.1016/j.future.2018.03.005.

[11] THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION, «Directive 2011/24/EU of the European Parliament and of the Council of 9 March 2011 on the application of patients' rights in cross-border healthcare». sep. 05, 2011. Accedido: jun. 25, 2019. [En línea]. Disponible en: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32011L0024

[12] X. Larrucea y I. Santamaría, «Correlations study and clustering from SPI experiences in small settings», *Journal of Software: Evolution and Process*, p. e1989, sep. 2018, doi: 10.1002/smr.1989.

[13] X. Larrucea y I. Santamaria, «Survival studies based on ISO/IEC29110: Industrial experiences», *Computer Standards & Interfaces*, vol. 60, pp. 73-79, 2018, doi: https://doi.org/10.1016/j.csi.2018.04.006.

[14] A.-L. Mesquida y A. Mas, «A project management improvement program according to ISO/IEC 29110 and PMBOK (R)», *JOURNAL OF SOFTWARE-EVOLUTION AND PROCESS*, vol. 26, n.º 9, SI, pp. 846-854, sep. 2014, doi: 10.1002/smr.1665.

[15] Executive Office of the President- White House, «Safeguarding Against and Responding to the Breach of Personally Identifiable Information», may 2007. https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2007/m07-16.pdf (accedido may 07, 2020).

[16] A. Wheeler y M. Winburn, «Application Data in the Cloud», en *Cloud Storage Security*, Elsevier, 2015, pp. 23-55. doi: 10.1016/B978-0-12-802930-5.00002-2.

[17] S. Weiss, «Privacy threat model for data portability in social network applications», *International Journal of Information Management*, vol. 29, n.º 4, pp. 249-254, ago. 2009, doi: 10.1016/j.ijinfomgt.2009.03.007.

[18] A. Narayanan y V. Shmatikov, «Myths and fallacies of "Personally Identifiable Information"», *Communications of the ACM*, vol. 53, n.º 6, pp. 24-26, jun. 2010, doi: 10.1145/1743546.1743558.

[19] S. J. Y. Go, R. Guinto, C. A. M. Festin, I. Austria, R. Ocampo, y W. M. Tan, «An SDN/NFV-Enabled Architecture for Detecting Personally Identifiable Information Leaks on Network Traffic», en *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)*, Zagreb, Croatia, jul. 2019, pp. 306-311. doi: 10.1109/ICUFN.2019.8806077.

[20] Y. Liu, H. H. Song, I. Bermudez, A. Mislove, M. Baldi, y A. Tongaonkar, «Identifying Personal Information in Internet Traffic», en *Proceedings of the 2015 ACM on Conference on Online Social Networks - COSN '15*, Palo Alto, California, USA, 2015, pp. 59-70. doi: 10.1145/2817946.2817947.

[21] L. Sweeney, «k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY», *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, n.º 05, pp. 557-570, oct. 2002, doi: 10.1142/S0218488502001648.

[22] K. LeFevre, D. J. DeWitt, y R. Ramakrishnan, «Mondrian Multidimensional K-Anonymity», en *22nd International Conference on Data Engineering (ICDE'06)*, Atlanta, GA, USA, 2006, pp. 25-25. doi: 10.1109/ICDE.2006.101.

[23] X. Li y Z. Zhou, «A generalization model for multi-record privacy preservation», *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, n.º 7, pp. 2899-2912, jul. 2020, doi: 10.1007/s12652-019-01430-y.

[24] A. Machanavajjhala, D. Kifer, J. Gehrke, y M. Venkitasubramaniam, «*L* -diversity: Privacy beyond *k* -anonymity», *ACM Transactions on Knowledge Discovery from Data*, vol. 1, n.º 1, p. 3, mar. 2007, doi: 10.1145/1217299.1217302.

[25] N. Li, T. Li, y S. Venkatasubramanian, «t-Closeness: Privacy Beyond k-Anonymity and l-Diversity», en *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, abr. 2007, pp. 106-115. doi: 10.1109/ICDE.2007.367856.

[26] X. Xiao, «Privacy Preserving Data Publishing», PhD Thesis, The Chinese University of Hong Kong (People's Republic of China), 2008.

[27] O. Gkountouna, S. Angeli, A. Zigomitros, M. Terrovitis, y Y. Vassiliou, «k m -Anonymity for Continuous Data Using Dynamic Hierarchies», en *Privacy in Statistical Databases*, vol. 8744, J. Domingo-Ferrer, Ed. Cham: Springer International Publishing, 2014, pp. 156-169. doi: 10.1007/978-3-319-11257-2_13.

[28] C. Dwork, «Differential Privacy», en *Automata, Languages and Programming*, vol. 4052, M. Bugliesi, B. Preneel, V. Sassone, y I. Wegener, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1-12. doi: 10.1007/11787006_1.

[29] S. Chen, A. Fu, J. Shen, S. Yu, H. Wang, y H. Sun, «RNN-DP: A new differential privacy scheme base on Recurrent Neural Network for Dynamic trajectory privacy protection», *Journal of Network and Computer Applications*, vol. 168, p. 102736, oct. 2020, doi: 10.1016/j.jnca.2020.102736.

[30] D. Sánchez, J. Domingo-Ferrer, S. Martínez, y J. Soria-Comas, «Utility-preserving differentially private data releases via individual ranking microaggregation», *Information Fusion*, vol. 30, pp. 1-14, jul. 2016, doi: 10.1016/j.inffus.2015.11.002.

[31] A. Friedman y A. Schuster, «Data mining with differential privacy», en *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*, Washington, DC, USA, 2010, p. 493. doi: 10.1145/1835804.1835868.

[32] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, y L. Xiong, «Publishing set-valued data via differential privacy», *Proceedings of the VLDB Endowment*, vol. 4, n.º 11, pp. 1087-1098, ago. 2011, doi: 10.14778/3402707.3402744.

[33] S. Garfinkel, J. M. Abowd, y C. Martindale, «Understanding database reconstruction attacks on public data», *Communications of the ACM*, vol. 62, n.º 3, pp. 46-53, feb. 2019, doi: 10.1145/3287287.

[34] R. Wirth, «CRISP-DM: Towards a standard process model for data mining», en *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000, pp. 29–39.

[35] R.-C. Härting y A. Sprengel, «Cost-benefit considerations for Data Analytics - An SME-Oriented Framework enhanced by a Management Perspective and the Process of Idea Generation», *Procedia Computer Science*, vol. 159, pp. 1537-1546, 2019, doi: 10.1016/j.procs.2019.09.324.

[36] S. Huber, H. Wiemer, D. Schneider, y S. Ihlenfeldt, «DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model», *Procedia CIRP*, vol. 79, pp. 403-408, 2019, doi: 10.1016/j.procir.2019.02.106.

[37] C. Dwork, F. McSherry, K. Nissim, y A. Smith, «Calibrating Noise to Sensitivity in Private Data Analysis», en *Theory of Cryptography*, vol. 3876, S. Halevi y T. Rabin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265-284. doi: 10.1007/11681878_14.

[38] N. Rodríguez-Barroso *et al.*, «Federated Learning and Differential Privacy: Software tools analysis, the Sherpa.ai FL framework and methodological guidelines for preserving data privacy», *Information Fusion*, vol. 64, pp. 270-292, dic. 2020, doi: 10.1016/j.inffus.2020.07.009.

[39] European Commission, «OpenNCP». https://ec.europa.eu/cefdigital/wiki/display/EHNCP (accedido oct. 01, 2018).

[40] M. Staffa *et al.*, «An OpenNCP-based Solution for Secure eHealth Data Exchange», *Journal of Network and Computer Applications*, vol. 116, pp. 65-85, ago. 2018, doi: 10.1016/j.jnca.2018.05.012.

[41] M. Lezzi, M. Lazoi, y A. Corallo, «Cybersecurity for Industry 4.0 in the current literature: A reference framework», *Computers in Industry*, vol. 103, pp. 97-110, dic. 2018, doi: 10.1016/j.compind.2018.09.004.

[42] X. Larrucea, M. Moffie, S. Asaf, y I. Santamaria, «Towards a GDPR compliant way to secure European cross border Healthcare Industry 4.0», *Computer Standards & Interfaces*, vol. 69, p. 103408, mar. 2020, doi: 10.1016/j.csi.2019.103408.

[43] J. J. Hathaliya, S. Tanwar, S. Tyagi, y N. Kumar, «Securing electronics healthcare records in Healthcare 4.0 : A biometric-based approach», *Computers & Electrical Engineering*, vol. 76, pp. 398-410, jun. 2019, doi: 10.1016/j.compeleceng.2019.04.017.

[44] Health Level Seven, «The HL7 v3 CDA Release 2 Standard, 04 2017», oct. 02, 2018. http://www.h17.org/implement/standards/product_brief.cfm?product_id=7 (accedido oct. 02, 2018).

http://doi.org/10.18239/jornadas_2021.34.59

Agente de recomendación para mejorar la privacidad de los usuarios cuando usan gestión de identidades federada

Carlos Alberto Villarán y Marta Beltrán
ORCID: 0000-0003-2602-6850 y 0000-0002-1689-7479
ETSII, Universidad Rey Juan Carlos
28933 Móstoles, Madrid
carlos.villaran@urjc.es y marta.beltran@urjc.es

***Resumen*—Los esquemas de gestión de identidades y accesos federados se utilizan habitualmente en entornos web, cloud, móviles o del Internet de las Cosas porque permiten a los recursos, aplicaciones o servicios delegar la responsabilidad de la autenticación y de la autorización en proveedores de identidades externos. Además, facilitan la experiencia de los usuarios evitando que tengan que gestionar una cuenta independiente para cada recurso, aplicación o servicio. Pero una de las limitaciones principales para la evolución de estos esquemas es la desconfianza que los usuarios finales tienen en lo que se refiere a la gestión de su privacidad y a la protección de sus datos personales, especialmente dado que la mayor parte de los proveedores de identidades son "sociales"(Facebook, Google, Twitter, Apple, etc.) y proporcionan el servicio de manera aparentemente gratuita. En este trabajo se propone la utilización de un nuevo agente de recomendación, el Privacy Advisor, en los flujos de autenticación y autorización. En concreto se especifican las capacidades que debe poseer este nuevo agente, su arquitectura, su integración con las especificaciones federadas existentes y el diseño de los módulos que lo componen. También se realiza la implementación de un primer prototipo sencillo que permite validar la propuesta realizada en casos de uso reales basados en OpenID Connect.**

***Index Terms*—Esquemas federados, Gestión de identidades y accesos, OpenID Connect, Privacidad, Protección de datos**

Tipo de contribución: *Investigación original*

I. Introducción

La gestión de identidades y accesos es uno de los aspectos más importantes en la ciberseguridad actual. Las tradicionales soluciones distribuidas y centralizadas para este problema han dado paso, en los últimos años, a soluciones federadas en las que los propietarios de recursos, aplicaciones y servicios así como los usuarios finales confían en un tercero, un proveedor de identidades, para que resuelva la Identificación, la Autenticación, la Autorización y la Auditoría (IAAA) en cada acceso. Además, aunque en la identidad auto-soberana se están realizando avances, sin un uso todavía muy generalizado, se va a tener que depender durante bastante tiempo de un tercero que provea la identidad en el uso de servicios. Esta evolución ha sido necesaria dado el contexto en el que surgen estos problemas en la actualidad, cada vez más heterogéneo, escalable y distribuido.

Aunque el esquema federado presenta muchas ventajas desde el punto de vista de la seguridad (y algún inconveniente como algunos trabajos de investigación han puesto de manifiesto [1], [2]), comienzan a surgir reservas en cuanto a las amenazas que pueden suponer para la privacidad de los usuarios. Especialmente teniendo en cuenta que los principales proveedores de identidades en la actualidad son grandes multinacionales tecnológicas como Facebook, Google, Twitter, Apple o las operadoras de telecomunicaciones.

De hecho, diferentes trabajos de investigación como [2], [3], [4] y [5] han demostrado también las amenazas que estos proveedores de identidades pueden materializar en lo que se refiere a la privacidad de los usuarios finales: falta de control en la compartición de datos personales con terceros, brechas de datos, elaboración de perfiles y etiquetado, geolocalización, etc.

En este trabajo se propone la introducción de un nuevo agente en las federaciones de identidades, el Privacy Advisor. La función de este nuevo agente es garantizar, desde fuera y sin obligar a realizar cambios importantes en las especificaciones que se usan en la actualidad (OAuth [6], OpenID Connect [7], etc.), la privacidad y la protección de datos de los usuarios finales. En concreto, en este trabajo se asocia la función del Privacy Advisor con la de un sistema de recomendación, es decir, se intenta dotar a los usuarios finales de herramientas que les ayuden en su toma de decisiones. La idea es que el usuario esté en el centro de estas decisiones y que pueda hacerse responsable de la protección de su privacidad. Las principales contribuciones de esta investigación son: 1) Proponer una arquitectura modular para el Privacy Advisor y la manera de integrarlo, prácticamente sin esfuerzo, con las especificaciones federadas actuales para facilitar su adopción. 2) Especificar como principales funcionalidades del Privacy Advisor la revisión de patrones de diseño, la comprobación de certificaciones de privacidad, la revisión de políticas de uso y privacidad, la utilización de sistemas de reputación y el análisis de cumplimiento de mejores prácticas de seguridad y privacidad. 3) Implementar un primer prototipo del Privacy Advisor que incorpore estas funcionalidades cuando se trabaja con OpenID Connect (la base de las soluciones que ofrecen actualmente Facebook, Google, Twitter o Apple por mencionar algunos de los ejemplos más significativos) y validarlo.

El resto de este artículo se estructura de la siguiente manera. En la sección II se resume el estado del arte necesario para comprender el resto del artículo y, a partir de los conceptos más importantes que se introducen, se discute la motivación de la investigación realizada. En la sección III, se identifican las capacidades del Privacy Advisor a alto nivel, se propone una arquitectura modular para este nuevo agente y se discute la mejor manera de integrar su uso con los esquemas federados

más extendidos en la actualidad. La sección IV profundiza en la funcionalidad de los módulos que componen el Privacy Advisor, mientras que la sección V se centra en analizar la implementación del primer prototipo y en validarla. Finalmente, la sección VI presenta las principales conclusiones obtenidas y destaca las líneas de trabajo futuro más interesantes.

II. Estado del arte y motivación

II-A. Esquemas de gestión de identidad federada

Los esquemas de gestión de identidad federada delegan la gestión de la identidad de los usuarios en proveedores de identidad (*Identity Provider*, IdP). En este tipo de esquemas, un usuario que quiere acceder a un servicio (normalmente denominado *Relying Party*, RP) que requiere autenticación y/o autorización, puede hacerlo sin necesidad de crearse una cuenta para ese servicio, simplemente realizando su autenticación en cualquiera de los IdP donde ya esté dado de alta. Siempre y cuando la RP soporte la integración con ese IdP en concreto.

Este tipo de modelo evita que el usuario tenga que recordar un nombre de usuario y una contraseña para cada servicio que utiliza. Puede utilizar la identidad dada de alta en un IdP, mejorando no sólo su experiencia como usuario, sino también la seguridad, ya que la superficie de ataque es menor y además tiene que recordar un menor número de contraseñas, por lo que éstas pueden ser de mayor complejidad. Además, las contraseñas no se comparten con ningún servicio, el usuario las introduce directamente en la interfaz del IdP, que es donde se validan. Un adversario que quisiera robar credenciales de acceso estaría forzado a hacerlo en los IdP, si comprometiera un servicio, no obtendría la contraseña de los usuarios. Desde el punto de vista de las RP, estos esquemas también tienen ventajas, ya que permiten proporcionar flexibilidad a los usuarios y además, externalizan una función que suele ser compleja y que implica una responsabilidad, en principio, en entidades que tienen la capacidad de asumirla corriendo menos riesgos.

La especificación OAuth 2.0 [6] resuelve la autorización con un esquema federado. Se basa en la emisión de tokens de acceso (*Access Token*) por el IdP, que tienen asociados unos usos (*scopes*) restringidos. En su RFC se han definido cinco flujos de comunicación distintos dependiendo de cómo sea el escenario del consumo del servicio, donde los dos más utilizados son el *Authorization Code Grant* y el *Implicit Grant*. En el *Authorization Code Grant*, la RP redirige al *user agent* al IdP para que se autentique y autorice el consumo del servicio según los *scopes* que necesite. El IdP devuelve un código de autorización al usuario, que éste re-envía automáticamente a la RP. Por último, la RP envía este código al IdP junto con sus credenciales de autenticación y el IdP le devuelve el token de acceso. Este token solo podrá utilizarse para los *scopes* que se hayan definido cuando el usuario lo autorizó. El *Implicit Grant* es idéntico, salvo que en lugar de devolver un código de autorización, el IdP directamente devuelve el token de acceso al usuario que se lo envía automáticamente a la RP. Este flujo está en desuso porque puede provocar problemas de seguridad.

OAuth 2.0 resuelve la autorización, pero no realiza gestión de identidades ni resuelve la autenticación. Por este motivo surgió la especificación OpenID Connect [7]. El flujo que define es idéntico a los utilizados en OAuth 2.0 salvo porque el IdP no sólo puede devolver un token de acceso, sino también un token con la identidad del usuario (*ID Token*) que permite que sea autenticado. Este *ID Token* es un JSON Web Token (JWT) [8] que puede firmarse y/o cifrarse.

II-B. Privacidad en esquemas de gestión de identidad federados

Los sistemas de gestión de identidad federada presentan una serie de ventajas relativas a la seguridad que ya se han comentado. Sin embargo, diferentes trabajos e investigaciones han demostrado una serie de carencias desde el punto de vista de la privacidad. Hay que tener en cuenta que el IdP concentra una gran cantidad de información personal de sus usuarios (normalmente sin cifrar), que puede verse involucrada en una brecha de datos. Además, este tipo de esquemas no contempla la privacidad en el proceso de propagación de los datos hacia terceros y tampoco tiene definido un método para evaluar al servicio (la RP) al que el IdP va a entregar los datos del usuario. Existe, por tanto, un descontrol en la forma en la que se propagan los datos personales de los usuarios.

En [4] y en [9] se proponen distintas soluciones de cifrado de los datos personales en el IdP para que éste no pueda conocer los datos personales de los usuarios que confían en él. Sin embargo, estas soluciones implican la implementación de un nuevo protocolo, por lo que no se han extendido ya que, tanto las RP como los IdP actuales, deberían implementarlo y cambiar completamente la manera en la que trabajan en la actualidad. En [5] y [10] se centran en resolver el cifrado de los datos de usuarios en el IdP haciendo uso de protocolos estándar, en [5] con SAML 2.0 y en [10] con OpenID 2.0. Este tipo de soluciones protegen la privacidad de los datos de los usuarios cuando están almacenados en el IdP pero no se tiene en cuenta el proceso de propagación de datos personales a terceros.

[3] y [11] añaden al uso de cifrado un sistema de reputación. En [11] el usuario puede definir las políticas de privacidad para el control de la propagación de los datos personales. Adicionalmente, se informa al usuario de los riesgos del uso del servicio mediante un sistema de reputación basado en el cumplimiento de estas políticas. Sin embargo, se realizan modificaciones en el protocolo OpenID Connect para añadir estas nuevas funcionalidades y en [3] se indica que la configuración de políticas es compleja y que la definición de una única política no podría cubrir todas las necesidades de un usuario. Por otro lado, en [3] el usuario cifra sus datos personales antes de enviarlos al IdP, de tal forma que éste no conozca su contenido. Pero se añade, de nuevo, un sistema de reputación para controlar la difusión de los datos personales de los usuarios. El IdP es el encargado de recopilar la información del sistema de reputación para mostrársela al usuario cuando acepta autenticarse. Por lo tanto, el usuario debe confiar en el IdP, que tiene que cumplir el modelo *honest but curious*, ya que tiene capacidades para alterar la información devuelta por el sistema de reputación.

En [2], no sólo se cifran los datos personales de los usuarios en el IdP, si no que se propone el uso de un nuevo agente, el *Privacy Arbiter*, como sistema de protección de la privacidad de los usuarios. Una ventaja de la propuesta

realizada en este trabajo es que la incorporación de este nuevo agente prácticamente no altera el flujo definido por OpenID Connect, es decir, su uso no requiere modificaciones significativas en la definición del protocolo, lo que facilita su adopción. El presente trabajo se basa en este enfoque pero intentando superar las limitaciones que todavía presenta, ya que se considera que las funcionalidades del *Privacy Arbiter* no cubren muchas de las necesidades que surgen en los escenarios actuales, que necesitan que el usuario final se involucre en la toma de las decisiones que afectan a su propia privacidad.

II-C. Motivación

Como se puede observar, los esquemas federados hacen que los usuarios finales se registren en los proveedores de identidades proporcionando información personal más o menos sensible. A partir de ese momento, los usuarios pierden el control de lo que ocurre con esos datos y los proveedores no son transparentes en los procesos de transferencia o compartición con otros agentes que forman parte de la federación. Por ejemplo, puede ser que los datos de un usuario se verifiquen en un tercero (una dirección de email, por ejemplo) o que sean compartidos con una RP (nombre, apellidos o dirección postal, por ejemplo, para auto-completar un formulario).

A esto hay que sumarle que no todos los agentes que forman parte de una federación presentan los mismos niveles de respeto a la privacidad y la protección de datos. Una RP, por ejemplo, puede cumplir con los principios de minimización o desvinculación, mientras que otra no. Y por último hay que mencionar que dentro de una federación, y dada esta heterogeneidad, puede ocurrir que una RP o un proveedor de identidades aprovechen la información que se obtiene de los flujos de autenticación y de autorización para realizar perfilado de los usuarios o incluso geo-localización. Sin embargo, otros agentes son completamente contrarios a este tipo de prácticas o las tienen prohibidas por el espacio geográfico en el que operan o en el que se encuentra su sede.

Los esquemas de gestión de identidades y accesos no incorporan en su especificación capacidades que lidien con estas amenazas de manera explícita. No parece que pretendan hacerlo en el futuro, ya que muchos aspectos que tienen que ver con la privacidad y la protección de datos se dejan abiertos y en manos de los proveedores de identidades y de las RPs. En este punto es en el que diferentes investigaciones han propuesto soluciones parciales y específicas para proteger los datos personales de los usuarios mediante diferentes mecanismos de cifrado. Aunque en muy pocos casos se resuelve la minimización en la recogida de datos, se resuelve el control de los usuarios finales sobre sus propios datos o se garantiza la transparencia de los flujos de estos datos.

En este trabajo se aborda la propuesta de mecanismos genéricos y ambiciosos que añadan todas estas capacidades desde fuera, desde agentes externos que ofrezcan a los usuarios finales la posibilidad de estar en el centro de las decisiones importantes. Los requisitos que se persiguen para estos nuevos mecanismos son la facilidad de uso para todos los agentes de la federación (sobre todo para los usuarios finales, que no querrán sacrificar usabilidad por privacidad), facilidad de integración con los flujos estándar (para que los proveedores de identidad y las RPs no tengan que cambiar de manera sustancial sus flujos de trabajo e implementaciones) y escalabilidad (dado el potencial tamaño de las federaciones en los contextos en los que se usan estos esquemas en la actualidad).

III. Privacy Advisor

III-A. Capacidades

El *Privacy Advisor* (PAdv) realiza funciones de recomendación relativas a la privacidad con respecto al recurso, aplicación o servicio que el usuario quiere utilizar (la RP). El PAdv tiene que proporcionar información útil al usuario de manera que éste pueda conocer los niveles de privacidad y protección de datos que le ofrecen cada una de las partes implicadas en el proceso y tomar así las decisiones más adecuadas.

Las capacidades de este nuevo agente se pueden resumir en:

1. Capacidad para recopilar información actualizada y completa sobre los niveles de privacidad y protección de datos que ofrece un recurso, aplicación o servicio desde diferentes fuentes de conocimiento complementarias.
2. Capacidad para mostrar esta información al usuario de manera sencilla e ilustrativa, para que le sea realmente útil en su toma de decisiones.
3. Capacidad de ofrecer un mayor detalle sobre los resultados de la recomendación a aquellos usuarios que así lo demanden (los más avanzados o más concienciados). Es decir, capacidad para trabajar con diferentes granularidades o niveles de detalle en la recomendación que ofrece.

Estas capacidades permiten asegurar la transparencia y el control por parte del usuario, dos de los principios o pilares fundamentales para la privacidad. Este trabajo, por motivos de espacio, se centra fundamentalmente en el primer pilar mencionado.

III-B. Arquitectura propuesta

Para garantizar estas capacidades, este trabajo propone una arquitectura completamente modular, tal y como se puede ver en la figura 1. De esta forma, cada módulo se puede ejecutar de forma distribuida e incluso diferentes módulos pueden pertenecer o ser ofrecidos por compañías u organizaciones distintas. Además, el proceso de añadir nuevos módulos es mucho más sencillo, ya que se procura que no exista un acoplamiento fuerte entre ellos.

Como se puede observar en la figura 1, se propone un Módulo Principal que es el encargado de recibir la petición de recomendación acerca de un determinado servicio, de realizar peticiones al resto de módulos y de devolver una respuesta al usuario. Para ello, necesita enviar consultas a cada módulo solicitando información sobre el servicio investigado. Este módulo almacena también las preferencias relativas a privacidad de los usuarios que usan el *Privacy Advisor*. El resto de módulos propuestos son los siguientes: Revisión de patrones de diseño, Certificaciones de privacidad, Revisión de políticas de uso y privacidad, Sistema de reputación y Cumplimiento de consideraciones seguridad y privacidad. El diseño propuesto para cada uno de ellos se explica más adelante, en la sección IV de este artículo.

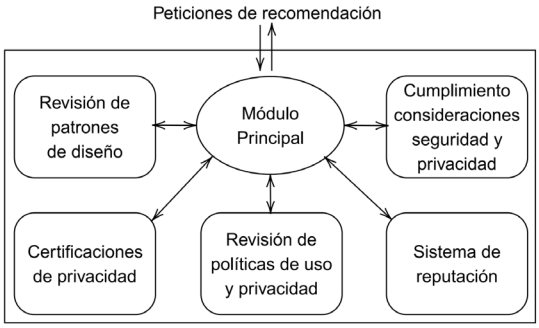


Figura 1. Arquitectura del Privacy Advisor

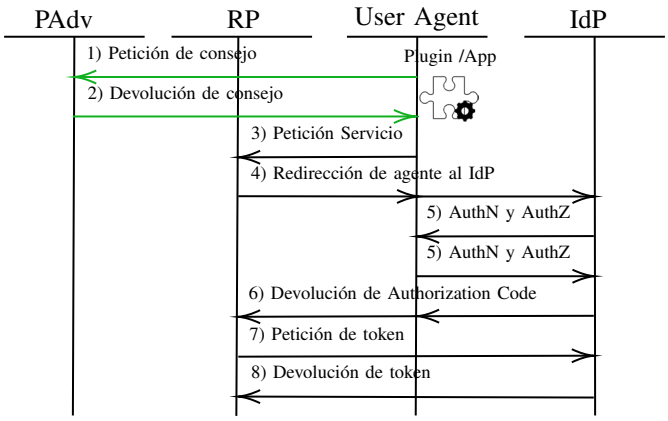


Figura 2. Ejemplo de flujo de autenticación con PAdv y OpenID Connect

III-C. Integración con los esquemas de gestión de identidades federados

La incorporación del *Privacy Advisor* no modifica los flujos definidos por las especificaciones federadas. Con nuestra propuesta, cuando un usuario quiere utilizar una de ellas para conectarse a una RP, se generan dos flujos. Por un lado, se genera una petición al *Privacy Advisor* solicitando una recomendación para el usuario (paso 1 de la figura 2). Esta petición se genera mediante un *plugin* instalado en el navegador o mediante una aplicación, en el caso de los dispositivos móviles. El PAdv responde con la recomendación al usuario sobre el servicio solicitado (paso 2 de la figura 2). Estos dos pasos pueden ejecutarse anteriormente o en paralelo al proceso de autenticación (paso 3 y posteriores de la figura 2), siguiendo el estándar de OpenID Connect para obtener el *token*. En el ejemplo mostrado en la figura, se usa un flujo con *Authorization Code*, pero esta propuesta se podría utilizar con todos los flujos especificados en OpenID Connect y en otras especificaciones similares. De hecho, se podría utilizar también con Mobile Connect y con el resto de especificaciones federadas, ya que se ejecuta por completo antes de comenzar la autenticación del usuario y sin que ésta se vea afectada.

Por lo tanto, el *Privacy Advisor* puede trabajar de dos maneras en función de las preferencias de los usuarios. En la primera, hasta que el usuario no revisa la recomendación y acepta no puede iniciar el proceso de autenticación. Este proceso es más intrusivo pero garantiza que el usuario vea la recomendación de privacidad y la tenga en cuenta antes de seguir con el flujo. En la segunda, la recomendación del PAdv se realiza en paralelo con la autenticación (por ejemplo, en otra pestaña o ventana del navegador). De esta forma, el proceso es completamente independiente al proceso de autenticación pero no obliga al usuario a revisar la recomendación.

IV. Diseño de los módulos que componen el Privacy Advisor

Según la arquitectura propuesta, todos los módulos del *Privacy Advisor* se han diseñado de manera independiente, de tal forma que se puedan añadir, modificar o sustituir de manera individual sin afectar al funcionamiento de este agente. Cada uno de los módulos recibe solicitudes de diferente naturaleza sobre una RP concreta y devuelve un resultado a un módulo coordinador de todo el proceso, el Módulo Principal.

IV-A. Revisión de patrones de diseño

Distintos trabajos han puesto de manifiesto la existencia de patrones de diseño que intentan confundir al usuario final en lo que se refiere a su privacidad y a la protección de sus datos para el beneficio de la organización a la que pertenece el recurso, servicio o aplicación que desea utilizar. Estos patrones se conocen como patrones oscuros o *dark patterns* [12]. Se apoyan, fundamentalmente, en la redacción de textos ambiguos, en el uso de opciones pre-marcadas o de colores con el fin de distraer o de ocultar opciones que no se desea que se identifiquen con facilidad (como la baja de un servicio, por ejemplo).

En [13] se identifican los patrones más utilizados: interrumpir a los usuarios con tareas que se lanzan de forma repetida (como *pop-ups*), realizar ciertos procesos de una forma más complicada de lo que sería necesario, intentar ocultar información relevante o manipular la interfaz para confundir al usuario y forzarle a realizar acciones no necesarias para poder utilizar el servicio. Además, en [14] se analizan los aspectos psicológicos que hacen que la utilización de este tipo de patrones sea tan efectiva.

Tener este tipo de información disponible de manera explícita para el usuario final en el *Privacy Advisor* permite al usuario estar más atento durante su interacción con la RP, evitando que ésta consiga sus objetivos mediante la utilización de estos patrones. O incluso le permitirá tomar la decisión de no realizar esta interacción y optar por una alternativa más respetuosa con su privacidad. Por lo tanto, este módulo debe identificar claramente qué patrones de diseño oscuros se usan en la RP.

En este trabajo se propone que este módulo identifique los siguientes tipos de patrones oscuros [12]:

- *Tricky Questions*: Preguntas cuya formulación lleva a confusión, de tal forma que la respuesta del usuario no implica lo que él cree.
- *Sneak into Basket*: Cuando al añadir un producto a la cesta se añaden también productos adicionales que no se han escogido de manera explícita. Se suelen utilizar opciones pre-marcadas como *radio button* y *checkbox*.
- *Roach Motel*: Cuando el alta en un servicio es sencilla, pero la baja es muy complicada.
- *Privacy Zuckering*: El usuario comparte más información personal de la que desea mediante alguna artimaña. Por ejemplo, se esconde en los términos de uso que la

información proporcionada en un sitio se cederá a un *data broker*.
- *Price Comparison Prevention*: La comparación de productos es lo suficientemente difícil para que el usuario no pueda decidir objetivamente.
- *Misdirection*: El diseño intenta evitar que el usuario se fije en aquellas cosas que no le interesan al propietario del recurso, servicio o aplicación.
- *Hidden Costs*: Se cargan costes ocultos que solo aparecen en la pantalla de resumen justo antes de realizar el pago.
- *Bait and Switch*: El usuario realiza cierta acción no deseable para él creyendo que hacía otra distinta.
- *Confirmshaming*: Buscan el sentimiento de culpabilidad del usuario en la redacción del texto para evitar que realice ciertas acciones, como darse de baja.
- *Disguised Ads*: Los anuncios están camuflados en el contenido para que el usuario pulse sobre ellos sin darse cuenta.
- *Forced Continuity*: Pretende cobrar un servicio en la tarjeta de crédito tras un periodo de prueba, sin avisar al usuario de ello.
- *Friend Spam*: El usuario da permisos de acceso a su lista de contactos para un servicio y, además, se aprovecha para enviar correos de *spam* a dicha lista.

IV-B. Comprobación de certificaciones de privacidad

Existen diferentes tipos de sellos de confianza y certificaciones que implican que se cumplen unos requisitos en cuanto al respeto a la privacidad y a la protección de los datos. Todos estos sellos y certificaciones suelen obligar a la realización de auditorías externas, tanto iniciales como periódicas, cuando se procede a la renovación de estas certificaciones.

Por ello, la recolección de este tipo de información en el *Privacy Advisor* puede ser útil para el usuario final, demostrando ciertos niveles de cumplimiento en la RP, contrastados además por terceros. La principal función de este módulo es, por tanto, listar los sellos y certificaciones de la RP, si están activos, y qué implica el haberlos conseguido.

En este trabajo proponemos verificar las siguientes certificaciones:

- *European Privacy Seal* (EuroPriSe) [15] certifica que cumple con las normas de la Unión Europea para la protección de datos, focalizándose en la navegación de los usuarios en las partes públicas de la web. La certificación se basa en una serie de dominios, donde auditores acreditados revisan *cookies*, *hosting* de la web, transferencias internacionales de datos, entre otros.
- La empresa TrustArc [16] dispone de nueve certificaciones que cubren distintos aspectos y estándares relativos a privacidad. Estas certificaciones tienen un componente de dependencia con respecto a *TrustArc Privacy & Data Governance ("P&DG") Framework*, que cumple con las guías de privacidad de la OECD, el framework de privacidad de APEC, el Reglamento General de Protección de Datos (RGPD), ISO27001, HIPAA, entre otros.
  - *APEC Cross Border Privacy Rules (CBPR)*: Certifica el libre flujo de PII entre los participantes del programa pertenecientes a América y Asia. Además, se comprueba la correcta protección de la PII en términos de seguridad y privacidad.
  - *Enterprise Privacy & Data Governance Practices*: La organización que lo obtiene, tiene el cumplimiento de los estándares incluidos en el *TrustArc Privacy & Data Governance Framework*.
  - Data Collection: Certifica que la recolección de PII es respetuosa con la privacidad cuando se actúa como tercera parte en una web o aplicación móvil.
- ePrivacy [17] se centra en el ámbito europeo basándose en el RGPD y el *Interactive Advertising Bureau* (IAB *Europe OBA Framework*). Se pueden obtener las siguientes certificaciones:
  - ePrivacySeal: Se centra en la revisión del cumplimiento del RGPD (ePrivacySeal EU) e incluye particularidades de otras leyes de protección de datos como la suiza (ePrivacySeal CH) o la alemana (ePrivacySeal DE).
  - ePrivacyApp: Es equivalente a la certificación ePrivacySeal pero centrada en aplicaciones móviles.

IV-C. Revisión de políticas de uso y privacidad

Las políticas (o condiciones) de uso determinan las condiciones que deben aceptar los usuarios finales para poder acceder a un recurso, servicio o aplicación. Además, la política de privacidad determina los usos que se va dar a los datos proporcionados por el usuario, incluido el tratamiento de los datos personales. Es común ver en un mismo documento la política de uso y la de privacidad. La revisión de este tipo de políticas es, por tanto, fundamental para que el usuario conozca (y, a ser posible, comprenda) qué tipo de consentimiento está proporcionando al aceptar esta política en lo que se refiere al tratamiento de sus datos personales. Sin embargo, en [18] se presentan los resultados de un estudio que demuestra que más del 80 % de los participantes invierten menos de un minuto en la lectura de la política de uso o privacidad. Además, más del 90 % de los usuarios aceptó los términos abusivos.

El *Privacy Advisor* puede ser útil para que el usuario analice este tipo de documentos (o un resumen o conclusión comprensible para él o ella), pero se trata de un reto importante, ya que su redacción suele ser muy formal y no son fácilmente procesables por una máquina. Algunos trabajos previos ya han avanzado en esta dirección. Un buen ejemplo es el proyecto P3P [19], que permite crear políticas de privacidad en un formato estándar y procesable por una máquina. Esta iniciativa no ha tenido una gran aceptación, sin embargo, sí lo están teniendo otras que descargan políticas de privacidad en uso, las procesan con técnicas de *machine learning* y/o mediante discusión y debate en comunidades de expertos y realizan una valoración, anotando sus ventajas e inconvenientes.

En este trabajo, se propone que la funcionalidad de este módulo del *Privacy Advisor* sea tener una valoración general con una etiqueta, en términos de privacidad, de las políticas de la RP sin la necesidad de tener que leer el documento completo. Esta etiqueta se corresponde con un valor de A a E, donde A es la mejor valoración posible y E la peor. Asociado a esta etiqueta hay un color (siguiendo el estándar del semáforo, reconocido internacionalmente): verde

(cumple), amarillo (cumple parcialmente), rojo (no cumple) o gris (no especificado, no se puede determinar, en base a las preferencias configuradas por el usuario final en el Módulo Principal del PAdv). De esta forma, el usuario puede ver la valoración obtenida por la RP y saber rápidamente el nivel de privacidad que ofrece. Además, puede disponer de información adicional sobre los motivos de esta valoración, obteniéndose este detalle a través de servicios especializados.

IV-D. Sistema de reputación

El módulo del sistema de reputación permite conocer el nivel de confianza que se puede tener en el servicio que se desea utilizar, mediante valoraciones de terceros. En [20] se propone un sistema de reputación para SAML 2.0 con la inclusión de un nuevo *statement*. Este sistema permite obtener la valoración de otros RP o IdP. Pero además, podrían existir aplicaciones que se dedicaran de forma exclusiva a reportar el nivel de confianza que tienen registrado.

En este trabajo se propone una solución basada en [20]. El Módulo Principal hace una petición REST de valoración de reputación de un servicio V_i a otros IdP o RP en los que confía. Éstos contestan en formato JSON con una valoración (de 0 a 100) y el detalle de esta valoración. Además, el PAdv no tiene por qué confiar de igual forma en las respuestas de los IdPs o RPs, por lo que se propone un sistema de pesos, con un peso P_i por elemento (con valores entre 0 y 1), donde cuanto mayor es el peso, más fiable es para el PAdv. La valoración final se obtendría con la siguiente fórmula:

$$Reputation = \frac{\sum_{i=1}^{n} V_i * P_i}{n} \quad (1)$$

La funcionalidad de este módulo permite al usuario informado determinar si el servicio que desea utilizar es suficientemente confiable. El usuario podrá ver OK (en verde) o No OK (en rojo), en función de si supera o no el mínimo de reputación, configurable en las preferencias de privacidad del PAdv.

IV-E. Cumplimiento de consideraciones seguridad y privacidad

Las especificaciones de OAuth 2.0 [6], OpenID Connect [7] y JSON Web Token (JWT) [8] en las que se basan los sistemas federados para la gestión de identidades disponen de apartados completos relativos a seguridad y privacidad. En estos apartados se recomiendan mejores prácticas para mitigar amenazas y vulnerabilidades como las encontradas en [1] o [2].

Se incluye también, con respecto a las directivas y normativas europeas, el cumplimiento de eIDAS y PSD2. Hay que recordar que eIDAS es relativa a la certificación de entidades de emisión de certificados a nivel europeo y PSD2 es relativa a los servicios de pago.

La principal funcionalidad de este módulo del *Privacy Advisor* es comprobar que se cumplen las mejores prácticas de seguridad y privacidad definidas en las especificaciones, así como las exigencias que reocgen las normativas mencionadas.

V. Implementación del primer prototipo y validación

Para la validación de la propuesta realizada se ha programado un *plugin* en el navegador Google Chrome que abre una nueva pestaña cuando el usuario pulsa en el enlace de inicio de sesión con el IdP. En esta pestaña aparece toda la información relativa al servicio que va a utilizar. El flujo de autenticación con el IdP se ejecuta en paralelo, mediante OpenID Connect, y se ha comprobado que no se ve afectado por la incorporación del PAdv.

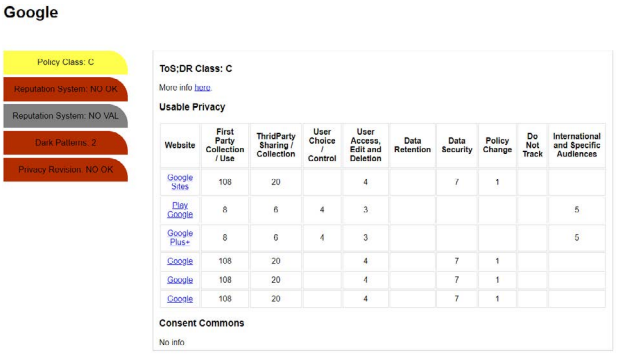


Figura 3. Ejemplo de recomendación realizada por el PAdv

La implementación del PAdv se ha realizado utilizando Python. El Módulo Principal recibe la petición HTTPS del usuario final y, una vez consultados el resto de módulos, compone la página HTML con la información para mostrar al usuario, tal y como muestra la figura 3. En la vista inicial se muestra únicamente la parte izquierda, donde se puede ver de forma muy rápida el resumen de los criterios evaluados por el PAdv, incluyendo códigos de color si cumplen (verde), cumplen parcialmente (amarillo), no cumplen (rojo) o no están definidos (gris) los criterios mínimos definidos por el usuario. Si se pulsa en cualquiera de los aspectos evaluados para la RP, se muestra información detallada incluyendo enlaces a las fuentes donde se ha obtenido la información o el enlace al certificado, por ejemplo, para el módulo de certificaciones de privacidad. En el ejemplo de la figura 3, si se deseara ampliar la información sobre la clasificación obtenida de ToS;DR habría que pinchar en el enlace, que te redirigiría a la figura 4. Lo mismo ocurriría con los enlaces de *Usable Privacy*, que redirigirían al sitio web de donde se ha obtenido la información y donde hay un mayor detalle todavía.

A continuación se describe brevemente cómo se ha implementado cada uno de los módulos del PAdv y cómo se ha podido validar que el diseño propuesto es posible de llevar a la práctica de manera útil, escalable y eficiente:

- Revisión de patrones de diseño: Para su integración en el *Privacy Advisor*, se han categorizado de forma manual un total de 50 patrones oscuros obtenidos del *Hall of shame* propuesto en [12]. Se ha extraído, para cada uno, el tipo de patrón oscuro, la web donde se ha detectado el patrón y un enlace al tweet explicativo en la cuenta del proyecto. Cuando se pregunta a este módulo por cierta RP, éste devuelve el conjunto de patrones oscuros detectados en formato JSON.
- Certificaciones de privacidad: Se ha identificado la información de las compañías certificadas en los propios sitios web de las certificaciones, para tener información lo más completa y actualizada posible de las certificaciones emitidas vigentes. Para EuroPrise y ePrivacy se ha utilizado la librería BeautifulSoap [21] para procesar el

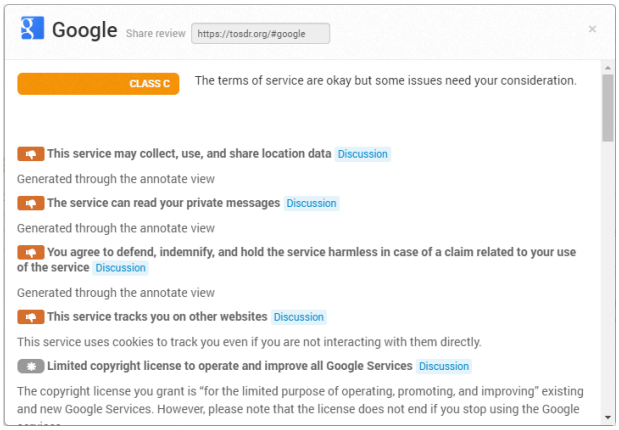


Figura 4. Ejemplo de detalle en ToS;DR

HTML de la página donde están publicadas. Para obtener las certificaciones TrustArc se ha consumido un recurso PHP que devuelve las certificaciones de una compañía y se ha procesado su salida en JSON. Para EU-U.S. and Swiss-U.S. Privacy Shield se realizado la extracción componiendo un JSON vía JQuery [22] en el navegador. Esta información extraída se ha consultado procesando el JSON con Python.

- Revisión de políticas de uso y privacidad: En este primer prototipo del Privacy Advisor se han utilizado tres proyectos ya existentes: *ToS;DR*, *Usable Privacy* y *Consent Commons* para demostrar su viabilidad. En *Terms of Service; Didn't Read* (ToS;DR) [23] una comunidad se encarga de evaluar los términos de servicio y políticas de privacidad para etiquetarlo en una clase. Las clases van desde la letra A hasta la letra E, donde A es la más respetuosa y E la peor. En el caso en el que la información sea insuficiente para obtener una calificación, se indica como "*No class yet*". Además, si se desea, se ofrecen con detalle los aspectos positivos y negativos que han llevado a la comunidad a darle esa calificación. ToS;DR tiene una cantidad limitada de webs analizadas, por lo que no siempre hay una clasificación disponible. Para realizar la integración de esta información, cuando sí está disponible, en el *Privacy Advisor*, ToS;DR expone una API que devuelve la información en formato JSON y además, bajo petición del usuario, se puede ofrecer la versión web para el detalle de una forma visual, tal y como muestra la figura 4.
  Usable Privacy [24] inspecciona las políticas de privacidad de páginas web con revisiones realizadas por personas y mediante técnicas de inteligencia artificial. Las revisiones realizadas por personas han servido como entrenamiento para un modelo basado en *machine learning*. En las revisiones se indican aspectos relacionados con la privacidad de los usuarios y los categorizan en uso de los datos, cesión a terceros o recolección de terceros, elecciones para los usuarios, acceso, modificación y borrado de datos personales, retención de los datos, seguridad en los datos personales, cambios de política, *Do Not Track* y la contemplación de tratamientos especiales para grupos concretos, como niños. Como aspecto negativo, hay que destacar que la última evaluación de las políticas por personas fue en 2015 y en 2017 con inteligencia artificial, por lo que la información disponible no está actualizada.
  Por último, *Consent Commons* [25] permite, mediante iconos, la visualización de los tratamientos relativos al RGPD que se le van a aplicar al usuario. Estos iconos están agrupados en varias categorías, cada una con un propósito:
  - Finalidad del tratamiento: Es una parte fundamental que indica el uso que se va a hacer con la PII.
  - Legitimación: indica el motivo por el que es legítimo el tratamiento de PII.
  - Cesión: Indica qué tipo de cesión de datos se va a realizar, si se fuera a hacer.
  - Derechos: Indica los derechos que tiene el usuario a ejercer.
  - Transferencias internacionales: Indica si se realizan transferencias internacionales de PII y si se realizan a países con un nivel de protección equivalente.
  - Almacenamiento: Indica si los datos se almacenan en el Espacio Económico Europeo.

  El uso de Consent Commons es bajo una licencia Creative Commons, siendo gratuita y siendo obligatorio mencionar el proyecto Consent Commons. Se trata de un proyecto incipiente (del año 2019), por lo que se ha tomado como ejemplo la propia página de Consent Commons, que hace uso de ella, ya que todavía no está muy extendido.
  En este módulo, se muestra en una primera vista la clasificación obtenida en ToS;DR y con el color verde de fondo si cumple, amarillo si cumple parcialmente y rojo si no obtiene la evaluación mínima según las preferencias establecidas por el usuario. Si el usuario desea mayor información, se muestra un enlace a la información obtenida en ToS;DR como la figura 4. También, se muestra una tabla con el número de coincidencias de cada categoría según el modelo de *machine learning* de *Usable Privacy*, así como un enlace a la fuente y los iconos de tratamiento de datos de *Consent Commons*.
- Sistema de reputación: Para realizar esta validación se ha implementado un ecosistema de 2 RP de ejemplo con la capacidad de responder peticiones de valoración de reputación de un servicio sobre el que se pregunte. Esta valoración consiste en puntuación, que toma valores entre 0 y 100, según el nivel de reputación y la justificación de esta evaluación, en formato de campo de texto libre. Adicionalmente, en el Módulo Principal se ha asociado a cada RP un peso distinto según la confianza que tiene sobre éstos. Con ambas respuestas de las RP y el peso se calcula la valoración total según la fórmula (1). Una vez devuelto el valor al usuario, el módulo principal lo compara con el valor mínimo de reputación que ha elegido el usuario. Si es superior, se muestra OK con el fondo verde y en caso contrario NO OK con fondo rojo. Si se desea un mayor detalle, se pueden ver las justificaciones de cada RP pulsando sobre la evaluación.
- Revisiones centradas en la privacidad: A modo de ejemplo, y de nuevo con el objetivo de validar el diseño propuesto, este primer prototipo se centra en comprobar

dos de las buenas prácticas que se incorporan en la especificación de OpenID Connect. De momento el PAdv revisa que no se envíe la cabecera *Referer*, vulnerabilidad detectada en [1], ya que si al cargar la página de la RP, en la devolución de información personal al usuario, ésta contiene enlaces a terceros, la información personal se enviará a éstos. En la especificación de OpenID Connect, se indica que las respuestas que contengan información sensible deberán tener la cabecera HTTP *Cache-control* con valor *no store* y la cabecera *Pragma* con valor *no-cache*. Esta información se ha obtenido de forma manual mediante escaneo de tráfico. Se revisa también que la negociación del cifrado de la comunicación es segura. Esta información se ha obtenido con pysslscan [26]. Además, tal y como se ha comentado anteriormente, se debe comprobar que el certificado ha sido emitido por una autoridad de certificación de confianza, como las registradas en eIDAS, haciendo uso del API [27]. Con todas estas evaluaciones se emite una valoración donde cada cumplimiento es un punto. Al usuario se motrará OK si supera el mínimo de puntos según sus preferencias o NO OK en caso contrario. En el detalle, tiene disponible la evaluación de cada requisito.

Además, se ha realizado una primera validación con un total de 151 participantes con el objetivo de poder probar la eficacia del Privacy Advisor. Tras utilizarlo, se les ha pedido que calificaran la utilidad de cada uno de los módulos, así como el Privacy Advisor como agente que ayude al usuario a conocer cómo trata su privacidad durante el proceso de autenticación. La calificación está basada en el uso de estrellas, valorando del del 1 al 5, donde el 5 es la mejor y 1 la peor. El Privacy Advisor ha obtenido un total de 4,28 sobre 5. La calificación de los módulos ha sido para revisión de patrones de diseño 4,31 sobre 5 , Certificaciones de privacidad 3,77 sobre 5, Revisión de políticas de uso y privacidad 4,01 sobre 5, Sistema de reputación 3,82 sobre 5 y Cumplimiento de consideraciones seguridad y privacidad 3,99 sobre 5. Por último se ha valorado, utilizando el método explicado anteriormente, la utilidad de los códigos de color para una evaluación rápida del resultado obtenido por el Privacy Advisor, obteniendo un 4,28 sobre 5.

VI. Conclusiones y trabajo futuro

Una solución para que los usuarios finales confíen más en las federaciones para la gestión de identidades y en el denominado login social es la introducción, en estas federaciones, de agentes que ejerzan la función de sistemas de recomendación y que permitan a los usuarios tomar decisiones informadas acerca de la protección de su privacidad.

En este trabajo se ha propuesto el diseño del *Privacy Advisor*, un agente modular cuyas principales funciones son la revisión de patrones de diseño, la comprobación de certificaciones de privacidad, la revisión de políticas de uso y privacidad, la utilización de sistemas de reputación y el análisis de cumplimiento de mejores prácticas de seguridad y privacidad. La implementación de un primer prototipo de este agente ha permitido validar su utilidad desde el punto de vista de la facilidad de uso, de integración con un esquema federado muy extendido (OpenID Connect) y de escalabilidad, así como su viabilidad.

La línea más interesante de trabajo futuro es garantizar que los usuarios puedan hacer uso del PAdv manteniendo su identidad anónima con técnicas basadas en *Zero Knowledge Proof* o en *Group Signature*.

Referencias

[1] Manuel Urueña, Alfonso Muñoz, and David Larrabeiti, "Analysis of privacy vulnerabilities in single sign-on mechanisms for multimedia websites," *Multimed Tools Appl (2014) 68:159–176*, Jul. 2012.

[2] Jorge Navas and Marta Beltrán, "Understanding and mitigating OpenID Connect threats," *Computers & Security*, vol. 84, pp. 1–16, Jul. 2019.

[3] R. Weingärtner and W. Carla Merkle, "A design towards personally identifiable information control and awareness in OpenID Connect identity providers," *2017 IEEE International Conference on Computer and Information Technology*, 2017.

[4] Martin Schanzenbach and Georg Bramm, "reclaimID: Secure, Self-Sovereign Identities using Name Systems and Attribute-Based Encryption," *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering*, 2018.

[5] David Núñez and Isaac Agudo, "BlindIdM: A privacy-preserving approach for identity management as a service," *Springer-Verlag Berlin Heidelberg 2014*, Mar. 2014.

[6] "RFC 6749: The OAuth 2.0 Authorization Framework," https://tools.ietf.org/html/rfc6749, Oct. 2012.

[7] "OpenID Connect Core 1.0 incorporating errata set 1," https://openid.net/specs/openid-connect-core-1_0.txt, Nov. 2014.

[8] "RFC 7519: JSON Web Token (JWT)," https://tools.ietf.org/html/rfc7519, May 2015.

[9] Bernd Zwattendorfer, Daniel Slamanig, Klaus Stranacher, and Felix Hörandner, "A Federated Cloud Identity Broker-Model for Enhanced Privacy via Proxy Re-Encryption," *FIP International Federation for Information Processing 2014*, 2014.

[10] D. Nuñez, I. Agudo, and J. Lopez, "Integrating OpenID with proxy re-encryption to enhance privacy in cloud-based identity services," *2012 IEEE 4th International Conference on Cloud Computing Technology and Science*, 2012.

[11] J. Werner and C. Merkle Westphall, "A Model for Identity Management with Privacy in the Cloud," *Post-graduate Program in Computer Science Federal University of Santa Catarina P.O. Box 476, 88040-970, Florianópolis, SC, Brazil*, 2016.

[12] "Dark Patterns," https://www.darkpatterns.org/.

[13] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs, "The Dark (Patterns) Side of UX Design," *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing*, Apr. 2018.

[14] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher, "Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns," *Proceedings on Privacy Enhancing Technologies 2016, 4 (2016), 237–254*, Jul. 2016.

[15] "European Privacy Seal (EurPriSe)," https://www.european-privacy-seal.eu, Accedido May. 2021.

[16] "TrustArc," https://www.trustarc.com, Accedido May. 2021.

[17] "ePrivacy," https://www.eprivacy.eu/, Accedido May. 2021.

[18] J. A. Obar and A. Oeldorf-Hirsch, "The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services," *Information, Communication & Society*, vol. 23, no. 1, pp. 128–147, 2020, publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2018.1486870.

[19] "Platform for Privacy Preferences (P3P) Project," https://www.w3.org/P3P/, Accedido May. 2021.

[20] Rosa Sánchez, Florina Almenares, Patricia Arias, Daniel Díaz-Sánchez, and Andrés Marín, "Enhancing Privacy and Dynamic Federation in IdM for Consumer Cloud Computing," *Consumer Electronics, IEEE Transactions on*, vol. 58, pp. 95 –103, Mar. 2012.

[21] "Beautiful Soup - Python Library," https://www.crummy.com/software/BeautifulSoup/, Accedido May. 2021.

[22] "JQuery," https://jquery.com/, Accedido May. 2021.

[23] "Terms of Service; Didn't Read - ToS;DR," https://tosdr.org, Accedido May. 2021.

[24] "Usable Privacy," https://www.usableprivacy.org, Accedido May. 2021.

[25] "Consent Commons," https://consentcommons.com, Accedido May. 2021.

[26] "pysslscan - Python Library," https://pysslscan.readthedocs.io/en/latest/, Accedido May. 2021.

[27] "Trusted List Browser," https://webgate.ec.europa.eu/tl-browser/swagger-ui.html, Accedido May. 2021.

http://doi.org/10.18239/jornadas_2021.34.60

Una revisión de: Darknets: Interconexiones y dark markets

Carlos Cilleruelo*,Luis de-Marcos†, Javier Junquera-Sánchez‡, y José-Javier Martínez-Herráiz§
Cátedra de Investigación ISDEFE-UAH sobre Ciberseguridad, TIC y Avance Digital. Universidad de Alcalá
Edificio Politécnico. Campus Universitario, Ctra. Madrid-Barcelona km, 33, 600 Alcalá de Henares. Madrid. Spain
ORCID: *0000-0001-7107-8655, †0000-0003-0718-8774, ‡0000-0002-4597-6539, §0000-0002-2351-7163

***Resumen*—Dos de las darknets más populares son Tor e i2p. Ambas redes se han convertido en una vía para cometer actos delictivos. Esto evidencia la necesidad de estudiarlas de cara a poder identificar y luchar contra estas actividades. En este trabajo se analiza las conexiones que existen entre dos de estas darknets, Tor e i2p, y las implicaciones para los dark markets. Para ello se ha creado el primer dataset que combina datos de las dos redes. Este dataset se compone de un listado de más de 49 mil servicios alojados en darknets. Además, durante la creación de este dataset se descubrió que existían referencias y conexiones entre ambas darknets y sus servicios. Esto significa que no es posible explorar de forma completa una darknet, sin tener en cuenta otras redes similares. Las darknets funcionan como un ecosistema y existen vías y enlaces claros entre ellas.**

***Index Terms*—Tor, i2p, Darknet, Análisis de Grafos, Dataset**

Tipo de contribución: Resumen extendido del artículo, 'Interconnection between darknets,' in IEEE Internet Computing, doi: 10.1109/MIC.2020.3037723. Article in press.

I. Introducción

En los últimos años, la exploración de las darknets, redes ocultas, se ha convertido en algo de gran importancia para los gobiernos y Fuerzas y Cuerpos de Seguridad del Estado [1]. Se conoce como darknet al espacio de Internet que busca ocultar y proveer de anonimato, a través del uso de técnicas criptográficas y el uso de diferentes tecnologías de enrutamiento [2].

"The Onion Router"[1] (Tor) es la darknet más popular. Tor ha sido auditado e investigado [3] y tiene más de 2 millones de usuarios diarios y 6000 nodos o relays [2].Los nodos aseguran el funcionamiento de la red y proporcionan privacidad. El número de nodos existentes es un indicativo del apoyo y existencia de una comunidad activa en torno a Tor. Una de las características más importantes de Tor es la posibilidad de crear sitios web y servicios llamados servicios ocultos, hidden services. Estos hidden services solo se pueden acceder si estamos conectados a la red Tor. Originariamente la idea de estos servicios era evitar la censura y facilitar la libertad de expresión. Sin embargo se han convertido en un espacio para actividades delictivas, como el tráfico de drogas [4].

Otra darknet bastante popular y activa es 'Invisible Internet Project', i2p. Al igual que los servicios ocultos de Tor, i2p ofrece eepsites, que son los sitios web y servicios disponibles en esta darknet. No hay estadísticas públicas sobre usuarios y servicios diarios, pero i2p recibe actualizaciones y desarrollos periódicos. Al igual que ocurre con Tor, los eepsites se convirtieron en un lugar para actividades ilegales [5].

[1]https://www.Torproject.org/
[2]Tor Metrics: https://metrics.torproject.org/

El objetivo principal de este artículo es analizar la interconexión entre las darknets i2p y Tor, demostrando la necesidad de rastrear ambas redes para obtener y estudiar sus estructuras y servicios. Un resumen de las contribuciones de este artículo son:

- Creación y publicación de un dataset que combina dominios de i2p y Tor y sus conexiones. Este es el primer conjunto dataset con información combinada de la red Tor e i2p. Este dataset se encuentra públicamente accesible en GitLab, https://gitlab.com/ciberseg-uah/interconection-between-darknets-dataset
- Se ha realizado una clasificación de los dominios más relevantes en Tor e i2p, en términos de su posición e influencia, basándonos en los datos recopilados. Esto demuestra que cada darknet desempeña un papel diferente ofreciendo servicios específicos, y enfatiza la necesidad de rastrear y estudiar diferentes darknets. El estudio de una darknet servirá para ampliar y encontrar mas información referentes a otras redes del mismo tipo.

II. Dataset de dominios de Tor e i2p

De cara a estudiar la conexión de Tor e i2p se ha construido un dataset de dominios y relaciones. Este dataset, al tener relaciones, nos permite la construcción de grafos. Ha de tenerse en cuenta que la creación del dataset ha supuesto un proceso de dos años y la combinación de múltiples métodos de recolección de información, mayoritariamente distintas técnicas de crawling. Este dataset contiene 49249 dominios (2687 dominios de i2p y 46562 dominios de Tor) y un total de 304673 relaciones entre dominios.

Existen otros datasets relacionados con darknets como Ahmia dataset[3] o DUTA-10k [6]. Sin embargo, ninguna de estos datasets contiene relaciones entre redes. Además si comparamos el tamaño de nuestro dataset con los existentes podemos apreciar que DUTA-10K tiene 10 mil dominios onion, hidden services, frente a los 46 mil hidden services de nuestro dataset. Asimismo se debe tener en cuanta que estas cifras no incluyen el hecho de que nuestro conjunto de datos contiene también dominios de i2p. En el caso de los datos referentes a i2p no se puede comparar nuestro dataset con otros, ya que actualmente no hay datasets públicos o estadísticas de esta red.

III. Metodología

III-A. Construcción de grafos

Teniendo un dataset con relaciones entre dominios, que dominio apunta a otro, es posible construir un grafo dirigido. Los

[3]https://ahmia.fi/

Tabla I
MÉTRICAS DE LAS REDES (GRAFOS), NODOS DE I2P, NODOS DE TORY NODOS DE TOR E I2P COMBINADOS

Métrica	i2p (eepsites)	Tor (hidden services)	i2p + Tor (eepsites + hidden services)
Nodos	2687	46562	49249
Aristas	13857	282270	304673
Grado medio	5.517	6.062	6.186
Densidad	0.002	<0.001	<0.001
Longitud media del camino	2.769	4.356	4.412
Diámetro	8	11	12
Componentes conectados	11	616	328

nodos representan dominios, y las aristas representan enlaces entre dominios. Siendo cada nodo del grafo dominios de i2p y Tor. Ya que nuestro interés es estudiar la interconexión entre darknets se han excluido los enlaces a la clearnet, la red de internet abierta. Además, el grafo creado no incluye los enlaces duplicados (mismo origen y mismo destino). El estudio se centra en las conexiones que existen entre diferentes nodos y sus respectivas darknets.

IV. RESULTADOS

La *Tabla I* presenta métricas de i2p y Tor, y las métricas de los datos combinados. Podemos ver que el tamaño de las redes difiere significativamente. Aunque Tor tiene muchos más nodos y aristas, las diferencias para el grado medio, la longitud media del camino y el diámetro no son muy grandes.

Otro de los análisis realizados ha sido utilizando el algoritmo PageRank. Este algoritmo permite medir la importancia de un nodo en función del número y la calidad de los enlaces hacia él. PageRank devuelve la probabilidad de que un usuario que haga clic al azar en los enlaces llegue a un sitio. La *Tabla II* presenta los nodos con mayor PageRank. Los dominios pertenecientes a Tor ocupan las primeras posiciones. El dominio i2p mejor posicionado está en la posición diez. Los resultados muestran que los dominios Tor son más importantes que los dominios i2p. Los dominios clasificados en las posición dos (DarkNet Light), tres (FreshOnions) y cuatro (un listado de dominios onion) también están bien posicionados utilizando métricas como in-degree, closeness y betweenness. Sin embargo el dominio más importante es Daniel's Hosting, que también tiene un alto grado de entrada (rango 17), grado de salida (rango 5) y betweenness (rango 3). Los resultados sugieren que este sitio es el dominio más importante y que probablemente recibirá más visitas que cualquier otro. Por otro lado DarkNet Light es la lista más importante de servicios de onion en Tor, y probablemente recibirá muchas más visitas que otros recopilatorios de enlaces que le siguen en el ranking. En el caso de i2p identiguy.i2p es el dominio i2p mejor clasificado (puesto 10). Siendo un servicio de salto, jump service, similar al sitio de salto stats.i2p.

V. DARK MARKETS Y DATOS MÉDICOS

El análisis de los datos recopilados de distintas darknets facilita la identificación de distintos tipos de actividades ilegales presentes en darknets. Muchas filtraciones de datos hacen uso de darknets para publicar y vender anónimamente información robada. En nuestro análisis no encontramos un único mercado o web ilícita altamente popular o altamente referenciado en la red. No existe un solo mercado principal, dark market, en estas redes, sino que hay multitud de ellos.

Tabla II
NODOS CON EL MAYOR PAGERANK

Rango	Dominio	Red	PageRank
1	dhosting4xxoydyaiv...syd.onion	Tor	3492
2	pejjyyh7rhv5ctyu.onion	Tor	2897
3	zlal32teyptf4tvi...syd.onion	Tor	1478
4	onionsnjajzkhm5g.onion	Tor	1129
5	donionsixbjtiohce2...ead.onion	Tor	1077
...			
10	identiguy.i2p	i2p	778

El servicio mayoritariamente ofrecido en estos dark markets es la venta de droga, pero es necesario destacar la venta de medicamentos y datos médicos, como números de las seguridad social.

VI. CONCLUSIONES

Mediante el análisis de grafos, es posible demostrar que de cara a obtener más información de una darknet es necesario investigar otras redes. Además, este estudio ayuda a comprender el posicionamiento de los principales sitios en Tor e i2p y su influencia en la red. Nuestro análisis devuelve valores similares para la mayoría de las medidas, lo que sugiere que ambas darknets son estructuralmente similares. La reducción del número de componentes conectados, cuando todos los nodos se incluyen en un único grafo, sugiere que los nodos i2p desempeñan un papel importante para hacer que más nodos Tor sean accesibles.

Por último, los dominios Tor más importantes son los sitios web de alojamiento e índice de sitios. Y los dominios i2p más importantes son sitios de salto. Dado que Tor no ofrece motores de búsqueda efectivos, los sitios web que ofrecen listados de dominios resuelven parcialmente esta limitación.

AGRADECIMIENTOS

Este trabajo ha recibido financiación del programa de investigación e innovación Horizonte 2020 de la Unión Europea bajo el acuerdo de subvención nº 826284 (ProTego)

REFERENCIAS

[1] M. Chertoff and T. Simon, "The Impact of the Dark Web on Internet Governance and Cyber Security," Feb. 2015. [Online]. Available: https://www.cigionline.org/publications/impact-dark-web-internet-governance-and-cyber-security

[2] G. Owen and N. Savage, "The Tor Dark Net," Centre for International Governance Innovation and the Royal Institute of International Affairs, Sep. 2015. [Online]. Available: https://www.cigionline.org/publications/tor-dark-net

[3] K. Loesing, S. J. Murdoch, and R. Dingledine, "A case study on measuring statistical data in the tor anonymity network," in *International Conference on Financial Cryptography and Data Security*. Springer, 2010, pp. 203–215.

[4] D. S. Dolliver, "Evaluating drug trafficking on the tor network: Silk road 2, the sequel," *International Journal of Drug Policy*, vol. 26, no. 11, pp. 1113–1123, 2015.

[5] M. Wilson and B. Bazli, "Forensic analysis of i2p activities," in *2016 22nd International Conference on Automation and Computing (ICAC)*. IEEE, 2016, pp. 529–534.

[6] M. W. Al-Nabki, E. Fidalgo, E. Alegre, and L. Fernández-Robles, "Torank: Identifying the most influential suspicious domains in the tor network," *Expert Systems with Applications*, vol. 123, p. 212 – 226, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417419300296

Sesión de Investigación B5: Privacidad
http://doi.org/10.18239/jornadas_2021.34.61

Diversec Voting - Votación remota distribuida para una seguridad en profundidad

Marino Tejedor-Romero, David Orden, Ivan Marsa-Maestre, Javier Junquera-Sánchez
Universidad de Alcalá, Escuela Politécnica
E-mail: marino.tejedor@edu.uah.es, david.orden@uah.es, ivan.marsa@uah.es, javier.junquera@uah.es
ORCID: 0000-0002-1679-0161, 0000-0001-5403-8467, 0000-0002-5529-2851, 0000-0002-4597-6539

***Resumen*—En las últimas décadas se están proponiendo muchos sistemas de e-voting, que están suscitando un gran interés. No todos los problemas están resueltos todavía, especialmente si hablamos de sistemas remotos. En este trabajo se propone Diversec, un sistema de e-voting remoto y distribuido que utiliza el esquema de secreto compartido de Shamir, operaciones en el cuerpo de Galois y mixnets, y que ofrece verificación del voto de extremo a extremo. Los partidos participan como nodos en la red, protegiendo los intereses de su opción y garantizando la seguridad del proceso, al tener intereses enfrentados. El modelo de amenazas es lo más conservador posible y ni los actores más privilegiados tienen la capacidad de vulnerar la privacidad o la integridad de los votos. Se hace especial énfasis en la seguridad en profundidad, superponiendo diferentes mecanismos para ofrecer garantías incluso en condiciones más adversas de las esperadas.**

***Index Terms*—E-voting, Votación, Votación remota, Votación verificable, Distribuido, Secreto compartido, Shamir, Cuerpo Finito, Galois, Interpolación polinómica**

Tipo de contribución: Investigación original

I. Introducción

En las últimas décadas, el campo de estudio de la votación electrónica o *e-voting*, indispensable para el desarrollo de un concepto más abstracto conocido como *e-democracy*, ha suscitado un amplio interés. Por una parte, las votaciones son un pilar clave para el correcto funcionamiento de los sistemas de gobierno democrático-representativos vigentes en la actualidad. Por otra, el desarrollo digital tan explosivo que hemos sufrido recientemente hace posible que podamos plantearnos enfoques alternativos a los procesos electorales utilizados hasta ahora [1]. En este momento contamos con hardware muy potente a precios aceptables y con un sector de desarrollo en un estadio muy avanzado de la profesión. Todo ello abre la posibilidad de integrar la automatización al concepto de votación. Ambos son requisitos fundamentales, pero no suficientes. Dado lo crítico de los aspectos de la sociedad humana gestionados por medio de votaciones, cualquier sistema en este ámbito que quiera tener éxito debe proporcionar garantías y mecanismos de seguridad. Este ha sido el esfuerzo de muchos investigadores durante los últimos años, y es la principal motivación de este trabajo.

Comparativamente, los sistemas de *e-voting* presentan ventajas inmediatas frente al proceso manual. Permiten ahorrar gastos, en la medida en que en un proceso manual hay muchas personas y muchos recursos involucrados en proporcionar las garantías que dan legitimidad a los resultados. También permiten ahorrar tiempo, procedente de los procesos de preparación y de organización previa. En resumen, ir a votar resulta más barato y mucho más sencillo. Se podría esperar incluso que el proceso de votación resulte más cómodo para el usuario. Estos beneficios directos ya podrían aplicarse a las elecciones de cualquier estado que despliegue elecciones o referéndum, pero además nos hacen pensar en más posibilidades. Actualmente sería inviable intentar someter regularmente a votación física por la ciudadanía las decisiones que toma un parlamento, aunque fueran únicamente las más relevantes entre el conjunto de medidas. Así, el delegar el poder en representantes en un parlamento o un senado es una solución adecuada a una escasez de recursos, que podría replantearse si fuera viable someter un mayor número de cuestiones a la votación ciudadana. Con un sistema automatizado de *e-voting*, la reducción de costes y de tiempo necesario permitirían aumentar la cantidad de votaciones que se pueden organizar en un plazo fijo y con un presupuesto fijo. De forma secundaria, la democracia podría evolucionar gracias al *e-voting*, volviéndose más participativa para la ciudadanía, avanzando hacia un modelo más cercano a la democracia directa.

Existen numerosas publicaciones a este respecto, e incluso ha habido elecciones documentadas que han utilizado medios digitales para automatizar el proceso [1], [2]. Esto no quiere decir que se hayan adoptado ni sean aceptadas. En prácticamente todas estas pruebas ha habido, por lo menos, polémica. En el peor de los casos han sufrido hasta fraude electoral. Es por esta razón por lo que, además de la infraestructura digital que hace posible una votación automatizada, se precisan técnicas criptográficas para asegurar confidencialidad e integridad de cada voto, entre otras propiedades de la información. El diseño correcto de un sistema de votación electrónica, para que provea de garantías al votante y al resultado, es muy importante de cara a la confianza de los usuarios. Especialmente cuando la aceptación legitima los resultados obtenidos, y en un momento en el que la popularidad de estos sistemas electrónicos no es totalmente favorable.

No partimos de cero, pues ya existen numerosas propuestas de sistemas de *e-voting* que usan criptografía para solucionar los problemas mencionados anteriormente, aunque todavía se encuentran carencias, inconvenientes o posibles mejoras. En este marco se propone Diversec Voting, un nuevo sistema de *e-voting* con un enfoque alternativo, tanto a nivel técnico como en las premisas. Los pilares sobre los que se construye esta propuesta son el secreto compartido de Shamir [3] y las propiedades del cuerpo de Galois [4]. El sistema es de

tipo remoto (o REV, del inglés *Remote electronic voting*) y verificable de extremo a extremo (o E2EV, del inglés *End to end verifiable*). También es de tipo distribuido, dados los problemas de confianza que podría presentar un sistema centralizado. Pero además busca dar una solución criptográfica simple y en profundidad. Simple, porque así llegará a más público y será más fácil de comprender. En profundidad, porque es un proceso lo suficientemente crítico como para cubrir las propiedades básicas de un sistema de votación electrónica desde diferentes ángulos. El mayor aspecto diferencial respecto al resto de propuestas es el enfoque de seguridad basado en el modelo de actores y amenazas. Se busca ofrecer un nivel de garantías máximo sin condiciones o con las mínimas condiciones. En otras palabras, se busca ofrecer un sistema en el que todos los actores involucrados estén protegidos y seguros incluso bajo la suposición de que los actores privilegiados tienen intereses maliciosos. Se busca, a la vez, superponer los mecanismos de seguridad, según indica el principio de seguridad en profundidad, para cubrir posibles vulnerabilidades.

II. Estado del arte

II-A. Propiedades básicas

Las propiedades que se esperan de un sistema de *e-voting* han sido ya debidamente documentadas en la literatura [5]. Desde las propiedades más evidentes que ya se intuyen de los sistemas de votación establecidos de urnas y papeletas, como puede ser el anonimato del voto o la imposibilidad de emitir más de un voto por cada participante legítimo, hasta otros requisitos que son propios de los sistemas electrónicos. Vamos a repasar la lista de propiedades:

- **Voto secreto**: Solo el votante debe saber cuál es la opción votada. En un sistema de votación electrónica, este requisito implica que debemos salvaguardar la confidencialidad de la información respecto a otros votantes, respecto a atacantes, y además respecto a la propia entidad que organiza el proceso de votación.
- **Legitimidad**: Solo se permite el voto de participantes registrados previamente. Es decir, en el resultado final no debe haber ningún voto emitido por alguien no autorizado. En los sistemas electrónicos, el proceso de autenticación garantiza este requisito.
- **Elegibilidad**: Al igual que no se permiten votantes no legítimos, los votantes legítimos tienen derecho a emitir únicamente un voto.
- **Precisión**: El recuento final de votos refleja todos los votos legítimos, sin alteraciones desde que se emiten hasta que se publica el recuento. Esta propiedad se corresponde con otra dimensión básica de la seguridad de la información, la integridad.
- **Verificabilidad**:
  - **Individual**: El votante puede verificar el correcto funcionamiento del proceso electoral en el que ha participado. Se distinguen la verificación extremo a extremo, del voto en el mismo recuento, y la verificación del proceso, en la que se supervisan las diferentes fases del protocolo.
  - **De terceros**: El resultado de la votación es auditable por terceros independientes, que determinan con la información accesible si ha habido corrupción o no.

  Idealmente, la verificación individual o de terceros no debería violar el resto de las propiedades, no debería revelar votos ni empeorar las garantías del sistema.
- **Resistencia a la coacción**: Un votante es capaz de votar la opción que desee incluso bajo coacción, amenaza o chantaje de un tercero. Una forma popular de obtenerlo es la imposibilidad de demostrar el voto propio. De esta forma, un tercero no podría comprobar a ciencia cierta la decisión de la víctima. La compra de votos se evita de forma indirecta, diluyendo la seguridad del atacante de que el voto de la víctima es el deseado.
- **Resistencia a errores**: El sistema es capaz de desenvolverse correctamente, sin que el resultado de la votación se vea comprometido de ninguna manera, incluso cuando ciertos actores introducen corrupción. Los casos de error están previamente acotados y delimitados. Es un requisito importante, puesto que en un protocolo de *e-voting* participan muchos agentes. Debemos estar preparados para la entrada de corrupción en el sistema, sea intencional o no, desde el diseño.
- **Disponibilidad**: El sistema debe estar completamente accesible durante el proceso de votación.

Como se especifica en [5], estos requisitos no son estrictamente independientes. Por una parte, existen relaciones de dependencia, como pasa entre la verificabilidad de terceros, la verificabilidad individual y la precisión. Por otra, existen ciertas incompatibilidades. Por ejemplo, el voto secreto choca de cierta manera contra la posibilidad de verificar tu propio voto. Igual que la capacidad para la verificación individual frente a la resistencia a la coacción. Tener la información suficiente para trazar tu voto a través del sistema suele conllevar la capacidad de prestar esta información a un tercero y, en consecuencia, demostrar el voto. Por ello, esta relación es típicamente la más complicada de satisfacer.

El objetivo es lograr un compromiso entre todas estas características, adaptado a la situación actual, al modo de operar que tiene el público para votar, y a las ventajas que estimemos más valiosas de los protocolos de *e-voting* frente al sistema tradicional de voto con papel y urnas.

II-B. Propiedades adicionales

Además de las características ya clasificadas y reconocidas, se han expresado ciertas preocupaciones sobre los sistemas electrónicos a la hora de llevarlos a la práctica. Por ejemplo, sobre las garantías del sistema. Propuesto un protocolo de *e-voting* cualquiera, un votante tiene pocas formas de saber qué está pasando más allá de su interacción. En este sentido, es estrictamente necesario para la confianza del ciudadano que la formalización del protocolo sea abierta y pública (*open-source*). Sin embargo, el hecho de que el procedimiento y sus características sea de conocimiento público no garantiza que el resto de componentes del sistema con los que se interactúan se adhieran a este procedimiento [6]. Ésta es la base de muchos ataques en ciberseguridad. Los atacantes

crean clientes personalizados y/o *exploits* con los que se salen del protocolo habitual y aprovechan vulnerabilidades, lo que implica que resulte temerario un modelo de amenazas en el que sea necesario confiar en un comportamiento no malicioso por parte de las autoridades. Aquí es donde hace énfasis Diversec.

Bajo esta suposición, la única forma de garantizar la confianza y la seguridad del votante junto a la integridad del resultado se logra mediante el diseño. Un sistema de *e-voting* resistente a la corrupción de la propia organización tiene que cumplir los siguientes requisitos:

- El votante debe enviar únicamente la información mínima para que todo el procedimiento se complete correctamente. El factor clave es que, con esta cantidad de datos, una organización corrupta no debe poder comprometer la privacidad ni la integridad del voto.
- Una organización corrupta no puede falsear de forma convincente el resultado de una votación.

Estas condiciones confieren al votante la posibilidad de participar sin la necesidad de comprobar el software que ejecutan sus interlocutores, un concepto denominado "independencia del software" [5]. El énfasis en que el sistema sea robusto frente a los abusos de los administradores es, además, bastante poco frecuente entre los sistemas informáticos, dentro y fuera de la ciberseguridad. Los administradores suelen tener el control absoluto de lo que pasa en el sistema, salvo excepciones. Un ejemplo muy reciente de estas excepciones es el protocolo Signal de mensajería instantánea, que utiliza cifrado de extremo a extremo, aislando el canal de los administradores [7].

II-C. Protocolos y sistemas de e-voting hasta el momento

Las primeras propuestas que han tenido repercusión van en la línea de ThreeBallot [8], PunchCard [9] y Scantegrity [10]. En ellas se mezclan papeletas físicas y sensores para automatizar con seguridad unas elecciones.

Posteriormente encontramos protocolos remotos y verificables, más cercanos a nuestra propuesta. En lo referente a técnicas, encontramos tres grandes grupos de protocolos [5]:

- ***Mixnet***: Es una arquitectura en capas en la que se encadenan servidores *proxy*. En cada salto de la mixnet, se permuta el origen de los datos, con el objetivo de que después resulte imposible trazar un mensaje determinado. Es el principio detrás de redes como Tor.
- **Firma ciega**: Es una operación criptográfica en la que una entidad firma un mensaje sin revelar el contenido firmado.
- **Cifrado homomórfico**: Son técnicas de criptografía en las que se opera con la información cifrada, sin revelar su contenido. Suele implicar manejar datos de forma agregada

Además de las técnicas, vamos a revisar algunos de los sistemas relevantes en este momento.

- **Helios**: Helios [6] es una plataforma web de e-voting. Es un sistema centralizado que intenta proporcionar las dos propiedades más fundamentales, integridad y anonimato. Se prioriza una integridad absoluta por encima del voto secreto en el caso más desfavorable. Helios ignora el problema de la coacción de manera deliberada, argumentando que es un problema inherente al hecho de poder votar de forma remota. Los procesos de verificación son interactivos.
- **Civitas/JCJ**: Civitas/JCJ es otra aproximación al *e-voting* remoto, que sí que emplea medidas contra la coacción [11], [12], a través de la duplicidad de identidades. En una fase de registro, se provee una identificación válida, además de otras falsas, que después deben ser eliminadas del recuento final. Su modelo de amenazas es más permisivo. Bajo un análisis más crítico, estas suposiciones más relajadas están en contra de la propiedad comentada anteriormente: las garantías de los votantes.
- **Sistemas blockchain**: Lejos de reducir la complejidad, la tendencia más reciente busca propuestas de e-voting con blockchain [13]. Cada sistema tiene diferentes propiedades, pero suelen arrastrar las desventajas del modelo *Proof of Work*.

III. Modelo de actores y amenazas

En una propuesta de sistema en ciberseguridad es crítico especificar qué amenazas se esperan desde los actores del sistema y bajo qué suposiciones el sistema propuesto mantiene las propiedades que ofrece. Es especialmente importante en este caso por lo explicado previamente: las consecuencias de un ataque son lo suficientemente graves como para hacer indispensable un diseño basado en la seguridad en profundidad. El objetivo de Diversec es ofrecer un sistema seguro incluso bajo la suposición de que los actores más privilegiados intentan manipular las elecciones a su favor con todos los medios posibles.

En esta propuesta existen tres tipos de actores: administración, partidos y votantes. Los votantes son los que menos privilegios tienen, capaces de emitir su voto tras autenticarse y verificarlo. Los partidos son nodos del sistema distribuido, y van a manejar información de forma activa. La administración no participa de forma activa como los partidos, pero es la encargada de publicar y configurar todos los parámetros de la sesión de votación. Además, es esperable que la administración dependa de uno de los partidos.

Los votantes apenas tienen margen de actuación. La administración podría emitir información falsa, pero este ataque es fácilmente detectable por todos los que la reciben. Los partidos son los que tienen un mayor margen de ataque (como veremos más adelante) y un mayor interés. Sin embargo, gracias a las propiedades de las funciones *hash*, del secreto compartido de Shamir y de la *mixnet*, se logra que el sistema sea completamente seguro en confidencialidad y en integridad a la vez, con la única excepción de que todos y cada uno de los partidos sean corruptos y cómplices entre sí. Se considera que esta situación es bastante inverosímil dada la actualidad política, y que sería de una gravedad extrema que sobrepasaría los objetivos de esta propuesta. El comportamiento que no es tan inverosímil podría ser la creación de un partido cuyo único objetivo sea impedir la votación, bloqueando el proceso.

Como último apunte, hay que destacar que la propuesta revela necesariamente a los partidos atacantes. Al ser estos

entidades públicas, la protección es doble. Por una parte, el ataque es detectable y no va a tener éxito. Por otra, la demostración pública, criptográfica e inequívoca de que un partido determinado ha intentado manipular unas elecciones menoscabaría sustancialmente su reputación.

IV. Descripción del sistema

IV-A. *Inicialización*

En la primera fase del protocolo, la administración establece todos los parámetros del sistema y publica toda la información necesaria. Vamos a enumerar todos los actores que participan, los parámetros que se publican y su notación.

- **Partidos**: Participan P partidos, denotados por $p_i, 1 \leq i \leq P$, de los cuales se publican:
  - Direcciones de red.
  - Todos los pares de claves necesarios.
  - Su corte en el eje X, para el esquema de secreto compartido de Shamir. Por simplicidad, $x = i$ para cada p_i.
- **Opciones**: Cada uno de los elementos sobre los que se vota, representados por un entero desde 0 hasta $O - 1$. Incluyen el voto en blanco y cada uno de los partidos, además de otras opciones alternativas (no son candidaturas de un partido). O es necesariamente un número primo. Si la cantidad de opciones no fuera un número primo, entonces se selecciona el primo inmediatamente superior añadiendo votos en blanco como relleno. Por ejemplo, se podría publicar esta relación para $O = 7$:
  - 0: Voto en blanco
  - 1: Partido 1
  - 2: Partido 2
  - 3: Partido 3
  - 4: Partido 4
  - 5: Alternativa 5
  - 6: Voto en blanco
- **Votantes**: V ciudadanos listados en un censo público como $v_j, 1 \leq j \leq V$. Están identificados por una clave pública reconocida por la administración.
- **Orden de la *mixnet*** (M): Los partidos forman un anillo, mientras que el orden M es una permutación del conjunto de partidos que define su colocación en el anillo. Por simplicidad, asumimos que el orden es creciente. Denotamos por M_i al orden M rotado de forma que el último elemento sea i.
- **Reto de identificación** (R): Reto aleatorio para evitar ataques de repetición. Los votantes lo utilizarán en la siguiente fase para construir su identificador. Este reto es común a todos los votantes pero único para cada votación.

IV-B. *Generación del voto*

Trabajamos en el cuerpo finito definido por los elementos de $\mathbb{F}_O$, en módulo O. El votante genera un polinomio de grado $P-1$, $v(x) = a_0+a_1x+a_2x^2+\ldots+a_{(P-1)}x^{(P-1)}$. El término independiente a_0 representa la opción escogida por el votante, mientras que los demás coeficientes $\{a_1, a_2, \ldots, a_{(P-1)}\}$ son enteros aleatorios del cuerpo finito $\mathbb{F}_O$. Dado $v(x)$, se calculan todos los P puntos $v(i)$, uno para cada partido. En la Figura 1 se puede ver un ejemplo.

Por otra parte, el votante genera un identificador llamado ID concatenando el reto de identificación y una parte aleatoria lo suficientemente grande. Este identificador I_j es único para cada votante, y le va a permitir verificar su voto de forma anónima, por lo que debe recordarlo y mantenerlo privado.

IV-C. *Cifrado de los puntos*

Cada punto se procesa de la siguiente manera, según refleja la Figura 2:

1. Cada punto se compone de sus coordenadas X e Y, y del ID.
2. Se obtiene la secuencia de la *mixnet* M_i para empezar a iterar desde el último elemento (es decir, i) hasta el primero.
3. Para cada índice m el votante toma el estado actual del punto y lo cifra para el partido p_m con su clave pública.
4. En cada paso el votante guarda una huella *hash*, desde el texto en claro hasta que está cifrado P veces.

Tras este proceso, cada punto se ha cifrado utilizando las claves públicas de todos los partidos según el orden determinado por su M_i. El resultado se denomina C_i.

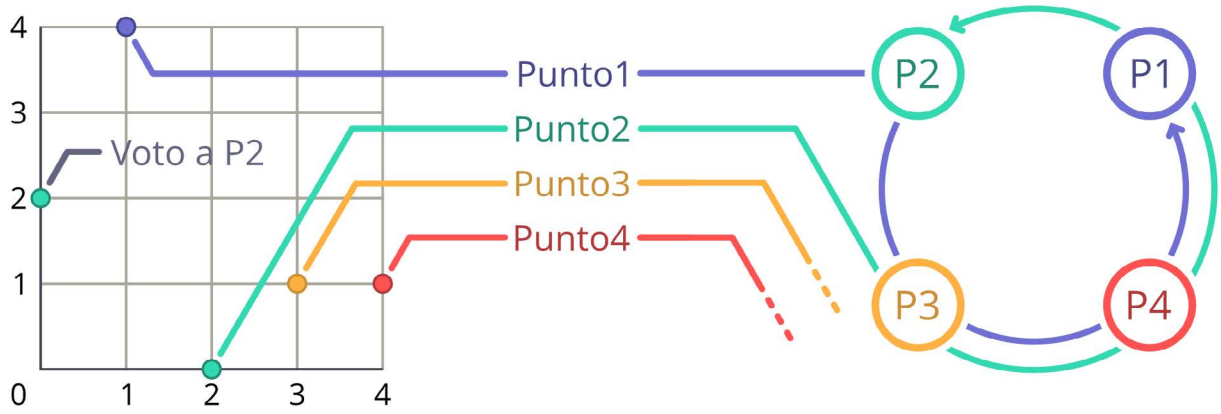


Figura 1: Generación del polinomio y distribución de los puntos

Como es habitual en los esquemas de criptografía asimétrica, el cifrado que se va a utilizar es, en la práctica, híbrido. Esto implica que el contenido se cifra utilizando cifrado de bloque o de flujo (simétricos), se cifra la clave utilizando criptografía asimétrica, y por último se mandan ambas partes.

Los P puntos de cada votante (uno por cada partido) han pasado por $P+1$ estados (desde el texto plano hasta haber sido cifrado P veces). Esto implica que cada votante va a almacenar la huella de $P^2 + P$ estados diferentes. Además, estas huellas *hash* no son simples. Según el sistema de cifrado mixto, cada mensaje se compone de criptograma y clave cifrada. El votante va a guardar huellas por separado de ambas partes. Se ilustra un ejemplo en la Figura 3.

El fin de estas huellas es verificar cada uno de los puntos durante la *mixnet*. Como se explicará más adelante, los partidos publican de forma agregada las huellas de todos los puntos que han descifrado, por lo que cada votante puede comprobar de forma anónima su punto. Se distinguen las huellas del criptograma y de la clave para poder recuperar en tiempo real un punto sin necesidad de revelar su contenido, en caso de ataque.

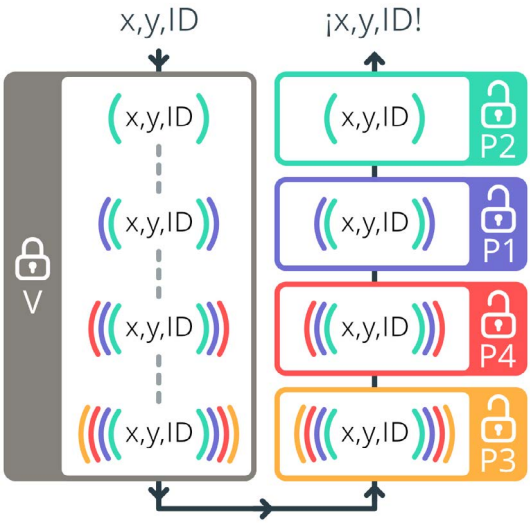


Figura 2: Ruta de cifrado (votante) y descifrado (partidos)

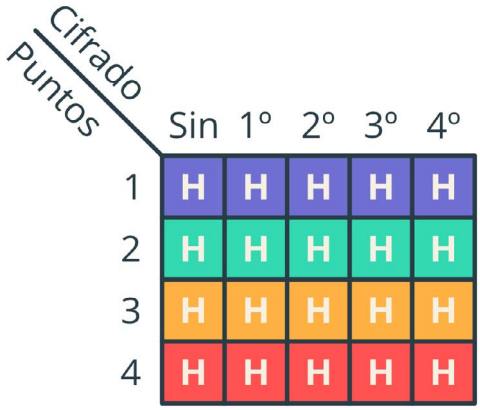


Figura 3: Tabla de huellas para cada votante

IV-D. Autenticación

Se utilizará un protocolo de autenticación de clave pública gracias al censo público, en el que consta toda la información necesaria de cada votante, de forma análoga al censo de una votación con urnas. Por ejemplo, valdría el método *public-key* del protocolo SSHv2, documentado en la RFC4252 [14], pero esta fase puede depender de los métodos considerados seguros en cada momento. Otra alternativa es utilizar un protocolo de transporte con opción de autenticación simétrica, como TLS 1.3 [15].

IV-E. Emisión

Una vez el votante queda autenticado, los partidos aceptan únicamente un punto cifrado de cada votante en esa misma conexión autenticada. De acuerdo con el esquema de cifrado explicado anteriormente, los puntos cifrados deben llegar al partido que le corresponde tras pasar por todos los nodos (en el orden cíclico establecido en la inicialización del sistema), de forma que solamente el último conozca el contenido. Por ello, cada punto cifrado C_i se envía al primer elemento de M_i.

IV-F. Mixnet

La fase de *mixnet* comienza cuando todos los votantes han emitido sus puntos. Llamamos V' al número de votantes legítimos que han decidido ejercer su derecho a voto, donde $V' \leq V$. Este subconjunto de participantes es conocido, dadas las firmas de autenticación, que comparten todos los partidos y la administración, y dados los puntos enviados. Los partidos publicarán las firmas de autenticación a modo de demostración, pero no los puntos. Este número debe coincidir a menos que algún votante se autentique sin emitir voto frente a todos o alguno de los partidos. Este caso se debe solucionar antes de continuar.

El objetivo de esta fase es que los puntos cifrados circulen por el anillo y lleguen a su destinatario, de acuerdo al orden M_i con el que el votante cifró cada punto. En cada salto, ocurren los siguientes pasos (se ilustra un ejemplo en la Figura 4):

1. El partido en cuestión recibe una lista de puntos cifrados del partido anterior en el anillo. Todos son puntos cifrados con una misma coordenada X, pero cada uno proviene de un votante.
2. Comprueba que lo recibido coincide con la lista de huellas *hash* publicadas por el partido anterior.
3. Descifra la capa más superficial de cada punto (que es la única que puede descifrar) recibido con su par de claves.
4. Comprueba que el contenido es otro criptograma para mandarlo al siguiente nodo. O que es el punto en texto claro, si nos encontramos en la última iteración.
5. Publica la lista de huellas de los puntos que ha descifrado, tanto de los criptogramas como de las claves cifradas. Es muy importante matizar que esa lista no revela de ninguna forma el orden de los puntos recibidos o de los enviados. Puede ser comprobada por el siguiente partido y por los votantes.
6. Si no es el final de la cadena, permuta de forma aleatoria el orden de los puntos cifrados y los reenvía al siguiente partido.

Después de P rondas, todas las capas de cifrado han sido eliminadas, y cada partido p_i tiene V' puntos en texto claro. Cada uno descifra la capa que puede, publica las huellas y transmite los puntos al siguiente partido permutados. Estas listas de huellas *hash* impiden que los partidos manipulen o sustituyan ninguno de los puntos que manejan. La Figura 5 ilustra un ejemplo.

IV-G. Publicación

Tras la última iteración de la *mixnet*, cada partido p_i conoce todos los puntos con x=i, junto a la coordenada y y el identificador que lo acompañan. Tras eliminar los posibles ataques de repetición (detectables gracias al reto de identificación) u otros tipos de ataque que impliquen datos mal formados, todos los partidos publican sus puntos.

Sin una medida de seguridad adicional, el último partido podría observar el resto de puntos y publicar una lista falsa que dé lugar al resultado que el partido atacante desee. Por esta razón, todos los partidos publican primero una huella que resuma todos los puntos. Solamente después de que todos los partidos hayan publicado su *hash* se publica la lista de puntos en claro. De esta manera se 'ata' la publicación de cada lista sin tener el conocimiento de las demás.

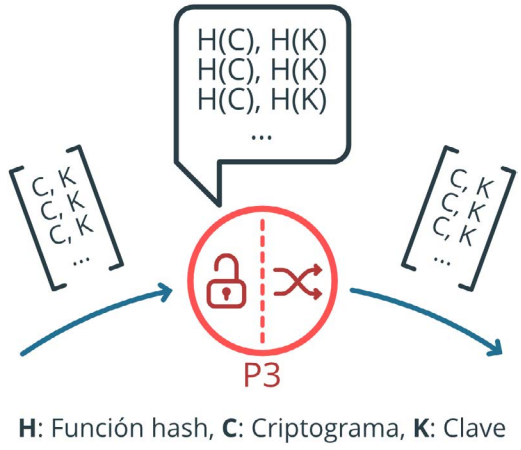


Figura 4: Fase de *mixnet*

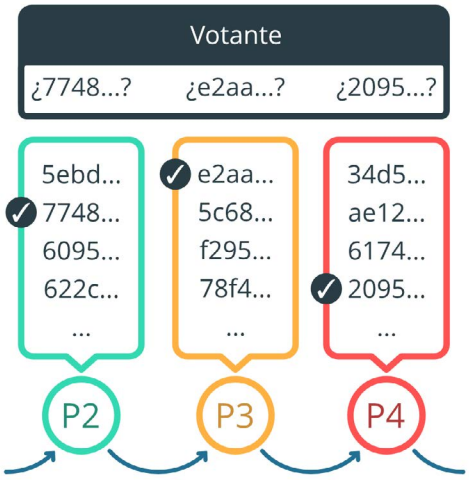


Figura 5: Verificación de huellas

A partir de este punto, todos los participantes de esta votación e incluso espectadores externos pueden recomponer el resultado a partir de los puntos. Para ello, deben:

1. Alinear todas las listas ordenando por los identificadores. Se ignoran los puntos que no pertenezcan a una misma función polinómica, es decir, puntos cuyo identificador no aparezca una única vez por lista. De lo contrario, un votante podría intentar un ataque de malformación de datos enviando sus cuatro puntos con diferentes ID.
2. Para cada identificador, se extraen los P puntos que definen la función polinómica que en su momento generó su votante.
3. Se interpola dicho polinomio y se extrae el término independiente, donde el votante codificó su voto.
4. El recuento es la lista de todos los identificadores junto al termino independiente extraído.

Esta lista agregada permite saber la cantidad de votos que tiene cada entero de la lista de opciones publicada en la inicialización. Los IDs de este recuento no aportan información sobre el origen del voto, salvo al propio votante, que puede verificar su voto y sus puntos de forma confidencial sobre este recuento. Se puede ver un ejemplo en la Figura 6.

V. Resultados

V-A. Análisis de seguridad

El sistema de e-voting Diversec cumple con todas las propiedades enunciadas anteriormente en la Sección II, contemplando ataques desde la administración, desde los partidos y desde los votantes.

El voto es secreto para el resto de votantes, para partidos y para la administración. Los puntos se cifran para todos los partidos, que solamente llegan a conocer una coordenada y un identificador del que no pueden trazar su origen. Los puntos se mantienen íntegros, además, por el 'rastro' de huellas que deja cada punto. La verificación se puede hacer en tiempo real, por lo que es imposible manipular la información que manejan los partidos. El identificador que se publica al final de la votación habilita la verificación de extremo a extremo mientras que la autenticación con clave pública y la consistencia en el número de firmas y puntos que circulan por la *mixnet* hace imposible que alguien vote sin estar en el censo, o que un votante vote dos veces.

Los votantes tienen la capacidad de emitir su voto y de verificarlo posteriormente. Los partidos actúan dentro de la *mixnet*, y la administración publica la información necesaria. En base a estas capacidades, vamos a enumerar los posibles comportamientos maliciosos que se podrían dar y cómo se frustran.

- **Publicación de información errónea por parte de la administración.**
  Dado que todos los parámetros son públicos, no puede mentir deliberadamente sobre un dato (independientemente de lo que pretenda) puesto que todos los participantes o los afectados serían conscientes.
- **Construcción maliciosa de votos y/o puntos**.
  Los votantes pueden escoger las coordenadas X e Y, y el ID. La coordenada X está limitada por el partido receptor,

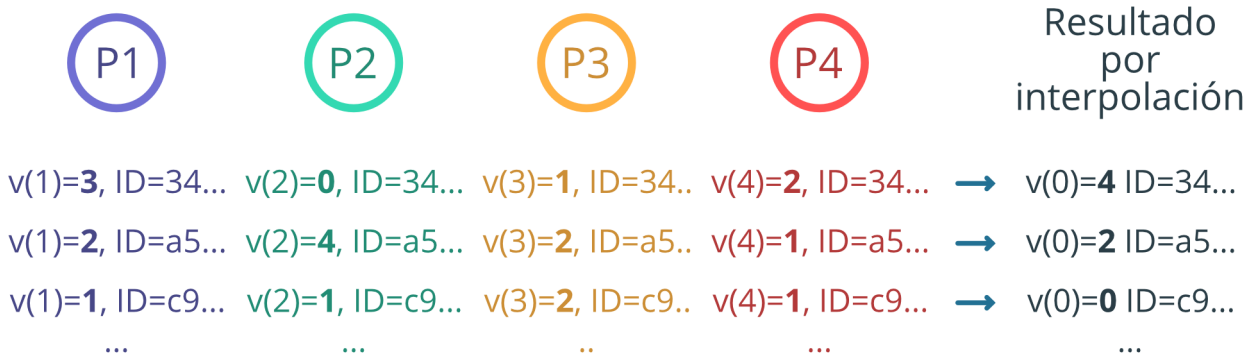


Figura 6: Publicación e interpolación del resultado

y la coordenada Y por el cuerpo finito. El ID es aleatorio. Si dicho votante intentara mandar cada punto con un ID diferente, sus participaciones serían eliminadas en el recuento final.

- **Manipulación de puntos durante la *mixnet*.** Este es el paso más crítico para la seguridad de los votos, pero está completamente cubierto. Cada partido es responsable de descifrar los puntos que recibe, publicar las huellas *hash* de estos puntos descifrados, y reenviar al siguiente nodo los puntos permutados. Los partidos podrían intentar manipular los puntos con el fin de cambiar el resultado final, pero eso resulta imposible. En primer lugar, los partidos no conocen el contenido del punto que descifran, a menos que sea la última etapa y eliminen la última capa. En segundo lugar, las propiedades del secreto compartido de Shamir hacen que el partido malicioso esté a ciegas. Alterar un punto sin el conocimiento del resto implica que el voto va a caer aleatoriamente en cualquier entero del espacio de opciones. Por último, las huellas publicadas hacen imposible que ninguno de los nodos sea capaz de alterar el contenido de los puntos sin que el votante que lo ha emitido ni los partidos contiguos lo detecten.
- **Denuncias de manipulación falsas**. El votante tiene el derecho de interrumpir la *mixnet* si ve que las huellas que tiene calculadas dejan de coincidir con la lista publicada por el partido en un momento dado. Para ello, debe indicar cuál es la huella en la lista del partido anterior que precede a la huella que falta. Sin embargo, el partido aludido puede demostrar fácilmente con la información que ha publicado y la que posee que ha descifrado correctamente el punto indicado y que se corresponde con una de las huellas que ha publicado si ha actuado correctamente. En esta demostración no se revela ningún tipo de información confidencial, porque especificamos que se separan las huellas del criptograma y de la clave cifrada del sistema mixto.

Dada la información que se publica, existen protocolos de resolución en tiempo real por los cuales todos los actores pueden interrumpir el sistema en caso de ataque, demostrarlo, resolver la incidencia sin vulnerar la privacidad ni la integridad del voto, y continuar.

Sin embargo, el análisis revela tres puntos vulnerables. Por una parte, la resistencia frente a la coacción puede vulnerarse si el votante quiere vender su identificador antes de que se revelen los resultados. De esta forma, este votante demuestra su resultado a un tercero. Es importante destacar que este ataque está restringido a la ventana de tiempo que transcurre desde el inicio de la votación hasta la publicación de los resultados. Después de este momento, con todos los identificadores publicados, ya no es posible. Por otra parte, la única manera de que el sistema de e-voting Diversec resulte inseguro es que todos los partidos sean corruptos y cómplices entre sí, puesto que la tensión entre intereses es uno de los mecanismos de seguridad. Esta situación límite no resulta problemática, pues es bastante improbable que todos los partidos del espectro ideológico y político sean conniventes entre sí. Si ocurriera, sería de una gravedad tan extrema que una votación carecería de sentido. La tercera condición conflictiva sería que se cree un nodo con la única intención de impedir que se desarrolle la votación. Se debe estudiar este caso más en profundidad, e intentar resolverlo.

V-B. Análisis computacional

La escalabilidad de un sistema de e-voting es crítica. La dimensión que va a crecer de manera sustancial es la cantidad de votantes V, mientras que el número de partidos P o el módulo del cuerpo de Galois O está limitado a cifras razonables. Respecto a V, la peor complejidad computacional que encontramos a lo largo de todas las fases es lineal, puesto que las operaciones son independientes (por ejemplo, descifrar cada punto o reconstruir para cada votante el voto a partir de los puntos de la función polinómica publicados). Por esta razón, el sistema tiene una buena escalabilidad. Respecto a los partidos, la complejidad es cuadrática. Este dato es todavía aceptable, porque aunque participaran todos los partidos en el sistema, este número no va a ser preocupante. Como ejemplo, en las últimas elecciones de la Comunidad Autónoma de Madrid participaron un total de 20 candidaturas.

Aunque la complejidad lineal sea favorable, si la situación lo requiriese (por simplicidad, logística, o incluso por restricciones de la ley electoral vigente en cada caso) siempre

existe la posibilidad de fragmentar la votación en diferentes subconjuntos.

VI. Conclusiones

Comparativamente, Diversec cumple con las propiedades esperadas, salvo la de resistencia a la coacción, que se ignora también en Helios [6] y muchos otros sistemas. Los pocos casos que incluyen mecanismos a este respecto no cumplen si los analizamos bajo un modelo de amenazas estricto.

Sin embargo, encontramos novedades y aportes destacables. A nivel técnico, el esquema de secreto compartido de Shamir es un recurso bastante menor y que no se había juntado antes con redes *mixnet*. Lo mismo ocurre con el uso del cuerpo de Galois para garantizar que tanto los puntos como el voto sean enteros válidos. Es decir, en la medida en la que utilizamos un cuerpo finito con la mínima cardinalidad posible para que todas las opciones estén representadas con un elemento de dicho cuerpo, evitamos que los atacantes introduzcan valores inválidos.

La arquitectura distribuida del sistema, que reparte la responsabilidad del tratamiento de la información entre entidades independientes y con intereses enfrentados, marca la diferencia respecto de otros sistemas de votación. Ya no existe un servidor central que concentra los datos y que puede resultar problemático para garantizar la seguridad de las elecciones frente a ataques desde dentro del propio servidor central. Incluso bajo la desconfianza total (que es el espíritu que se mantiene en todo este trabajo), la privacidad y la integridad quedan protegidas, a excepción de que todos los partidos sean corruptos entre sí, situación en la que ni siquiera tiene sentido dar legitimidad a unas elecciones. De esta forma, hemos llevado al mínimo posible la confianza que los votantes tienen que depositar en el sistema.

En resumen, Diversec es un sistema de e-voting sencillo pero seguro en profundidad. Los mecanismos de verificación son simples y claros, y los procesos se pueden entender de manera intuitiva. Esto es importante puesto que los votantes aceptarán más fácilmente un sistema que entiendan mínimamente antes que otro que resulte opaco o demasiado complejo para el público. No solamente es importante para el público no-experto, también es positivo para los técnicos y para los investigadores. La simplicidad de Diversec no repercute negativamente en sus propiedades, puesto que es una propuesta que cubre la privacidad y la integridad del voto, las dos dimensiones más críticas de la seguridad de la información en este caso, en profundidad. Dado lo crítico de las elecciones, en esta propuesta se solapan a la vez sistemas de seguridad que se encargan de dar garantías incluso cuando se de alguna vulnerabilidad no esperada.

Como trabajo futuro, Diversec queda abierto a posibles mejoras y modificaciones que impliquen la mejora del sistema. La prioridad en este sentido es intentar encontrar la solución al problema de la coacción en sistemas remotos, el cual es uno de los puntos que más dificultades presenta a los investigadores de ciberseguridad. Además, queda pendiente desarrollar un *software* que implemente el sistema aquí explicado. En un primer paso, se desarrollaría una versión simplificada con fines puramente demostrativos. Con la experiencia adquirida, en un segundo paso se desarrollaría el *software* final, orientado a votaciones reales, con todas las garantías de seguridad que ofrece.

VII. Agradecimientos

Marino Tejedor Romero ha sido financiado por una beca de iniciación a la investigación de la Universidad de Alcalá. David Orden ha sido financiado por el proyecto PID2019-104129GBI00 / AEI / 10.13039/501100011033 del Ministerio de Ciencia e Innovación. Ivan Marsa-Maestre ha sido financiado por el proyecto PID2019-104855RB-I00/AEI/10.13039/501100011033 del Ministerio de Ciencia e Innovación. David Orden e Iván Marsá han sido financiados por el proyecto SBPLY/19/180501/000171 de la Junta de Comunidades de Castilla-La Mancha y FEDER y por el proyecto UCeNet (CM/JIN/2019-031) de la Comunidad de Madrid y la Universidad de Alcalá.

Referencias

[1] J. P. Gibson, R. Krimmer, V. Teague, and J. Pomares, "A review of e-voting: The past, present and future," vol. 71, no. 7-8, pp. 279–286, 2016.

[2] Ü. Madise and T. Martens, *E-Voting in Estonia 2005. The First Practice of Country-Wide Binding Internet Voting in the World*. Gesellschaft für Informatik e.V., 2006.

[3] A. Shamir, "How to share a secret," vol. 22, no. 11, pp. 612–613, 1979.

[4] R. Lidl and H. Niederreiter, *Finite Fields*. Cambridge University Press, 1997.

[5] T. Peacock, P. Y. A. Ryan, S. Schneider, and Z. Xia, "Chapter 69 - Verifiable Voting Systems," in *Computer and Information Security Handbook (Second Edition)*, J. R. Vacca, Ed. Morgan Kaufmann, 2013, pp. 1103–1125.

[6] B. Adida, "Helios: Web-based Open-Audit Voting." in *USENIX Security Symposium*, vol. 17, 2008, pp. 335–348.

[7] K. Cohn-Gordon, C. Cremers, B. Dowling, L. Garratt, and D. Stebila, "A Formal Security Analysis of the Signal Messaging Protocol," vol. 33, no. 4, pp. 1914–1983, 2020.

[8] R. L. Rivest, "The ThreeBallot Voting System," 2006.

[9] S. Popoveniuc and B. Hosp, "An introduction to punchscan," in *IAVoSS Workshop on Trustworthy Elections (WOTE 2006)*. Robinson College United Kingdom, 2006, pp. 28–30.

[10] D. Chaum, A. Essex, R. Carback, J. Clark, S. Popoveniuc, A. Sherman, and P. Vora, "Scantegrity: End-to-End Voter-Verifiable Optical- Scan Voting," vol. 6, no. 3, pp. 40–46, 2008.

[11] S. Neumann, C. Feier, M. Volkamer, and R. Koenig, "Towards A Practical JCJ / Civitas Implementation," 2013, pp. 804–818.

[12] A. Juels, D. Catalano, and M. Jakobsson, "Coercion-resistant electronic elections," in *Towards Trustworthy Elections*. Springer, 2010, pp. 37–63.

[13] N. Kshetri and J. Voas, "Blockchain-enabled e-voting," *IEEE Software*, vol. 35, no. 4, pp. 95–99, 2018.

[14] T. Ylonen and C. Lonvick. (2020) The Secure Shell (SSH) Authentication Protocol. [Online]. Available: https://tools.ietf.org/html/rfc4252

[15] E. Rescorla. The Transport Layer Security (TLS) Protocol Version 1.3. [Online]. Available: https://tools.ietf.org/html/rfc8446

Sesión de Transferencia Tecnológica

Sesión de Transferencia Tecnológica
http://doi.org/10.18239/jornadas_2021.34.67

Aplicación de control de acceso y técnicas de Blockchain para el control de datos genéticos

Isabel Román
Dpto. Ingeniería Telemática
Universidad de Sevilla
Email:isabel@trajano.us.es
https://orcid.org/0000-0001-9508-0880

Germán Madinabeitia
Dpto. Ingeniería Telemática
Universidad de Sevilla
Email:german@trajano.us.es
https://orcid.org/0000-0001-6376-4620

Rafael Estepa
Dpto. Ingeniería Telemática
Universidad de Sevilla
Email:rafa@trajano.us.es
https://orcid.org/0000-0001-8505-1920

Jesús Díaz-Vedejo
Dpto. Teoría de Señal, Telemática y Comunicaciones
Universidad de Granada
E-mail: jedv@ugr.es
https://orcid.org/0000-0002-8424-9932

Antonio Estepa
Dpto. Ingeniería Telemática
Universidad de Sevilla
Email:aestepa@trajano.us.es
https://orcid.org/0000-0003-1841-3973

José Luis González-Sánchez
Fund. Computación y Tecnologías Avanzadas de Extremadura
E-mail: joseluis.gonzalez@cenits.es
https://orcid.org/0000-0002-0777-8427

Felipe Lemuz Prieto
Fund. Computación y Tecnologías Avanzadas de Extremadura
E-mail: felipe.lemus@cenits.es
https://orcid.org/0000-0001-7119-8630

***Resumen*—Este trabajo presenta una solución al reto de mejorar la trazabilidad del acceso a información genética almacenada en una aplicación propietaria a través del uso de blockchain. Para ello se realizan tres acciones: (a) se normaliza la estructura y acceso a los datos conforme al estándar sanitario FHIR; (b) se diseña una arquitectura normalizada de control de acceso a los datos en la que el paciente puede administrar las políticas de acceso a sus datos clínicos compatible con el RGDP; (c) se securiza mediante blockchain la trazabilidad del acceso a los datos.**

Los resultados de las tres acciones anteriores se integran en un demostrador o una aplicación piloto que tiene las siguientes características: (a) arquitectura SOA con interfaces normalizados de acceso que siguen el estándar FHIR; (b) cuenta con sistema distribuido de control de acceso de grano fino que sigue el estándard XACML/SAML; (c) utiliza blockchain de forma que se garantice la trazabilidad y la integridad de los registros de acceso al sistema.

***Index Terms*—Respuesta al reto 'Aplicar blockchain en el ámbito sanitario para la trazabilidad de información genética': Solución para Trazabilidad, Control de accesos, Blockchain**

Tipo de contribución: *Reto de Transferencia*

I. Introducción y objetivos

Uno de los retos científicos propuestos por la Fundación Computación y Tecnologías Avanzadas de Extremadura (COMPUTAEX) en las Jornadas Nacionales de Investigación en Ciberseguridad (JNIC) 2019/20 fué el titulado '*Aplicar blockchain en el ámbito sanitario para la trazabilidad de información genética*'. Este documento presenta los resultados que el grupo de investigación de Ingeniería Telemática de la Universidad de Sevilla ha logrado en la consecución de este reto.

La ficha del reto reza *El reto que se propone es analizar y diseñar un sistema basado en blockchain, ya sea público, privado o híbrido o incluso valiéndose de cadenas de bloques ya existentes (por ejemplo la red ethereum) que permita controlar y registrar de forma ineludible el acceso a información personal anonimizada especialmente sensible y garantice si se han producido cambios o no en dicha información. Recorriendo la cadena de bloques deberá ser posible revisar los accesos y si ha habido cambios garantizando de esta forma la confidencialidad e integridad de la información. En concreto, se propone el reto de analizar y diseñar un sistema blockchain de trazabilidad de información genética que garantice la seguridad, privacidad, autenticidad y anonimato de la información genética durante las etapas de secuenciación, almacenamiento, acceso, actualización y cancelación*.

Tras una reunión inicial con el retador se aclararon los términos del reto que, en términos generales, buscaba mejorar la seguridad en el proyecto HeritaGen desarrollado por COMPUTAEX y centrado en el dominio de la gestión de datos clínicos y genéticos. El foco de interés era controlar y registrar de forma ineludible el acceso a la información, especialmente sensible, y garantizar su integridad. Al mismo tiempo se buscaba diseñar una solución abierta, que facilitara la integración de la solución desarrollada en flujos de trabajo más complejos que implicaran la colaboración de diversas organizaciones.

Se desea que las soluciones de diseño incluyan los siguientes requisitos:

- Una interfaz normalizada de acceso a los datos. Esto facilita la reutilización de librerías y proyectos externos en el desarrollo de las aplicaciones que exploten los datos (tanto propias como de terceras partes).
- Facilitar la colaboración con organizaciones externas (secuenciadores de genoma externos, Servicio Extremeño de Salud, etc.) utilizando un formato común de acceso, una estructura de datos conocida y una trazabilidad basada en blockchain. En definitiva, estas dos ventajas

se podrían resumir como la facilidad para la integración en arquitecturas de servicios complejas.
- Facilitar la implementación de un sistema de control de acceso que cumpla con el Reglamento General de Protección de Datos (RGPD). Por un lado debe permitir una granularidad fina a la hora de definir y aplicar políticas de acceso a los datos en función de sus características. Por otro debe sentar las bases para capacitar al ciudadano en la autoadministración de las políticas de control de acceso a sus datos clínicos

A la luz de lo expuesto, el objetivo de este trabajo es presentar el diseño de una solución para la integración de servicios propietarios en una Arquitectura Orientada a Servicios (SOA) sanitarios estandarizada que incluya aspectos de seguridad en cuanto a la identificación y el control de acceso de usuarios. Este objetivo general se puede descomponer en tres objetivos parciales:

O1 Normalización del acceso a los datos conforme a un estándard sanitario
O2 Aplicación del control de acceso de grano fino centrado en el propietario de los datos
O3 Creación de un sistema de trazabilidad basado en blockchain para garantizar la integridad de registros de acceso y configuración.

II. Estado de la técnica

Los modelos para la provisión de soluciones software están en constante evolución. Sin embargo, en todos ellos aparece la exigencia de integración, dada la necesidad de reutilizar componentes (propios o de terceros). La integración puede resultar especialmente compleja cuando el servicio ofrecido da soporte a procesos empresariales que involucran a diversas organizaciones, cada una con distintas tecnologías, conforme a estándares y políticas organizativas diferentes. En este sentido, el dominio sanitario es un claro ejemplo en el que la integración resulta clave. No solo por la importancia del acceso a los datos de la Historia Clínica Electrónica (EHR, por sus siglas en inglés), que puede estar distribuida en diversos centros, sino porque es cada vez más frecuente que en un proceso asistencial participe más de una entidad. Las arquitecturas abiertas de servicios (SOA) son un referente para facilitar la integración de soluciones empresariales. Éstas permiten mantener un acoplamiento débil entre cada componente/servicio y facilitan la integración al proporcionar capacidades transversales y la posibilidad de incluir componentes que den soporte a flujos de negocio complejos. No resulta extraño, pues, que la mayoría de las soluciones sanitarias actuales sigan este paradigma. Aun así, la integración no es trivial y es necesario resolverla en distintos niveles o puntos de vista: empresarial/dominio, modelos información, interfaces computacionales, ingeniería/arquitectura y tecnología [1], [2], [3], [4].

Una de las claves de la integración es el uso de estándares. Existen dos estándares principales para la gestión de información sanitaria que definen las estructuras de datos que deben emplearse para el intercambio de información sanitaria (p.ej. observaciones, pruebas clínicas, medicamentos, citas, etc.) y otra información de interés, como la información demográfica (de pacientes, personal, organizaciones, etc.):

- ISO/CEN 13606 [5]: estándar usado en la comunidad europea basado en una simplificación el modelo OpenEHR con el objetivo de trasladar la información clínica a formato electrónico.
- HL7/CDA [6]: estándar creado por ANSI que goza de una amplia difusión en el dominio sanitario. Además de las estructuras de datos, define una interfaz de comunicación y acceso a recursos mediante la mensajería *Clinical Document Arquitecture* (CDA). Adicionalmente, HL7 también define otro estándar para el intercambio de datos sanitarios que sigue el paradigma REST llamado *Fast Healthcare Interoperability Resources* (FHIR). Así, FHIR proporciona una API RESTfull normalizada con la que realizar operaciones de tipo *create / read / update / delete* (CRUD) sobre recursos en el dominio sanitario. FHIR también proporciona un mecanismo para extender los recursos ya normalizados.

La mayor penetración de mercado del estándar HL7/CDA, así como la ventaja de tener definida una interfaz normalizada de acceso a los datos clínicos y la experiencia previa [7], [8], [9], [10] de algunos miembros del equipo de trabajo, sugieren la idoneidad de usar el estándar FHIR para el desarrollo del presente proyecto.

La normativa HL7 no presupone ninguna medida de seguridad ni control de acceso a los datos. Sin embargo, la seguridad es un aspecto relevante en cualquier SOA y por ello es habitual que las plataformas proporcionen capacidades para soportar algunas funcionalidades de seguridad como la gestión de identidad y control de accesos. Así, en lugar de reinventar soluciones propietarias para estos aspectos, los desarrolladores pueden interactuar con los componentes de la arquitectura que tienen estas responsabilidades. Para abordar el segundo de nuestros objetivos, en este trabajo se diseña la arquitectura de un sistema de control de acceso que incluye puntos de aplicación de políticas que gobernarán el acceso a la información. Además, el sistema diseñado debe tener en cuenta el cumplimiento de la legislación en relación a considerar quién es el propietario de los datos (RGDP). Por ello, la propuesta de control de acceso se debería basar en la premisa de que es el paciente el que administra las políticas de acceso a sus datos clínicos.

La mejor opción para diseñar la arquitectura de control de acceso es, de nuevo, recurrir a estándares bien reconocidos como el *framework* definido por la ITU-T en su norma X.810. En nuestro caso particular, es posible implementar los aspectos de *framework* relativos al control de acceso (norma X.812) en una arquitectura SOA utilizando el estándard SAML (Security Assertion Markup Language), que define un lenguaje de política de control de acceso basado en atributos de grano-fino. La combinación de SAML con XACML (*Extensible Access Control Markup Language*) permite diseñar una arquitectura distribuida compatible con X.812 que cuenta con los siguientes componentes principales:

- El punto de aplicación de políticas (PEP, *Policy Enforcement Point*), en el que los servicios delegan las tareas de control de acceso.
- El punto de administración de políticas (PAP, *Policy Administration Point*) que facilita la gestión de las políticas de control de acceso (edición, activación, desactivación,

combinación...). Como hemos señalado anteriormente, un aspecto de especial relevancia en nuestro caso de estudio es considerar quién es el responsable de la administración de políticas. Este trabajo se enfoca en propocionar una solución que permita al paciente (o ciudadano en general) gestionar las políticas que regirán el acceso a su información sanitaria y genética.
- El punto decisor de políticas (PDP, *Policiy Decision Point*), que discrimina las políticas que es necesario considerar y las evalúa, en función de los atributos aplicables, para tomar una decisión.
- El punto de Información de políticas (PIP, Policy Information Point), al que se puede recurrir para consultar los valores concretos de atributos aplicables en las reglas a evaluar.

En cuanto a las vistas más bajas, y dependientes de plataformas y tecnologías concretas, existen numerosas soluciones que pueden utilizarse como base para el despliegue de una SOA. Una de las principales implementaciones en software libre es la conocida como WSO2. Este proveedor integra soluciones de software libre y proporciona componentes interoperables que cubren distintas necesidades para desplegar una SOA.

Por último, este proyecto utilizará la tecnología BlockChain [11], [12] para proteger el cuaderno de bitácora (log) compartido entre diversas entidades y que no puede ser alterado. La aplicación de esta tecnología debe ser distribuida, por lo que sería posible realizar una implementación utilizando los centros de supercomputación a los que se tiene acceso. Existen numerosos proyectos para blockchain en código abierto, como *Hyperledger, Ethereum, Corda, Quorum, Openchain, etc.* que se serían aplicables a nuestro caso de uso.

III. Trabajo desarrollado

III-A. Situación de partida

El trabajo se debe iniciar a partir de una base de datos no relacional (MongoDB) preexistente donde se guardan colecciones de información en formato JSON. Las principales colecciones de información manejadas en este estudio son:
- Participantes en el estudio.
- Genética, que incluye documentos de genes y de riesgo de cáncer.
- Variantes, variaciones genéticas anotadas y otro tipo de información propia del proyecto.

A estas colecciones se accedía a través de una API Rest propietaria en una aplicación donde no se realizaba control de acceso.

Para el proyecto no se ha dispuesto de dicha base de datos, sólo de una pequeña muestra de documentos JSON que ejemplifican las principales colecciones e información manejadas

III-B. Arranque del proyecto y decisiones iniciales

Tras la toma de requisitos, identificación de los datos de interés, actores (roles) y restricciones impuestas por el RGPD, comenzó una primera fase de búsqueda de bibliografía científica y proyectos significativos que nos llevó a revisar más de 50 referencias bibliográficas. Tras ello se realizó una propuesta con las tres acciones u objetivos mencionados en la Sección I: (i) normalizar los datos, (ii) securizar el acceso a los mismos con un sistema de autorización basada en el propietario de los datos, así como facilitar la interacción con terceras partes, y (iii) permitir la trazabilidad del sistema con blockchain. Para cumplir los objetivos propuestos se plantea una planificación con tres tareas a realizar de forma secuencial:

1. Proporcionar una interfaz normalizada de acceso que siga el estándar FHIR [13];
2. Aplicar el marco de seguridad XACML/SAML [14], [15] como mecanismo de grano fino para el control de acceso. XACML permite un modelo distribuido donde se separan la decisión de acceso, del recurso al que se desea acceder, lo que permite la actualización directa de cualquier cambio de política de autorización generada por el cliente.
3. Utilizar blockchain para mejorar la trazabilidad del acceso a la información.

Dados los objetivos del proyecto y los recursos disponibles, se decidió limitar el alcance en la implementación a un piloto de pruebas o demostrador que facilite la toma de decisiones de cara a abordar un proyecto en el futuro con un alcance más ambicioso. La integración de la información disponible en un sistema completo de historia clínica electrónica (EHR) es una tarea compleja. Por ello, y en aras de disponer de una plataforma en la que poder comprobar los avances de la solución propuesta, para la implementación del piloto de pruebas se ha optado por las siguientes simplificaciones:
- Hacer sólo la integración como servidor elemental; es decir, se permiten peticiones de los clientes al sistema siguiendo la interfaz REST del estándar FHIR [16], pero no se desarrolla ningún flujo de trabajo complejo típico de la vista de empresa (p.ej. alta de paciente, etc.).
- Integrar en el sistema sólo algunos recursos REST a modo de demostración. No tiene sentido modelar e implementar como recursos REST todos los elementos de información disponibles en el proyecto hasta que no se tome una decisión sobre el futuro del mismo.

En cuanto a las tecnologías utilizadas se recurre a Carbon (de WSO2) como middleware y a las soluciones API Manager [17] e Identity Server [18] (de WSO2) para dar soporte a la identificación, control de acceso y actividades de auditoría. Se propone además el uso de blockchain [19] sobre los ficheros de *log* que trazan el acceso a recursos y la definición de políticas de usuario para garantizar la integridad y trazabilidad.

III-C. Diseño propuesto

En la figura 1 se resume el diseño del sistema propuesto. El paciente, propietario de los datos, utiliza una aplicación (PAP) con una interfaz simple que le permite definir las políticas de acceso a sus datos. Cualquier modificación de estas políticas generará un registro en la traza general del sistema (*log*).

Cuando un usuario con cierto rol (p.ej., Juan, en la figura anterior) quiere acceder a los datos, el PEP interceptará la URI con petición y le añadirá información de contexto (p.ej. fecha/hora) antes de consultar al PDP si puede o no cursar la petición. En caso afirmativo enviará la consulta al servidor que maneja los recursos solicitados en la URI, reenviando la respuesta al usuario. El fichero de *log* registrará

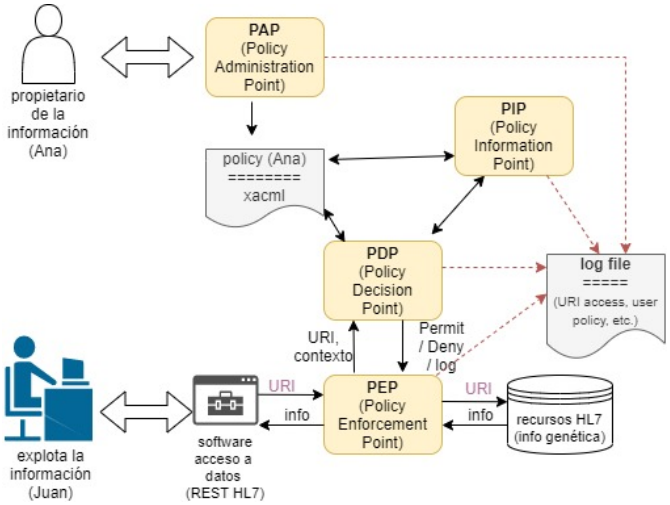


Figura 1. Arquitectura general de la solución

Tabla I
CAMPOS IDENTIFICADOS EN EL ESTÁNDAR.

Person (paciente)	Patient (paciente)
	OccupationalData
Person (madre)	Patient (madre)
	RelatedPerson (de paciente)
Person (padre)	Patient (padre)
	RelatedPerson (de paciente)
Person (conyuge)	RelatedPerson (conyuge)
Observation	Talla del paciente
	Peso del paciente
	Embarazos de la madre
	Deporte del paciente
	Tabaco del paciente
Condition	Muerte padre
	Muerte madre
	Una por enfermedad que tenga
Consent	Paciente
	Padre
	Madre
DiagnosticReport	Comentarios técnicos del paciente

todas las peticiones de PIP, PDP y PEP, lo que permitirá una trazabilidad completa de toda la actividad. Este fichero formará parte de la blockchain para garantizar su integridad.

A continuación se detallan los trabajos incluidos en cada una de las tres líneas de acción realizadas.

III-D. Integración de la información en un sistema normalizado de Historia Clínica

La primera tarea consiste en analizar cómo representar los recursos de nuestro servicio a través de los recursos FHIR normalizados y crear el conector que proporcione la API FHIR para gestionar los mismos desde los clientes. Las interfaces de un servicio SOA suelen estar más alineadas con la funcionalidad de la organización, reflejada en la vista empresarial. Podríamos referirnos a casos de uso como "Registrar un paciente" o "Reservar un quirófano". Sin embargo, la normalización de FHIR se da en términos de grano "fino", como operaciones CRUD sobre recursos. Por ello, en el proyecto, las API FHIR normalmente se encapsularán en servicios de grano más grueso, que a su vez se publicarán dentro de la SOA.

Siguiendo las muestras de datos a las que tiene acceso el equipo de trabajo, se definen 3 API que sigan el standard FHIR:

- API Participantes: se encarga de ofrecer todas las operaciones relacionadas con los datos de los participantes del proyecto. Se contemplan todas las operaciones CRUD.
- API Genética: comprende las operaciones que sirven la información tanto de genes como del cáncer. Al no ser propia, sino que es la unión de múltiples bases de datos disponibles, solo existe la opción de consulta.
- API Variantes: en este caso, información genética que sí es propia del proyecto, por ejemplo, las variantes genéticas anotadas.

Los recursos de la API Participantes están formados por 55 campos que han sido adaptados a los correspondientes del estándar FHIR HL7 tal y como se muestra en la Tabla I.

En la implementación se ha utilizado la librería *cryptography* de *Python* para la implementación de interfaces de bajo nivel para algoritmos criptográficos para ciertos campos, como son: enfermedades, DNI y Muestras (*samples*), ya que estas son las que relacionan los datos de los participantes con las secuenciaciones genéticas. El resto de información no es preciso cifrarla, pues la información genética está completamente anonimizada y la única manera de identificar las muestras es mediante la información ya cifrada.

Para la gestión de datos genéticos de pacientes se identifican distintas posibilidades. Existe un formato estandarizado ISO 25720 (ASUML) [20] de representación de este tipo de información. Pero también en FHIR se propone cómo podría incorporarse la información genética a la norma HL7 [21] y, por tanto, hacerla parte de la historia clínica del paciente, incluyendo el consejo genético. Siendo consecuentes con la elección que hemos hecho para la API de participantes esta sería la opción a elegir. FHIR DSTU2 incluye un perfil genético estandarizado aplicable a los recursos de tipo *Observation*, de modo que este recurso pueda incluir resultados de análisis genéticos de forma estandarizada. FHIR STU3 va un poco más allá, permitiendo mayor granularidad y menos ambigüedad al crear un nuevo recurso, *MolecularSequence*, que se usará para representar datos de secuencia genética relevantes para la práctica clínica y que podrá integrarse en otros recursos (entre ellos también *Observation*). Dado que no disponíamos de datos genéticos no se ha implementado esta parte en el demostrador.

Con respecto al piloto implementado con fines demostrativos, los recursos FHIR que se ponen disponibles corresponden a la API Participantes y han sido: *Patient*, *Person*, *Related* Person, *Observation*, *DiagnosticReport*, *OccupationalData*, *Condition* y *Consent*. No se han implementado recursos FHIR asociados a la API Genética y Variante al carecer de datos de ejemplo.

El piloto realizado incluye el desarrollo de una capa FHIR REST que devuelve los recursos anteriormente descritos de forma normalizada a partir de consultas a los documentos JSON proporcionados almacenados en MongoDB. Para implementar dicho servidor se ha recurrido a la API HapiFHIR [22], que está desarrollada usando el *framework* Spring y permite desarrollar servidores REST de forma relativamente sencilla. En nuestro caso los datos persistentes se almacenan en una base de datos no relacional (MongoDB), a la que se

accede a través de la API JPA. Por simplicidad el servidor sólo soporta las peticiones GET a los recursos FHIR *Person* y *RelatedPerson*. Entendemos que la implementación del resto de recursos del éstandar FHIR resulta una tarea sencilla, pero que no aporta mayor valor al piloto realizado.

Un componente fundamental en una SOA es el gestor de API, es decir, el gestor de las interfaces a los servicios que conforman la SOA. Éste da soporte tanto a los publicadores del servicio como a los clientes. A los primeros les facilita la creación, les permite publicar sus interfaces, monitorizar el uso que se hace de las mismas y configurar interceptores de las peticiones (API Gateway) que pueden tener cualquier función, como por ejemplo la facturación por el uso, el registro de eventos, aplicar seguridad, etc. A los clientes o consumidores les facilita la localización y la suscripción a APIs, así como la gestión de las aplicaciones cliente, por ejemplo la creación de claves y *tokens* para el acceso a los servicios.

Una vez que nuestro servidor es conforme a FHIR se publica en el Gestor de APIs de WSO2. Éste nos va a permitir publicar nuestro servicio a través de un proxy, es decir, publicar un punto de acceso controlado a los clientes de nuestra API FHIR.

III-E. Arquitectura del control de acceso

El diseño del modelo de control de acceso se ha basado en atributos (modelo ABAC[**?**]), que permite la especificación de políticas con cualquier nivel granularidad. Las reglas expresadas en las políticas, en base a las cuales se permite o deniega una solicitud de acceso, pueden considerar características referentes al objetivo, al iniciador, a la propia petición y a cualquier aspecto del contexto de la misma. En el contexto de aplicación que nos ocupa, el sanitario, los objetivos (elementos accesibles) se expresan como recursos FHIR, las características del iniciador con información demográfica de usuarios, atributos de organizaciones y de las aplicaciones cliente y los tipos de acceso corresponden a las operaciones REST [23], [24], [25].

Dado que el interés del reto está centrado en los aspectos de seguridad mencionados, los componentes WSO2 que hemos utilizado han sido el gestor de APIs (*API Manager*) y el servidor de identidad (*Identity Server*). Del primero nos interesan las capacidades para publicar servicios (como una API), para desarrollar otros nuevos y para controlar los accesos a través del *API Gateway*, que en nuestro caso actuará como PEP. Del segundo sus perfiles como PAP, PDP y PIP, todos ellos con interfaces estándar XACML/SAML.

En el trabajo desarrollado se ha configurado en el *API Gateway* un interceptor de tipo *Entittlement mediator*, que actúa como Punto de Aplicación de Políticas (PEP) en la arquitectura de control de acceso. El proceso se resume en la Fig. 2. El PEP intercepta la solicitud REST del cliente para acceder a un recurso (1) y consulta (empleando XACML/SAML) al Punto de Decisión de Política (PDP) (2), del que recibirá una respuesta (6) (No/Si/Si con obligaciones). Si la respuesta es afirmativa reenviará la petición del Usuario al servidor de datos FHIR, y la respuesta viajará de vuelta al cliente.

Tal y como se ha descrito en el apartado anterior, la arquitectura de control de acceso se basa en XACML/SAML y tiene cuatro componentes principales:

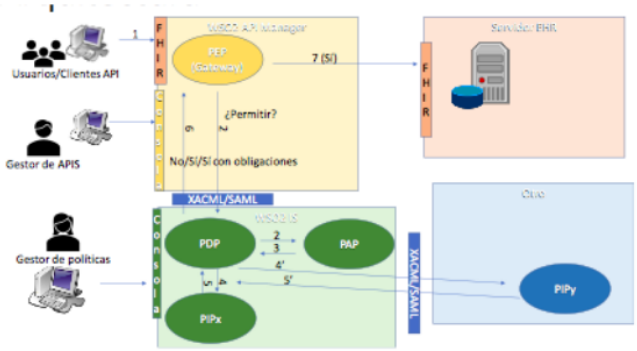


Figura 2. Arquitectura para el control de acceso en una SOA sanitaria

1. El punto de aplicación de políticas (PEP), del que hemos hablado anteriormente, y que como hemos comentado se implementa en este caso concreto en el APIGateway del API Manager.
2. El punto de decisión de políticas (PDP), que toma la decisión sobre permitir o no el acceso a un recurso por parte de una entidad determinada.
3. El punto de administración de políticas, que facilita la gestión (PAP) de las políticas de control de acceso (edición, prueba, publicación, etc. . .).
4. El punto de información de políticas (PIP), que facilita el acceso a la información requerida para evaluar las políticas.

Estos tres últimos componentes de la arquitectura están implementados, en este caso, en el Servidor de Identidad de WSO2. Por otro lado este componente también nos proporciona funcionalidad de gestor de claves (*Key Manager*) para los *token* de autorización de Oauth2, que permiten verificar la identidad del usuario. La Fig. 2 muestra el esquema final implementado.

Cada uno de los componentes se ha instanciado en una máquina virtual. En el demostrador se ha usado VirtualBox y se han creado tres instancias:

- una para la ejecución del servidor de FHIR.
- otra para instalar el WSO2 API Manager.
- y una tercera para el despliegue del WSO2 Identity Server.

Donde el paciente, dueño de los datos (genéticos o de cualquier otro tipo), utiliza la interfaz al PAP del Identity Server que le permite definir las políticas de acceso a sus datos. Cualquier actualización de estas políticas generará un registro en la traza general del sistema (*log*).

III-F. Soporte de auditoría (registro de eventos enriquecido con blockchain)

Para mejorar la seguridad y facilitar la trazabilidad de acceso a la información, proponemos un sistema BlockChain que se encargue de la protección de la información[1] referente a los accesos a recursos del sistema FHIR, así como de las tareas de gestión de políticas. En particular, de:

[1]No se observa utilidad práctica en la introducción de los datos clínicos (sean o no genómicos) del paciente en la cadena a custodiar en BlockChain, ya que con el esquema de acceso anterior, que desacopla el acceso de los datos, bastaría con que dichos datos estuvieran cifrados en uno o varios servidores (backup) propiedad de la entidad que obtuviera la cesión por parte del usuario.

- registro de accesos a recursos (URI) de pacientes (información genómica o de otro tipo). Los accesos pueden ser para la creación, lectura o escritura por parte de los iniciadores.
- registro de definición de políticas por parte del paciente.

Las ventajas de implementar este sistema serían:

- Permite la trazabilidad de las acciones ejecutadas sobre los recursos, aunque estuvieran realizadas por diferentes actores.
- Impide que la entidad que tiene la encomienda del acceso a los recursos pueda eliminar o modificar la trazabilidad de todas las acciones ocurridas sobre los recursos
- Evita el repudio por parte del dueño de los datos sobre sus propias políticas de seguridad, que formarán parte de la cadena custodiada por el BlockChain. Esto simplifica la responsabilidad en el caso de accesos inadecuados.

Se ha creado un piloto de pruebas utilizando librerías existentes de código libre. Una vez conocida la naturaleza de los datos a custodiar se evaluaron las diversas alternativas en el campo del software libre para implementar la cadena de bloque, de forma que dicha información no pudiera ser alterada. La solución elegida para el piloto demostrador fue HyperLedger [26].

En el piloto demostrador se configuraron todos los elementos del sistema y se cebaron con datos de ejemplo. Este demostrador permitió analizar en detalle los siguientes pasos:

- Inclusión de varias políticas de acceso, con la intención de mostrar accesos permitidos o denegados según el usuario que accede.
- Realización de diversas peticiones GET a un recurso Person del *backend*, empleando directamente *curl* al endpoint del API Manager que hace de proxy hacia el *backend* (servidor FHIR).
- Análisis del resultado, comprobando los accesos permitidos (con la recepción del recurso solicitado) o rechazado (con la recepción del mensaje correspondiente).
- Realización de un script que permite la ejecución automática de estas peticiones.
- Creación de la BlockChain del fichero *log*.

IV. Conclusiones

El control de acceso a los datos es un requisito básico en el dominio sanitario, como lo es la capacidad de integración e interoperabilidad de cualquier solución. En este trabajo se aborda el reto de lograr ambos requisitos a través del diseño y la implementación de un piloto donde se ha enriquecido a una aplicación propietaria dotándola de una arquitectura SOA con interfaces estándar propios del mundo sanitario (FHIR) y añadiéndole un sistema de gestión y control de accesos distribuido que usa XACML/SAML. Es compatible con el estándar X.812 y está basado en atributos, lo que permite personalizar la granularidad de las políticas y la gestión de las mismas por parte de los propios usuarios. El uso de las soluciones WSO2 ha permitido acelerar el proceso facilitando algunos aspectos esenciales que han reducido el coste de desarrollo. Por ejemplo, al usar el *Identity Server* se proporciona una interfaz de gestión de políticas de control de acceso, que facilita la capacitación del usuario para la autogestión de las mismas.

La utilización de BlockChain en el fichero de registros del sistema de control de accesos permite la trazabilidad de las acciones ejecutadas sobre los recursos e impide ataques a la propia trazabilidad por parte de un sólo actor, además de evitar el repudio por parte del dueño de los datos sobre sus propias políticas de seguridad.

Con el propotipo se busca obtener una evaluación preliminar sobre el coste y el beneficio de desarrollar el sistema completo antes de dedicar más recursos. En caso afirmativo, los siguientes pasos serían la implementación de los datos genéticos como recursos y la realización de pruebas de rendimiento de la cadena de bloques utilizada.

Agradecimientos

Referencias

[1] ITU-T, *Information technology – Open Distributed Processing – Reference Model: Overview*, Intenational Telecommunications Union ITU ITU-T Rec. X.901, 1997.

[2] ——, *Information technology – Open Distributed Processing – Reference model: Foundations*, Intenational Telecommunications Union ITU ITU-T Rec. X.902, 2009.

[3] ——, *Information technology – Open distributed processing – Reference model: Architecture*, Intenational Telecommunications Union ITU ITU-T Rec. X.903, 2009.

[4] ——, *Information technology – Open Distributed Processing – Reference Model: Architectural Semantics*, Intenational Telecommunications Union ITU ITU-T Rec. X.904, 1997.

[5] "Informática sanitaria. comunicación de la historia clínica electrónica." UNE-EN ISO 13606-1, 2013.

[6] HL7. (2011) Hl7 version 2 product suite. [Online]. Available: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=185

[7] M. Rosa, C. Faria, A. M. Barbosa, H. Caravau, A. F. Rosa, and N. P. Rocha, "A Fast Healthcare Interoperability Resources (FHIR) Implementation Integrating Complex Security Mechanisms," *Procedia Computer Science*, vol. 164, pp. 524–531, jan 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050919322628

[8] M. Rosa, C. Faria, A. M. Barbosa, A. I. Martins, A. F. Almeida, and N. P. Rocha, "A Platform of Services to Support Community-Dwelling Older Adults Integrating FHIR and Complex Security Mechanisms," *Procedia Computer Science*, vol. 160, pp. 314–321, jan 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050919317855

[9] N. Hong, A. Wen, D. J. Stone, S. Tsuji, P. R. Kingsbury, L. V. Rasmussen, J. A. Pacheco, P. Adekkanattu, F. Wang, Y. Luo, J. Pathak, H. Liu, and G. Jiang, "Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries," *Journal of Biomedical Informatics*, vol. 99, p. 103310, nov 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046419302291

[10] A. V. Karhade, J. H. Schwab, G. Del Fiol, and K. Kawamoto, "SMART on FHIR in spine: integrating clinical prediction models into electronic health records for precision medicine at the point of care," *Spine Journal*, jun 2020.

[11] M. Soni and D. K. Singh, "Blockchain-based security & privacy for biomedical and healthcare information exchange systems," *Materials Today: Proceedings*, feb 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2214785321011561

[12] M. Sookhak, M. R. Jabbarpour, N. S. Safa, and F. R. Yu, "Blockchain and smart contract for access control in healthcare: A survey, issues and challenges, and open issues," *Journal of Network and Computer Applications*, vol. 178, p. 102950, mar 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1084804520304045

[13] FHIRr4, *HL7 Fast Healthcare Interoperability Resources, Release 4)*, Health Level Seven International FHIR Management Group HL7 FHIR R4, 2018.

[14] OASYS, *Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML)*, OASYS Open OASIS eXtensible Access Control Markup Language (XACML) TC R2.0, 2005.

[15] OASIS, *eXtensible Access Control Markup Language (XACML)*, OASYS Open OASIS eXtensible Access Control Markup Language (XACML) TC R3.0, 2013.

[16] R. Saripalle, C. Runyan, and M. Russell, "Using HL7 FHIR to achieve interoperability in patient health record," *Journal of Biomedical Informatics*, vol. 94, p. 103188, jun 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046419301066
[17] WSO2. (2020) Api manager documentation 3.2.0. [Online]. Available: https://apim.docs.wso2.com/en/latest/
[18] ——. (2020) Api manager documentation 5.11.0. [Online]. Available: https://is.docs.wso2.com/en/latest/
[19] D. Di Francesco Maesa, P. Mori, and L. Ricci, "A blockchain based approach for the definition of auditable Access Control systems," *Computers & Security*, vol. 84, pp. 93–119, jul 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167404818309398
[20] "Health informatics – genomic sequence variation markup language (gsvml)," UNE-EN ISO 25720, 2009.
[21] HL7. (2018) Genomics implementer guidance. [Online]. Available: https://www.hl7.org/fhir/genomics.html
[22] H. FHIR. (2021) The open source fhir api for java. [Online]. Available: https://hapifhir.io/
[23] Y. K. Rivera Sánchez, S. A. Demurjian, and M. S. Baihan, "A service-based RBAC & MAC approach incorporated into the FHIR standard," *Digital Communications and Networks*, vol. 5, no. 4, pp. 214–225, nov 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352864817301803
[24] S. Pal, M. Hitchens, V. Varadharajan, and T. Rabehaja, "Policy-based access control for constrained healthcare resources in the context of the Internet of Things," *Journal of Network and Computer Applications*, vol. 139, pp. 57–74, aug 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1084804519301377
[25] K. Riad and J. Cheng, "Adaptive XACML access policies for heterogeneous distributed IoT environments," *Information Sciences*, vol. 548, pp. 135–152, feb 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0020025520309634
[26] T. L. Foundation. (2021) Open source blockchain technologies. [Online]. Available: https://www.hyperledger.org/

Sesión de Formación e Innovación Educativa

Sesión de Formación e Innovación Educativa
http://doi.org/10.18239/jornadas_2021.34.62

Guía de resolución de pruebas CTF para adquirir habilidades de seguridad informática y análisis forense

José Carlos Sancho Núñez
Universidad de Extremadura
jcsanchon@unex.es
ORCID: 0000-0002-4584-6945

Delia Mª Pablo Rodríguez
Mobbeel
Edificio Garaje 2.0. Cáceres
dpabloro@alumnos.unex.es

Andrés Caro Lindo
Universidad de Extremadura
andresc@unex.es
ORCID: 0000-0002-6367-2694

Resumen- **La falta de perfiles especializados y de necesidad formativa en el ámbito de la ciberseguridad, ha hecho que las competiciones del tipo *Capture the Flag* (CTF) se hayan incrementado en los últimos años. La finalidad de estas pruebas es identificar talento que incorporar a empresas del sector tecnológico de manera rápida. Debido a la importancia formativa dada a este tipo de pruebas cuyo enfoque es el Aprendizaje Basado en Retos, esta contribución analiza un total de 125 retos de CTF englobados en las disciplinas de esteganografía, análisis forense y hacking web. El objetivo es conocer si existen coincidencias o correlaciones en el proceso de resolución de los retos. Como resultado de dicho análisis profundo, se presenta una guía de resolución de pruebas CTF que, a modo de esquemas, sirva de apoyo a los usuarios o equipos para adquirir habilidades en seguridad informática y análisis forense.**

***Index Terms*- Capture The Flag, CTF, ciberseguridad, educación, innovación formativa, detección talento.**

Tipo de contribución: *Formación innovación*

I. Introducción

La ciberseguridad es una disciplina de la informática compleja y poco estudiada en las diferentes etapas educativas. Sin embargo, el aumento de ciberataques en los últimos años como indican los ciberincidentes gestionados por el Centro Criptológico Nacional (CCN-CERT) de los años 2016 (20.940), 2017 (26.500), 2018 (38.029) y 2019 (42.997) [1]–[4], ha puesto de manifiesto la necesidad de disponer de un número mayor de profesionales con conocimientos especializados en este ámbito.

Por este motivo, se ha puesto de moda utilizar las competiciones del tipo *Capture The Flag* (CTF*)* para identificar talento de forma rápida. Los tipos de retos denominados CTF*,* en español Captura la Bandera, consisten en resolver una prueba o desafío informático con el fin de encontrar la bandera que, en este caso, representa la solución. Este enfoque pedagógico consigue que los usuarios tengan que demostrar conocimientos específicos, investigar todas las posibilidades ante un problema, desarrollar un proceso de estudio y elegir el camino más adecuado para encontrar la solución.

La finalidad principal de esta contribución es descubrir si los retos CTF siguen un patrón de pasos para llegar a su resolución o si, por el contrario, no es posible descubrir dicha correlación y deben ser tratados de manera independiente. En el caso de se identifique la existencia de un patrón, la idea es crear un recurso de apoyo que ayude y facilite la resolución de los retos a usuarios o equipos que se enfrentan a competiciones del tipo CTF.

Como aportaciones fundamentales esta investigación analiza multitud de retos pertenecientes a las disciplinas de esteganografía, análisis forense y hacking web, e identifica patrones de resolución que presenta a modo de esquemas como apoyo a la resolución de retos CTF.

II. Capture The Flag

Para contextualizarse en el ámbito de los *Capture the Flag* es preciso conocer las distintas tipologías de competiciones que existen, las disciplinas que componen los retos y la multitud de recursos que existen para formarse en este entorno.

A. Tipologías de CTF

En función del estilo o enfoque que se le aplica a las pruebas de CTF se pueden clasificar en estilo *Jeopardy*, *Attack-Defense* o mixto, siendo este último la combinación de los anteriores.

- *Jeopardy*: Los usuarios o los equipos deben completar tantos desafíos de ciberseguridad como puedan de una selección proporcionada. Se van obteniendo puntos conforme se van resolviendo los retos, donde la puntuación de cada prueba es proporcional a la dificultad para encontrar la bandera. Generalmente, no se computa el tiempo necesario para superar cada desafío. El equipo o persona ganadora es la que más puntos suma.

- *Attack-Defense*: Los usuarios o los equipos disponen de una red o host con servicios vulnerables en ejecución, en la que disponen de un tiempo determinado para proteger los servicios vulnerables y/o para construir *exploits* (programas de ataque) que ataquen los servicios de otros usuarios o equipos. Se puntúan por un lado las acciones de ataque y, por otro, las de defensa, de manera que la puntuación total se obtiene protegiendo los servicios propios de un usuario o equipo y vulnerando los del resto de oponentes.

- Mixto: Se produce una combinación de las opciones anteriores, es decir, las competiciones mezclan los formatos existentes. Por ejemplo, es posible que en un desafío del tipo *Jeopardy* tengas un tiempo límite para resolver cada prueba.

B. Disciplinas de CTF

Los retos propuestos en los CTF suelen tener distinto nivel de complejidad y abordan diversas disciplinas dentro de la ciberseguridad. A continuación, se introducen las disciplinas más utilizadas en los CTF:

- Análisis Forense (*Forensics*): Se deben aplicar técnicas que analicen en detalle imágenes de memoria, de discos duros o capturas de red. Es la disciplina más utilizada en los CTF.

- Criptografía (*Crypto*): Es preciso aplicar técnicas que descifren representaciones lingüísticas de ciertos mensajes con el fin de hacerlos legibles y entendibles.

- Esteganografía (*Stego*): Requiere la aplicación de técnicas para desvelar información u objetos que han sido ocultos dentro de otros ficheros informáticos, de forma que a simple vista no se detectaba su presencia. Se considera la tecnología más ingeniosa para la ocultación de datos o información.

- Explotación de software (*Pwn*): Se utilizan técnicas de programación para ejecutar un código fuente que consigue vulnerar una aplicación y averiguar la información requerida.

- Ingeniería Inversa (*Reversing*): Se hace uso de técnicas que requieren la ejecución de un proceso inverso a como se acostumbra a realizar. Un ejemplo es la descompilación de ficheros ejecutables en formatos (*bin, exe, elf, APK, etc.*).

- Programación (*Professional Programming & Coding*): Precisan conocimientos avanzados de programación en diferentes lenguajes para crear *scripts* o programas que consigan una tarea determinada.

- *Hacking web*: Utilizan técnicas de *testing* de seguridad para descubrir y vulnerar aplicaciones web. Los ataques más frecuentes son *Inyección SQL, Cross-Site Scripting (XSS),* fuerza bruta*, CRLF, CSRF, etc.*

- Reconocimiento (*Recon*): Se realiza la búsqueda de la bandera en distintos sitios de Internet siguiendo las pistas proporcionadas como nombres de personas o lugares emblemáticos.

En la mayoría de CTF se produce una mezcla de las disciplinas explicadas.

C. *Recursos de pruebas CTF*

El auge de los últimos años en esta tipología de competiciones se ve reflejado en la existencia de multitud de recursos en Internet que permiten desarrollar habilidades en seguridad informática de manera legal y didáctica. A continuación, se clasifican los recursos en plataformas de competición, sitios web especializados y conferencias.

- Plataformas de competición: Existen plataformas de aprendizaje donde se alojan diversos retos de ciberseguridad y entornos virtuales con multitud de desafíos. Ejemplos de plataformas son: *picoCTF, CTF Time, Root Me, Hack Me, Hack The Box, CTF365, Atenea*, etc.

- Páginas web: Muchos sitios web recogen los *write-ups* de retos pertenecientes a eventos pasados. Un *write-ups* es la redacción justificada que detalla los pasos que un competidor ha seguido para resolver el reto correspondiente. Algunos sitios web en los que podemos encontrar estas redacciones son: *CTF Time, hackplayers, ironhackers, InfoSec, etc.*

- Conferencias: Los últimos años han acogido multitud de eventos con presentaciones de expertos, talleres y CTF, dirigidos a personas que desean iniciarse en la cultura "hacking". Ejemplos de conferencias que albergan CTF son: *ROOTCON, DEFCON* y *CyberCamp*.

III. CTF ORGANIZADOS

En los últimos años, la Universidad de Extremadura en colaboración con la empresa Viewnext S.L. ha organizado diferentes eventos del tipo CTF con el objetivo de concienciar y detectar talento en una disciplina de la informática tan especializada como es la ciberseguridad.

A. *Cybersecurity Challenge VIEWNEXT-UEx - 2018*

El *CyberSecurity Challenge VIEWNEXT-UEx* se llevó a cabo en el año 2018 dentro del evento *ForoCIBER,* un evento divulgativo sobre ciberseguridad y derecho tecnológico en Extremadura. Como muestra la Fig.1 la competición se desarrolló durante 72 horas de forma virtual, a través de la plataforma de formación del Campus Virtual de la Universidad de Extremadura.

Tuvo una participación de un total de 132 usuarios del ámbito nacional e internacional. De todos ellos, 11 finalizaron con éxito los 5 retos propuestos y 41 consiguieron superar alguno de los retos. Sobre la filiación de los participantes el 57% (75) fueron trabajadores cualificados del sector TIC o miembros de Fuerzas y Cuerpos de Seguridad del Estado, el 41% (54) estudiantes y el 2% (3) personas en desempleo.

Fig. 1. Espacio virtual utilizado para el *Cybersecurity Challenge.*

B. *Capture the Flag JNIC-VIEWNEXT - 2019*

El *Capture the Flag JNIC-VIEWNEXT* se celebró como actividad complementaria de las *V Jornadas Nacionales de Investigación en Ciberseguridad (JNIC)* celebradas en Cáceres del 5 al 7 de junio del año 2019.

Al igual que el CTF anterior se llevó a cabo de manera virtual durante 72 horas. En la Fig. 2 se puede observar el espacio utilizado para el desarrollo de la competición donde participaron un total de 97 usuarios.

En esta ocasión, ninguno de los usuarios superó todos los retos propuestos, que fueron realizados por sendos Trabajos Fin de Grado de estudiantes del Grado en Ingeniería en Informática en Ingeniería del Software [5], [6].

Fig. 2. Espacio virtual utilizado para el *CTF JNIC-VIEWNEXT.*

IV. Materiales

Entre las disciplinas expuestas con anterioridad se seleccionan las tres más habituales en este tipo de competiciones, como son esteganografía, análisis forense y hacking web.

La Tabla I recoge el número de retos seleccionados por disciplina, siendo un total de 125 pruebas de CTF analizados en esta contribución.

Tabla I
NÚMERO DE RETOS SELECCIONADOS PARA LA CONSTRUCCIÓN DE LA METODOLOGÍA. FUENTE: PROPIA.

Disciplina	Número de retos seleccionados
Esteganografía	33
Análisis forense	42
Hacking web	50
Total	125

En apartados posteriores se indican las referencias de los CTF analizados, de modo que se puedan contrastar las fuentes de información utilizadas.

V. Metodología

La metodología utilizada para la creación de un patrón de resolución de pruebas CTF se basa en analizar 125 retos y medir el número de veces (concurrencia) que se repita una acción entre todas las soluciones de los retos estudiados para cada disciplina concreta.

Por ejemplo, en un reto de esteganografía es preciso conocer el tipo y el formato real de cada fichero, ya que es frecuente que un fichero no sea del tipo que dice o parece ser. Por lo tanto esta acción se considera comprobar la extensión de los ficheros. Igualmente, ocurre con la búsqueda de archivos ocultos, el análisis de metadatos o descifrar mensajes encontrados.

Una vez conocida la concurrencia de las acciones para cada disciplina es preciso sintetizar y ordenar los resultados de manera que se consiga facilitar la interpretación de los pasos a seguir para resolver los retos de una disciplina.

Por ello, la síntesis de los resultados obtenidos se presenta para cada disciplina a modo de esquema, indicando los pasos o posibilidades que debe realizar el usuario o equipo en función del avance de la resolución de los retos.

VI. Resultados

En base a la concurrencia de las acciones realizadas para resolver los retos, se presentan los resultados de cada una de las disciplinas identificadas, junto a los diferentes esquemas creados que servirán de ayuda a los usuarios y equipos que se enfrentan a los retos de CTF.

A. Esteganografía

En esta disciplina se analizan un total de 33 retos. En primer lugar, se distinguen los tipos de archivo encontrados en los retos, debido a que en el proceso de análisis se ha podido comprobar que las acciones realizadas sobre ellos son diferentes.

Del análisis de los 33 retos, se identifican 27 ficheros de imágenes, 9 ficheros comprimidos, 9 ficheros de audio y 2 ficheros de vídeo.

Por tanto, cuando nos enfrentemos a retos de esteganografía lo más probable es que nos encontremos con imágenes, que se suelen proporcionar directamente o mediante un enlace para su descarga. Rara vez vamos a trabajar con vídeos, que además hay escasas herramientas para tratarlos, pero es posible que nos encontremos algún archivo de audio o .zip, que puede contener cualquiera de los tipos nombrados, siendo las imágenes el recurso más habitual.

De cada uno de los tipos de archivo encontrados se han estudiado detalladamente las acciones llevadas a cabo para su resolución, así como la forma o las distintas formas de realizarlas.

Con respecto a los archivos comprimidos encontrados en los retos de esta disciplina se tratan de archivos en formato .zip. Por lo tanto, tal y como la lógica indica, la acción más frecuente y primera acción que se debe realizar es descomprimirlo, para así extraer, analizar su contenido y seguir los pasos que correspondan a los otros tipos de ficheros que se indica más adelante.

En este tipo de ficheros, es posible que durante el proceso de descompresión o extracción de archivos nos soliciten una contraseña, en el caso de no disponer de ella se puede hacer uso para obtenerla de ataques por fuerza bruta o ataques con algún diccionario facilitado o al que llegamos tras alguna pista encontrada. Por ejemplo el famoso diccionario "Rockyou" que es un fichero que almacena más de 14 millones de contraseñas y que ha sido creado con las contraseñas de usuarios que se han ido filtrando con el paso del tiempo. En la Fig. 3 se observan los pasos que debemos seguir en el caso de encontrarnos con retos de esteganografía que contienen archivos comprimidos.

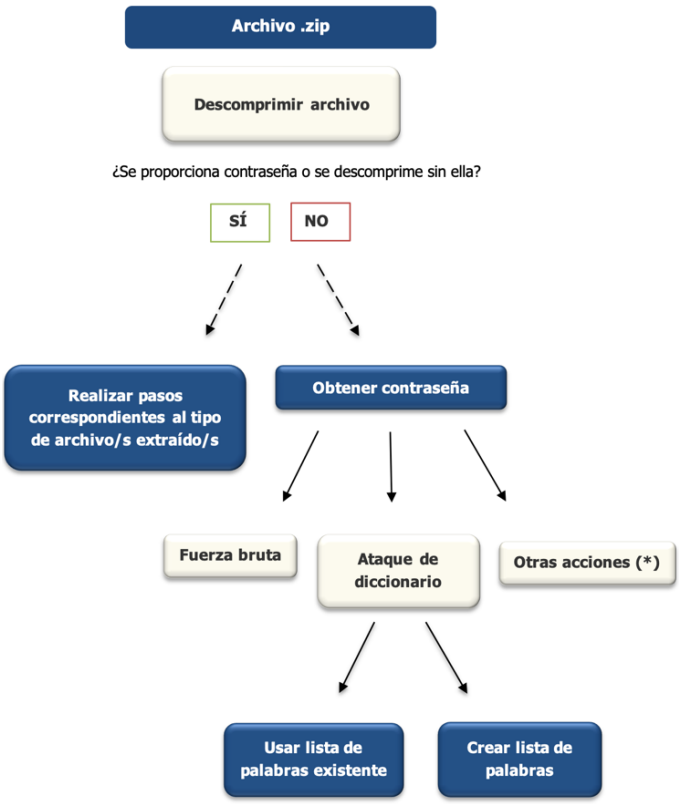


Fig. 3. Esquema de resolución de los retos que contienen archivos comprimidos.

Por otro lado, es frecuente que en los retos de esteganografía los usuarios se encuentren con ficheros de audio. Una de las propiedades que se aprovechan para ocultar la información secreta en los archivos de audio es que tienen un gran espacio. La sensibilidad del sistema auditivo humano hace que detectar la información oculta en los archivos de

audio sea una tarea difícil. Sin embargo, algunas distorsiones ambientales generales pasan desapercibidas para los oyentes en casi todos los casos. Los investigadores explotan estas propiedades para ocultar la información secreta utilizando señales de audio como portadoras.

Entre las acciones para procesar y obtener información de un archivo de audio, la acción más repetida y que más información proporciona se trata de analizar el espectrograma. Debido a que la vista del espectrograma de una pista de audio proporciona una indicación visual de cómo cambia la energía en diferentes bandas de frecuencia con el tiempo y es fácil representar información como palabras o números en el espectrograma. Por tanto, esta es la primera acción que se debe probar cuando un usuario se enfrenta a un reto de esteganografía con archivos de este tipo.

En segundo lugar, si el espectrograma no aporta información, la opción más útil se trata de detectar los tonos DTMF de este tipo de archivos. Los tonos DTMF (multifrecuencia de doble tono) son utilizados en los teléfonos para la marcación por tonos, que suenan cuando se presionan las teclas de número. Para obtener una cadena legible a través de los tonos DTMF se pueden utilizar detectores.

Si no logramos obtener nada tras las acciones anteriores, podemos tratar de tratar de modificar y jugar con sus características para obtener algún mensaje.

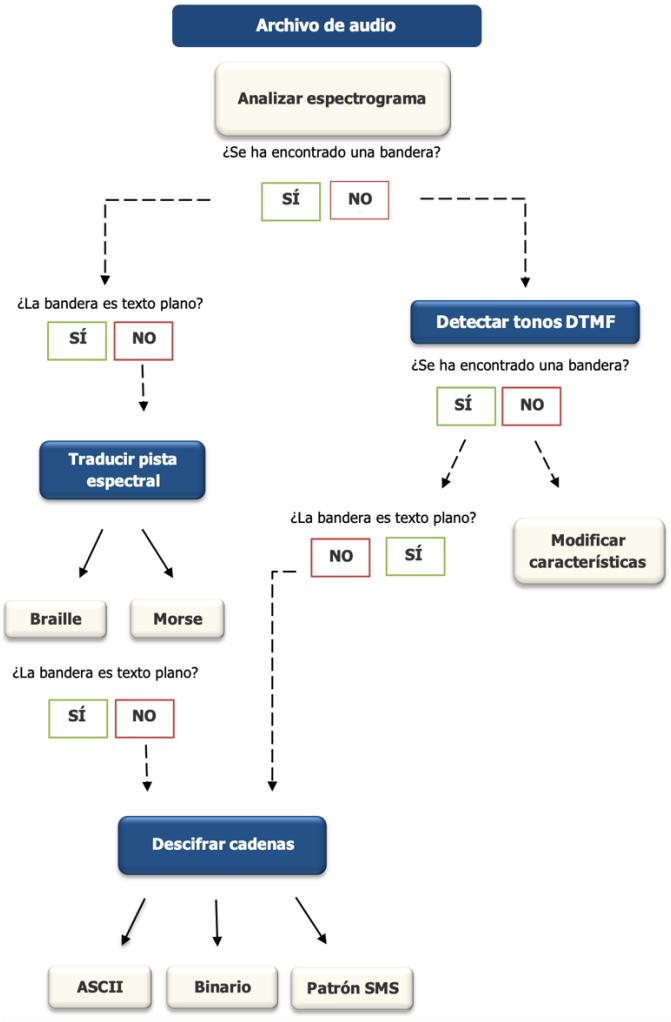


Fig. 4. Esquema de resolución de los retos que contienen archivos de audio.

La esteganografía en imágenes es la técnica más utilizada y de la que mayor número de retos se encuentran.

Los principios en los que se fundamentan las técnicas de ocultación son la modificación de la imagen no introduzca un ruido visual que levante sospechas a una persona que vea la imagen y que las modificaciones introducidas no proporcionen pistas adicionales a un estegoanalista. En este sentido se han publicado una gran variedad de técnicas para diferentes formatos gráficos de fichero (bmp, gif, jpeg, png, etc.).

Estudiando las acciones realizadas para procesar y obtener información de un archivo de imagen, la acción más repetida y que más información proporciona se trata de comprobar la extensión real del archivo proporcionado. Por tanto, esta es la primera acción que se debe probar cuando nos enfrentemos a un reto de esteganografía con archivos de este tipo.

Tras asegurarnos del tipo de archivo que nos proporcionan, la acción que más se ha realizado entre los casos estudiados, se trata de identificar archivos ocultos dentro de la imagen.

Otras operaciones comunes con imágenes son ocultar un mensaje o un texto que para poder verlo se necesita aplicar algunos filtros de color o jugar con los niveles de color.

Las Tablas II y III, que para facilitar su visualización aparecen en la siguiente página, recogen las acciones realizadas sobre los archivos de imagen encontrados en los retos de las disciplinas de esteganografía y de análisis forense.

En base a los resultados obtenidos en las Tablas II y III, la Fig. 5 muestra el proceso de resolución perteneciente a los retos de la disciplina de esteganografía que contienen archivos de imagen.

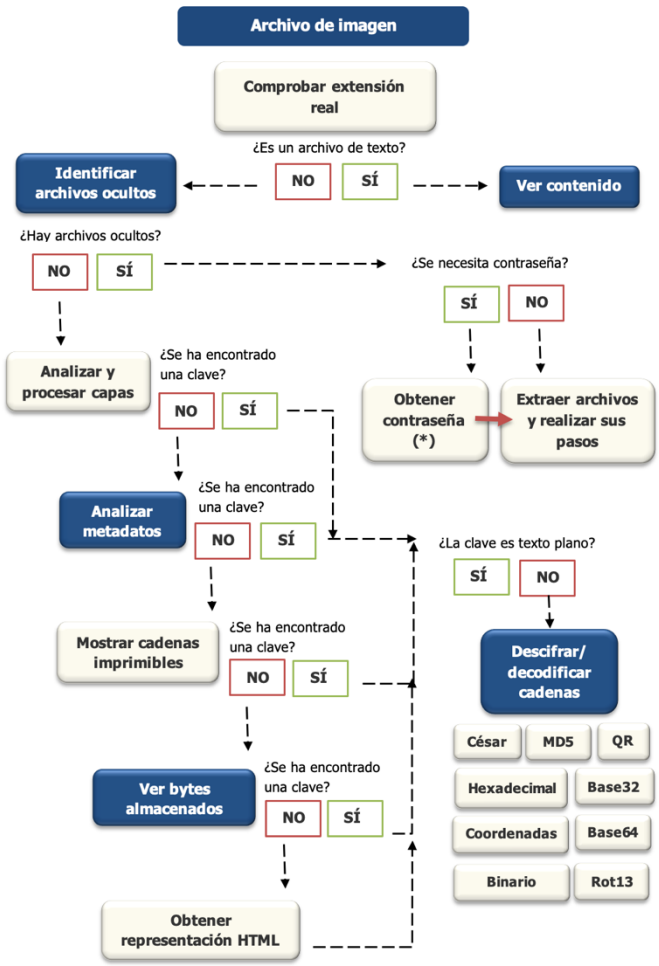


Fig. 5. Esquema de resolución de los retos que contienen imágenes.

Tabla II
LISTADO DE RETOS SOBRE ESTEGANOGRAFÍA QUE CONTIENEN FICHEROS DE IMÁGENES. FUENTE: PROPIA.

Referencia	Comprobar extensión	Imprimir contenido	Archivos ocultos	Analizar metadatos	Mostrar cadenas	Ver bytes	Analizar y procesar	HTML	Descifrar cadenas
[27]	✓	✓			✓				
[28]	✓				✓	✓			✓
[29]								✓	✓
[30]	✓	✓	✓			✓			
[31]				✓					✓
[32]		✓	✓						✓
[33]		✓	✓				✓		✓
[34]	✓			✓	✓				✓
[35]	✓	✓	✓	✓		✓			✓
[36]	✓	✓	✓	✓					✓
[37]	✓	✓	✓	✓			✓		✓
[38]	✓			✓					
[39]		✓	✓		✓				✓
[40]	✓			✓	✓		✓		✓
[41]							✓		✓
[42]		✓				✓			✓
[43]	✓						✓		
[44]	✓		✓	✓		✓	✓		✓
[45]			✓	✓			✓		
[46]	✓	✓	✓	✓					✓
[47]	✓		✓		✓				✓
[48]		✓	✓				✓		✓
[49]			✓				✓		
[50]	✓						✓		
[51]		✓	✓						✓
[52]	✓						✓		
[53]								✓	✓
Resultados	**15**	**12**	**14**	**10**	**6**	**5**	**11**	**2**	**19**

Tabla III
LISTADO DE RETOS SOBRE ÁNALISIS FORENSE QUE CONTIENEN FICHEROS DE IMÁGENES. FUENTE: PROPIA.

Referencia	Comprobar extensión	Imprimir contenido	Archivos ocultos	Analizar metadatos	Mostrar cadenas	Ver bytes	Analizar y procesar	Descifrar cadenas
[5]	✓					✓		
[6]	✓					✓		
[7]		✓	✓		✓		✓	✓
[8]	✓					✓		
[9]				✓				✓
[10]								✓
[11]			✓		✓			
[12]	✓				✓	✓		
[13]	✓							
[14]		✓	✓	✓				✓
[15]			✓					
[16]	✓		✓		✓		✓	✓
[17]						✓		
[18]					✓			
[19]			✓	✓	✓			✓
[20]	✓	✓	✓					
[21]			✓					
[22]	✓	✓	✓					✓
[23]					✓			✓
[24]	✓							
[25]	✓					✓		
[26]								✓
Resultados	**10**	**4**	**9**	**3**	**7**	**6**	**2**	**9**

B. *Análisis forense*

En esta disciplina se analizan 42 retos. De nuevo, se distinguen los tipos de archivo encontrados en los retos, encontrando 25 ficheros de imágenes, 18 ficheros comprimidos, 9 de capturas de paquetes, 7 imágenes de disco, 5 de volcado de memoria RAM y 2 ficheros de audio.

Con respecto a los ficheros de imágenes encontrados en la esta disciplina se deben seguir los pasos de resolución expuestos en la anterior, indicados en la Fig. 5.

A diferencia de la anterior que puede ser coincidente en varias disciplinas, los archivos PCAP (captura de paquetes) que almacenan el tráfico de red capturado solo van a aparecer en la disciplina forense. En este tipo, los paquetes de interés generalmente se encuentran mezclados con tráfico no relacionado, por lo que el análisis de clasificación y el filtrado de datos también es un trabajo necesario. Ejemplos de filtrado son los relacionados de dominios (filtro: DNS), las solicitudes de inicio de sesión (filtro: HTTP) o los paquetes que se envían desde una ip concreta (filtro: 'ip.addr == ipCorrespondiente').

Al analizar detenidamente el contenido de los campos de los paquetes de interés se pueden dar varias situaciones y, por tanto, las acciones a realizar variarán en función a ellas. Estas acciones van desde encontrar texto cifrado, enlaces externos, archivos incrustados o averiguar la contraseña de redes inalámbricas.

La Fig. 6 muestra gráficamente el proceso de resolución perteneciente a los retos de la disciplina de análisis forense que contienen capturas de paquetes.

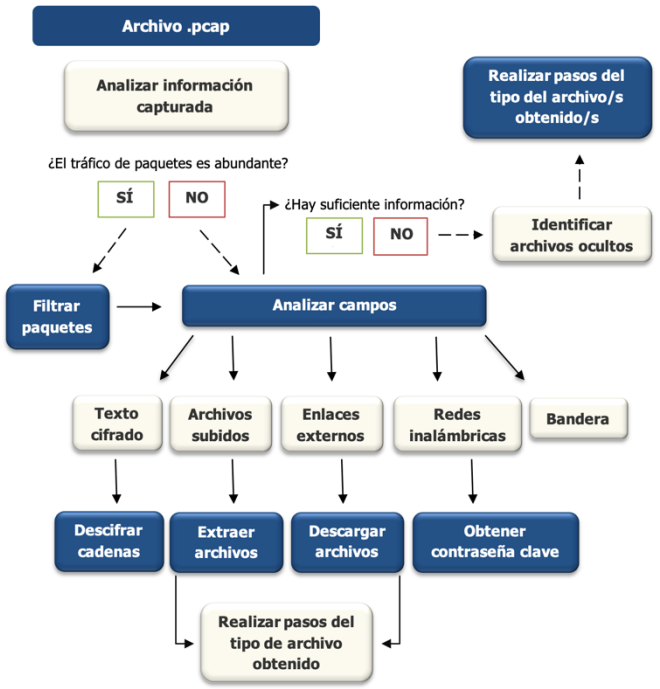


Fig. 6. Esquema de resolución para ficheros de capturas de paquetes.

Ocasionalmente, los desafíos forenses facilitan una imagen de disco completa. Por tanto, lo primero que se debe hacer en esta situación es montar la imagen de disco en un ordenador.

En estos casos, una de las acciones más frecuente es tener que recuperar archivos borrados recientemente tanto de forma automática como de forma intencionada, como muestra Fig. 7.

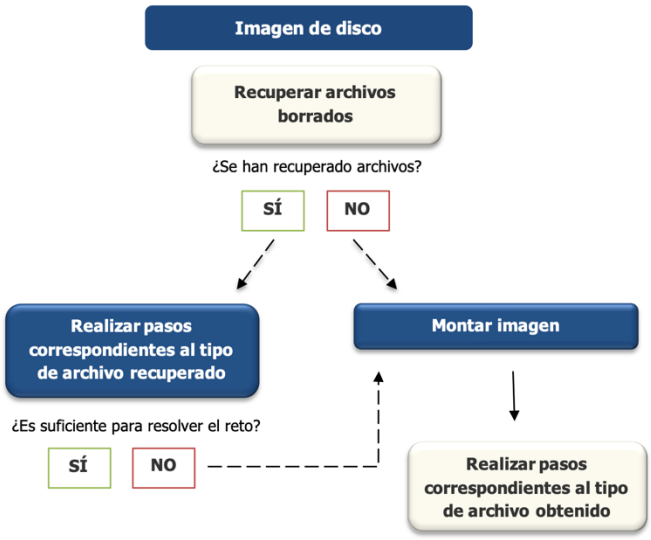


Fig. 7. Esquema de resolución de los retos de análisis forense que contienen imágenes de disco.

Las instantáneas de la memoria contienen información que solo existen en tiempo de ejecución (configuraciones operativas, código de *shell* de explotación remota, contraseñas y claves de cifrado, etc.). Por lo tanto, el análisis forense de volcado de memoria se ha convertido en una práctica popular en la respuesta a incidentes. En un CTF, es posible encontrar desafíos en los que se proporcionen una imagen de volcado de memoria y le asigne tareas para localizar y extraer un secreto o un archivo desde su interior.

La Fig. 8 muestra gráficamente el proceso de resolución perteneciente a los retos de la disciplina de análisis forense que contienen volcados de memoria.

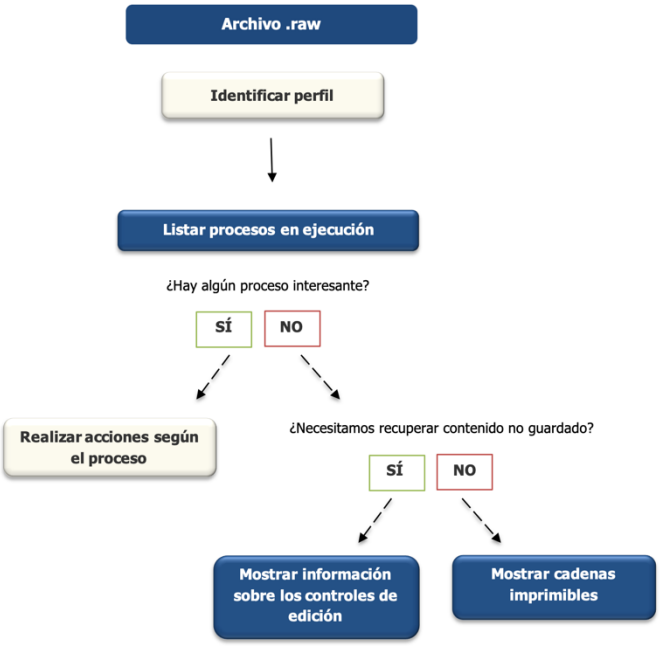


Fig. 8. Esquema de resolución de los retos de análisis forense que contienen volcados de memoria.

C. *Hacking web*

En esta disciplina se han seleccionado para la investigación un total de 50 retos. Las vulnerabilidades de aplicaciones web aparecen en los CTF como desafíos de seguridad donde el usuario necesita explotar un error para obtener algún tipo de privilegio de nivel superior.

Cuando aparecen retos de *hacking web* lo más probable es que, al acceder al dominio correspondiente, se encuentre una página web principal o con una página de inicio de sesión (Login Page).

Una página de inicio de sesión es una página de entrada a un sitio web que requiere identificación y autenticación del usuario, realizada regularmente ingresando una combinación de nombre de usuario y contraseña. El inicio de sesión proporciona acceso al sitio para el usuario y permite que el sitio web rastree las acciones y el comportamiento del usuario. Por lo tanto, se deben explorar los campos relacionados con la información que nos interesa, en este caso los campos relacionados con '*user*' o '*password*'.

La Fig. 9 muestra gráficamente el proceso de resolución perteneciente a los retos de la disciplina de hacking web que contienen páginas de inicio de sesión.

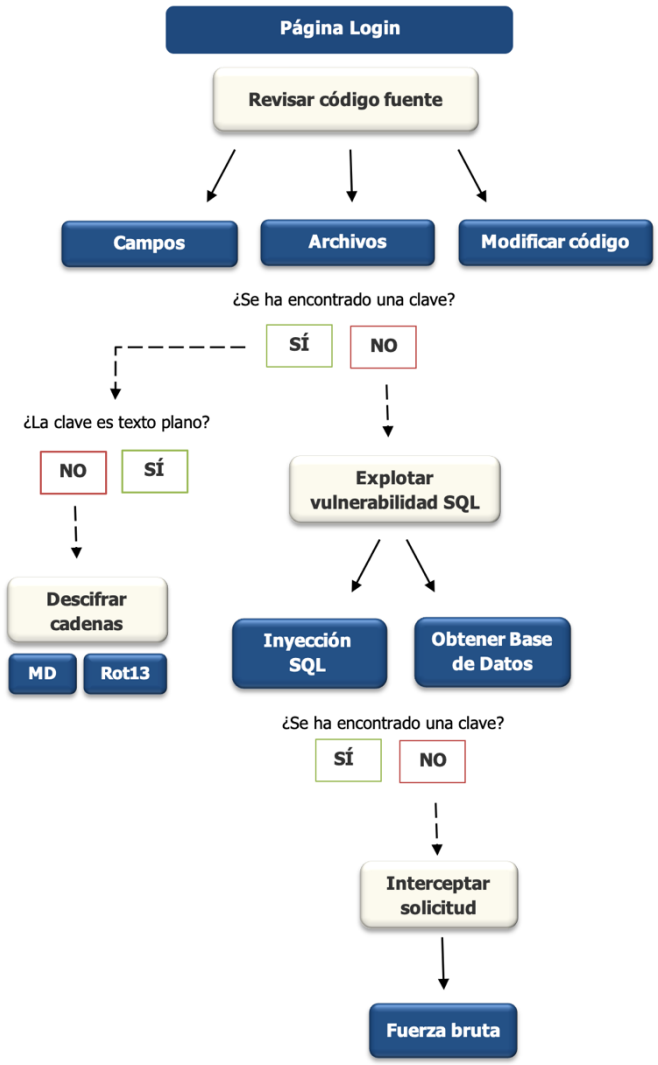


Fig. 9. Esquema de resolución de los retos de hacking web que contienen páginas login.

La página principal de un sitio web se trata de la página predeterminada que se muestra en él cuando un visitante solicita el sitio. El archivo index.html es el nombre más común utilizado y es el archivo que se selecciona por defecto si no se especifica otra página.

La Fig. 10 muestra gráficamente el proceso de resolución perteneciente a los retos de la disciplina de hacking web que contienen páginas principales.

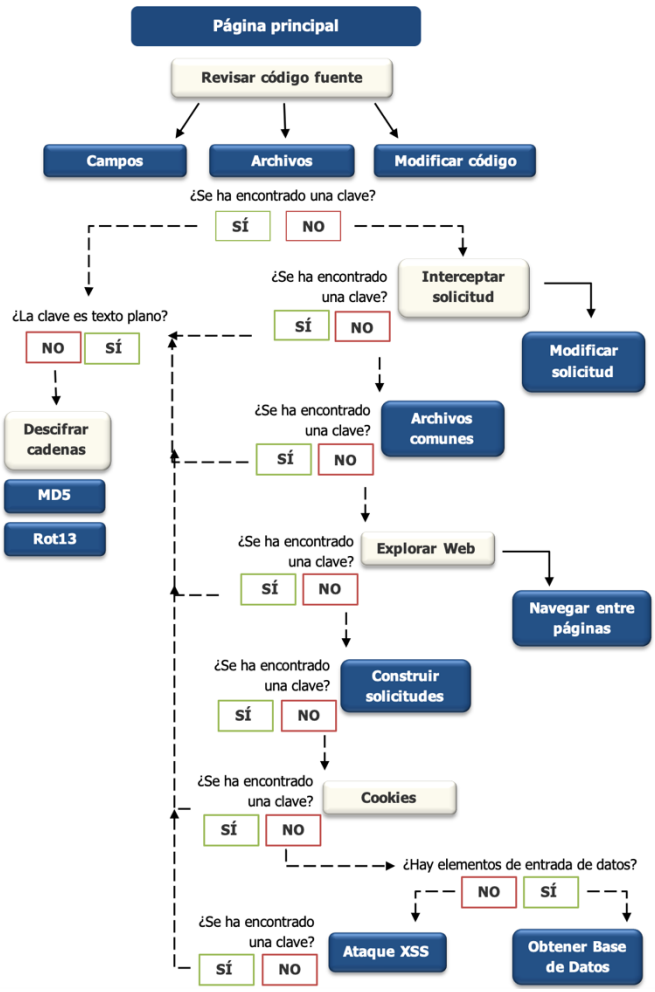


Fig. 10. Esquema de resolución de los retos de hacking web que contienen páginas principales

VII. Conclusiones

Tras analizar en profundidad las acciones de resolución de 125 retos de pruebas CTF correspondientes a las disciplinas de esteganografía, análisis forense y hacking web, se evidencia que existe para cada disciplina por separado una correlación entre las acciones de resolución de los retos. Incluso, como es en los ficheros de tipo de imagen se encuentran relacionados aún siendo de disciplinas distintas.

Dicha relación ha permitido la elaboración de una guía de pasos que se ha mostrado a modo de esquemas y se puede utilizar como apoyo para iniciarse, por un lado, en la participación de competiciones del tipo CTP y, por otro, en el aprendizaje en ciberseguridad basado en retos.

Referencias

[1] Centro Criptológico Nacional, "Ciberamenazas y Tendencias Edición 2017," 2017. [Online]. Available: https://www.ccn-cert.cni.es/informes/informes-ccn-cert-publicos/2221-ccn-cert-ia-16-17-ciberamenzas-y-tendencias-edicion-2017-resumen-ejecutivo-1/file.html.

[2] Centro Criptologico Nacional, "Ciberamenazas y Tendencias Edición 2020," 2020. [Online]. Available: https://www.ccn-cert.cni.es/informes/informes-ccn-cert-publicos/5377-ccn-cert-ia-13-20-ciberamenazas-y-tendencias-edicion-2020/file.html.

[3] Centro Criptologico Nacional, "Informe de Ciberamenazas y Tendencias 2018," 2018. [Online]. Available: https://www.ccn-cert.cni.es/informes/informes-ccn-cert-publicos/2835-ccn-cert-ia-09-18-ciberamenzas-y-tendencias-edicion-2018-1/file.html.

[4] Centro Criptológico Nacional, "Informe de Ciberamenazas y Tendencias 2019," 2019. [Online]. Available: https://www.ccn-cert.cni.es/informes/informes-ccn-cert-publicos/3776-ccn-cert-ia-13-19-ciberamenazas-y-tendencias-edicion-2019-1/file.html.

[5] M. T. Conejo Sequedo, “Trabajo Fin de Grado - Planteamiento de un CTF para practicar hacking ético e informática forense,” 2019.

[6] R. E. Maguiña Becerra, “Trabajo Fin de Grado - Planteamiento de un CTF para practicar hacking ético e informática forense,” 2019.

[7] SAUDI AND OMAN NCS, “Try to See me,” 2019. https://ctftime.org/writeup/13237 (accessed Feb. 27, 2020).

[8] CSAW-CTF, “Keep Calm and CTF,” 2015. https://github.com/Alpackers/CTF-Writeups/blob/master/2015/CSAW-CTF/Forensics/Keep-Calm-and-CTF/Readme.md (accessed Feb. 27, 2020).

[9] RITSEC, “Take it to the Cleaners,” 2019. https://abraxas.io/2019/11/20/ritsec-ctf-2019/#FORENSICS (accessed Feb. 27, 2020).

[10] PEACTF, “The wonderful wizard,” 2019. https://github.com/SinHackTeam/writeup/blob/master/peaCTF2019/Round 1/Forensics/The_wonderful_wizard-volken-writeup.pdf (accessed Feb. 27, 2020).

[11] NEVERLAN, “Open Backpack,” 2020. https://github.com/SinHackTeam/writeup/blob/master/NeverLAN CTF 2020/Forensics/OpenBackpack-volken-writeup.pdf (accessed Feb. 27, 2020).

[12] PEACTF, “We are E.xtr,” 2019. https://hackmd.io/@FlsYpINbRKixPQQVbh98kw/HkpJ6sEfr (accessed Feb. 28, 2020).

[13] STARCTF, “Doc. Holmes,” 2020. https://hackmd.io/@o7feX9hfREaSegDAEu2cAw/H17fC_FmI (accessed Feb. 28, 2020).

[14] PRAGYAN, “Up can be Down,” 2020. https://github.com/SinHackTeam/writeup/blob/master/PragyanCTF2020/Forensics/Up_can_be_Down-volken-writeup.pdf (accessed Feb. 28, 2020).

[15] TJCTF, “Mind blown,” 2019. https://ctftime.org/writeup/14660 (accessed Feb. 28, 2020).

[16] PRAGYAN, “Welcome,” 2019. https://ethan-world.tistory.com/entry/Pragyan-CTF-19-Write-up-Welcome?category=702270 (accessed Feb. 28, 2020).

[17] CSAW-CTF, “Flash,” 2015. https://github.com/Alpackers/CTF-Writeups/blob/master/2015/CSAW-CTF/Forensics/Flash/Flash option 4.md (accessed Feb. 28, 2020).

[18] XMASCTF, “Santa’s Forensics,” 2019. https://khroot.com/2019/12/21/x-mas-ctf-2019-write-ups/ (accessed Feb. 28, 2020).

[19] HACKERS4FUN, “W3llth3nn3v3rm1nd,” 2018. https://elmalodebatman.blogspot.com/2018/06/writeup-reto-h4f-forense.html (accessed Feb. 29, 2020).

[20] RITSEC, “Burn the candle on both ends,” 2018. https://medium.com/@thereallulz/writeup-ritsec-ctf-2018-forensics-by-thereallulz-eb06196e7ae1 (accessed Feb. 29, 2020).

[21] STEMCTF, “Forensics 100,” 2017. https://tobloef.com/ctf/mitre-ctf-2017 (accessed Feb. 29, 2020).

[22] PRAGYAN, “Pretty Peculiar Pokemon,” 2020. https://github.com/SinHackTeam/writeup/blob/master/PragyanCTF2020/Forensics/Pretty_Peculiar_Pokemon-volken-writeup.pdf (accessed Feb. 29, 2020).

[23] CYBERCAMP, “Free Wifi,” 2019. https://ironhackers.es/en/writeups/cybercamp-2019-free-wifi-forense/ (accessed Feb. 29, 2020).

[24] M. López Guerra, “Un día como analista forense,” 2018. https://www.hackplayers.com/2018/03/writeups-ctf-forociber2018.html (accessed Mar. 01, 2020).

[25] ICECTF, “Hardshells,” 2018. https://medium.com/@johnhammond010/icectf-2018-writeups-32df8e53facd (accessed Mar. 01, 2020).

[26] FIC, “5 Forensic,” 2018. https://malware.news/t/fic-2018-random-forensics-challenges-write-up/17730 (accessed Mar. 01, 2020).

[27] NET-FORCE, “Can you see me?,” 2015. https://resources.infosecinstitute.com/defeating-steganography-solutions-to-net-force-ctf-challenges-using-practical-steganalysis/#gref (accessed Jan. 31, 2020).

[28] NET-FORCE, “Another picture!,” 2015. https://resources.infosecinstitute.com/defeating-steganography-solutions-to-net-force-ctf-challenges-using-practical-steganalysis/#gref (accessed Jan. 31, 2020).

[29] NET-FORCE, “Nice colors eh?,” 2015. https://resources.infosecinstitute.com/defeating-steganography-solutions-to-net-force-ctf-challenges-using-practical-steganalysis/#gref (accessed Jan. 31, 2020).

[30] NET-FORCE, “Just a flag,” 2015. https://resources.infosecinstitute.com/defeating-steganography-solutions-to-net-force-ctf-challenges-using-practical-steganalysis/#gref (accessed Jan. 31, 2020).

[31] ATENEA, “Juegos en guerra,” 2018. https://ironhackers.es/writeups/resolviendo-los-retos-criptografia-y-esteganografia-de-atenea-ccn-cert-2-2/ (accessed Jan. 31, 2020).

[32] HACK THE BOX, “Hackerman,” 2019. https://medium.com/@KamranSaifullah/hackerman-stenography-challenge-solution-76407d50781f (accessed Jan. 31, 2020).

[33] HYPERION GRAY, “Stegonauts,” 2019. https://ctf.themanyhats.club/write-up-hyperion-gray-steganography-challenge/ (accessed Feb. 01, 2020).

[34] HACKPLAYERS, “Acid pirate,” 2017. https://www.hackplayers.com/2017/08/solucion-al-reto-21-acid-pirate.html (accessed Feb. 01, 2020).

[35] ATENEA, “Durin’s Gates,” 2018. https://ironhackers.es/writeups/resolviendo-los-retos-criptografia-y-esteganografia-de-atenea-ccn-cert-2-2/ (accessed Feb. 01, 2020).

[36] HACKERS4FUN, “Inmortals,” 2018. https://elmalodebatman.blogspot.com/2018/05/writeup-reto-10-ctf-hackers4fun-h4f.html (accessed Feb. 01, 2020).

[37] HACKERS4FUN, “Sure you can!,” 2018. https://elmalodebatman.blogspot.com/2018/05/writeup-reto-9-hackers4fun-h4f-la.html (accessed Feb. 01, 2020).

[38] BREAKIN, “Three Thieves,” 2016. https://github.com/objEEdump/breakin/tree/master/three_thieves (accessed Feb. 01, 2020).

[39] HACK THE BOX, “Da Vinci,” 2019. https://medium.com/write-ups-hackthebox/como-resolver-da-vinci-hackthebox-a95189008041 (accessed Feb. 01, 2020).

[40] Desconocido, “Random CTF challenge,” 2018. https://medium.com/@thereallulz/write-up-some-random-ctf-challenges-stego-part-1-be6e0c17fd4e (accessed Feb. 01, 2020).

[41] ICECTF, “Drumbone,” 2018. https://medium.com/@johnhammond010/icectf-2018-writeups-32df8e53facd (accessed Feb. 02, 2020).

[42] TJCTF, “Interference,” 2018. https://ctftime.org/writeup/10630 (accessed Feb. 02, 2020).

[43] WPICTF, “Jay-Peg,” 2018. https://www.da.vidbuchanan.co.uk/blog/WPICTF-2018-jay-peg-writeup.html (accessed Feb. 02, 2020).

[44] ISITDTU CTF QUALS, “Acronym,” 2019. https://volatilevirus.home.blog/2019/07/02/isitdtu-ctf-quals-acronym-writeup/ (accessed Feb. 03, 2020).

[45] ENCRYPTCTF, “Stressed out,” 2019. https://www.abs0lut3pwn4g3.cf/writeups/2019/04/04/encryptctf-stressedout.html (accessed Feb. 03, 2020).

[46] FWHIBBIT, “Elliot knows everything,” 2017. https://blog.kalrong.net/es/2017/06/15/follow-the-white-rabbit-ctf-elliot-knows-everything/ (accessed Feb. 04, 2020).

[47] HACK THE BOX, “The Beatles,” 2020. https://exp1o1t9r.com/2020/01/16/hackthebox-challenge-stego-beatles/ (accessed Feb. 04, 2020).

[48] HACK THE BOX, “Forest,” 2018. https://medium.com/write-ups-hackthebox/como-resolver-forest-hackthebox-3cb130e71f36 (accessed Feb. 04, 2020).

[49] TRY HACK ME, “STEGOsaurus,” 2019. https://blog.tryhackme.com/stegosaurus-writeup/ (accessed Feb. 05, 2020).

[50] HACK THE BOX, “Widescreen,” 2018. https://medium.com/write-ups-hackthebox/como-resolver-widescreen-hackthebox-806a944e1a53 (accessed Feb. 05, 2020).

[51] TRY HACK ME, “Stego Challenge,” 2019. https://www.embeddedhacker.com/2019/09/hacking-walkthrough-basic-steganography/ (accessed Feb. 06, 2020).

[52] ENCRYPTCTF, “Into The Black,” 2019. https://vijeta1.github.io/EncryptCTF-Writeups/#into-the-black (accessed Feb. 06, 2020).

[53] HACK THE BOX, “Hidden in colors,” 2019. https://medium.com/write-ups-hackthebox/cómo-resolver-hidden-in-colors-c6020b4ca333 (accessed Feb. 06, 2020).

http://doi.org/10.18239/jornadas_2021.34.63

Selección de competencias en ciberseguridad para la formación en la industria de defensa

Rafael Estepa
Universidad de Sevilla
ORCID: 0000-0001-8505-1920
rafaestepa@us.es

José María de Fuentes
Univ. Carlos III de Madrid
ORCID: 0000-0002-4023-3197
josemaria.defuentes@uc3m.es

Lorena González-Manzano
Univ. Carlos III de Madrid
ORCID:0000-0002-3490-621X
lorena.gonzalez@uc3m.es

Antonio Estepa
Universidad de Sevilla
ORCID: 0000-0003-1841-3973
aestepa@us.es

Jaime Domínguez
Universidad de Sevilla
ORCID: 0000-0002-0491-7911
jaime@us.es

Daniel Segovia-Vargas
Univ. Carlos III de Madrid
ORCID: 0000-0001-7811-3791
dansevar@ing.uc3m.es

Resumen- **En el contexto de la defensa escasean los profesionales capacitados en ciberseguridad, siendo aconsejable impulsar programas de formación específicos para la industria de este sector. Este es uno de los objetivos del proyecto europeo ASSETS+. En este trabajo presentamos uno de sus resultados preliminares: una lista de competencias que los profesionales de ciberseguridad deberían poseer para satisfacer las necesidades formativas de la industria de defensa europea. Para la realización de este listado, se ha seguido una metodología basada en dos fases. En primer lugar, se identifican aquellas tecnologías de ciberseguridad útiles para el sector de defensa mediante el análisis de múltiples fuentes de datos. En segundo lugar, se realiza una adaptación del listado de competencias en ciberseguridad del NIST (NICE) a aquellas aplicables a los perfiles de trabajo en la industria de la defensa, así como a las tecnologías identificadas en la fase 1. El resultado es una relación de competencias de tipo transversal, técnicas o exclusivas del sector de defensa que debería ser central en el diseño de futuros cursos de formación especializados.**

***Index Terms*- formación ciberseguridad, competencias ciberseguridad, ciberseguridad en defensa**

Tipo de contribución: *Formación innovación*

I. Introducción

Nuevos dominios tecnológicos tales como la inteligencia artificial, robótica o ciberseguridad están revolucionando el tradicional mundo de la defensa. En particular, el ciberespacio resulta de especial interés por constituir (junto a tierra, mar, aire, y espacio) un dominio de batalla donde se llevan a cabo multitud de operaciones en misiones 'no públicas'. Éstas se articulan habitualmente a través de grupos de atacantes patrocinados por Estados, como son las amenazas persistentes avanzadas, o incluso cuerpos especiales del Ejército que operan libremente en el ciberespacio mientras que intentan evitar que los adversarios lo hagan [1].

La importancia de la ciberseguridad dentro de la defensa se ve potenciada por el riesgo que supone la captura de información sensible militar (con la consiguiente pérdida de ventaja frente al adversario), ya se produzca en el ejército o en un subcontratista (denominado Defence Industrial Base - DIB-). Este riesgo se ha incrementado en los últimos años debido a la masiva incorporación de las TIC en productos y servicios militares, introduciendo nuevas vulnerabilidades que pueden ser explotadas por el adversario. No en vano, ciertas amenazas avanzadas (como las APT, por sus siglas en inglés *Advanced Persistent Threats*) ya han sido especialmente dirigidas para atacar este sector. Por todo ello, [2] es deseable que la DIB tenga, al menos, las competencias mínimas necesarias para proteger la información sensible, tal y como establece la norma NIST 800.171 para la DIB de Estados Unidos. Adicionalmente, sería muy deseable que los productos software y hardware producidos por la DIB cumplieran las buenas prácticas y los estándares de ciberseguridad descritos por los reguladores. Por lo tanto, una formación adecuada en este sentido forma parte de las necesidades esenciales de la industria.

La actual panorámica educativa actual adolece de algunos problemas. En primer lugar, el escaso número de trabajadores con las competencias y experiencia necesaria para desarrollar ciertas tareas de ciberseguridad provoca un mercado laboral desequilibrado y comporta inherentes debilidades en la seguridad de las organizaciones (en especial el sector DIB). En segundo lugar, la mayoría de los estudiantes están altamente especializados o fragmentados en 'sub-campos' (p.ej., forense, *hacking*, operaciones, auditoría). Para el acceso a cursos de especialización normalmente se exige como prerrequisito la formación en ingenierías TIC, lo que excluye a la mayoría de la fuerza del mercado laboral. Además, los programas educativos especializados en ciberseguridad orientada a defensa son escasos en la Unión Europea y, en no pocas ocasiones, utilizan tecnologías que no están completamente actualizadas. Actualmente es posible encontrar tres tipos de programas de formación en ciberseguridad: (a) programas de certificaciones internacionales en campos específicos de ciberseguridad (p.ej., ISC2, ISACA), (b) programas de postgrado de orientación más generalista (p.ej., Másteres Universitarios, normalmente como títulos de especialización de un grado en el ámbito TIC), y (c) formación en instituciones públicas de defensa (p.ej., policía, militar, …), donde se cubren algunos aspectos muy específicos de la ciberseguridad (p.ej. el ámbito forense). Animamos a los lectores interesados a profundizar en el diseño de currículos en ciberseguridad a leer el tutorial [3], que revisa las principales aproximaciones en el mundo académico y de la industria.

Por ello, uno de los retos a los que se enfrentan los países es la incorporación al sector DIB de personal técnico formado en estos dominios tecnológicos, en general, y en ciberseguridad, en particular. Esta necesidad ha sido identificada y está siendo actualmente atendida por un proyecto Erasmus+ denominado ASSETS+ [3], en el que se encuentran incursos los autores del presente trabajo.

La definición de los cursos requiere de la selección de aquellas competencias relacionadas con ciberseguridad que permitan afrontar con éxito temas y tecnologías de interés en la industria de defensa. En este artículo se presenta un primer resultado del citado proyecto: un listado de competencias que sirvan de base para la elaboración de futuros programas formativos para profesionales de la defensa.

Organización del artículo. La Sección II introduce los conceptos previos. La Sección III describe la metodología empleada. La Sección IV muestra la lista de competencias identificada. Finalmente, la Sección V describe las conclusiones y líneas de trabajo futuro.

II. Terminología y Marcos de referencia

La palabra *tecnología* se refiere a *métodos, sistemas y dispositivos, resultado de conocimiento científico, utilizados para propósitos prácticos* (diccionario Collins). Guiados por esta definición, en el proceso de diseño de cursos de formación será necesario examinar fuentes de información de distintos dominios (mercado, reguladores, academia/investigación, defensa) con la finalidad de identificar y clasificar tecnologías de ciberseguridad que puedan resultar clave en el dominio de la defensa. Sin embargo, realizar una clasificación o taxonomía en ciberseguridad es un reto debido a la falta de homogeneidad en la terminología empleada en conceptos similares a lo largo del tiempo por parte de diferentes actores. De esta forma, en ambientes académicos se emplean términos ligados a las áreas de formación. Por ejemplo, en [5] se revisan más de 100 artículos educativos en el campo de la ciberseguridad agrupados en las siguientes áreas: desarrollo seguro de software (incluyendo ingeniería inversa), monitorización y seguridad en red, ciberataques, malware, seguridad ofensiva y explotación, aspectos humanos, incluyendo privacidad, ingeniería social, legislación y ética e impacto social, criptografía y autenticación y autorización. Estas áreas podrían suponer una primera clasificación basada en temas que forman parte de los cursos de formación actuales. Sin embargo, en el ámbito de la investigación se emplea otra terminología diferente, basada en áreas y palabras clave de los artículos de investigación, que difieren según la revista y autor. Finalmente, en el ámbito de la industria, se suelen emplear términos y clasificaciones basadas en el uso de tecnologías de ciberseguridad, habitualmente cambiantes en el tiempo. Entendemos que las empresas DIB son quien, en última instancia, deben especificar los requisitos y tareas a realizar por los alumnos que reciban los cursos de formación específicos. Por ello, resulta conveniente ofrecer una taxonomía de las tecnologías de ciberseguridad basada en su aplicación en la industria como paso previo a la selección de aquellas que resulten de interés al mundo de la defensa.

Dado que los organismos de normalización y regulación proporcionan referencias de utilidad internacionalmente aceptadas, tanto en el ámbito de las taxonomías de ciberseguridad como en la definición de competencias y terminología, revisaremos a continuación los principales marcos encontrados.

A. Taxonomías en ciberseguridad

Existen distintos marcos conceptuales que agrupan o clasifican controles o procedimientos de ciberseguridad que permiten cubrir las necesidades de la industria. Uno de estos modelos, especialmente relevante en el ámbito de defensa, es el definido por NIST como Cybersecurity Maturity Model Certification (CMMC) [2]. Su objetivo es que aquellas empresas e instituciones (universidades o centros de investigación) que trabajan para el Ministerio de Defensa de EE.UU (unas 300.000) puedan acreditar un nivel de seguridad que les posibilite tener información sensible como contratos federales o información controlada no clasificada. Así, CMMC permite certificar a empresas en función del nivel de cumplimiento de los controles o buenas prácticas que utilicen. Las prácticas descritas en CMMC se agrupan en 17 dominios de capacidades diferentes (p.e.: *Access control, Incident Response, Awareness and Training, Recovery, Identification and Authentication*, etc.) que en sí podría suponer una posible taxonomía en el campo de la ciberseguridad.

Una posible crítica al modelo del CMMC es que se centra mucho en la protección del control de acceso la información (26 prácticas), auditorías (14 prácticas), respuesta a incidentes (13 prácticas) y proyección de comunicaciones y sistemas (27 prácticas). Por ello, un marco de referencia alternativo podría ser el descrito por NIST en su *Framework for Improving Critical Infrastructure Cybersecurity* (CSF) [6]. CSF define 260 controles o buenas prácticas para infraestructuras críticas agrupados en una taxonomía de 23 categorías que pertenecen a una de las 5 funcionalidades básicas de la ciberseguridad: identificar, proteger, detectar, respuesta y recuperación. En cualquier caso, ambas normas tienen una serie de limitaciones para nuestro trabajo: (I) La aplicación de dominios de seguridad (17 en CMMC) o categorías (23 en CSF) resulta difusa para la definición de tecnologías, mientras que el uso de dos niveles (prácticas en CMMC o subcategorías en CSF) ofrece un nivel de granularidad excesivo (entre 72-172 con CMMC y 260 en CSF). (II) Se basan en buenas prácticas bien conocidas para bastionar o reforzar sistemas, pero de manera generalista y no centrada en la industria de defensa.

B. Marcos de referencia en educación en ciberseguridad

Hay dos grandes marcos de referencia internacionales en formación para ciberseguridad: el ACM *cybersecurity Curricula Guideline* [7] y el propuesto por el NIST en 2020 denominado *National Initiative for Cybersecurity Education* (NICE) [7]. El primero está pensado para el diseño de Grados en ciberseguridad dentro del mundo académico, y divide los contenidos en 8 grandes áreas de conocimiento sobre las que define unidades de conocimiento, tópicos y resultados del aprendizaje. El segundo, sin embargo, define las competencias y conocimientos a adquirir por un alumno para realizar distintas tareas en el campo de la ciberseguridad, lo que facilita el diseño de cursos en función de las tareas que se

buscan para un perfil de trabajo determinado. Dado nuestro enfoque a la industria, y la facilidad que ofrece NICE para trasladar en áreas de conocimiento de otros marcos como el de ACM, adoptaremos NICE como marco de referencia. De dicho marco tomaremos las siguientes definiciones:

- Tarea (*Task*): actividad dirigida a conseguir los objetivos de la organización. La definición de la tarea debe realizarse en el lenguaje corporativo y debe incluir el trabajo que debe ser completado (ej. resolución de problemas de hardware). En la definición no se debe incluir el objetivo, sino la tarea.
- Competencia (*Skill*): es la capacidad para realizar una acción observable. En relación con una tarea, la competencia es la demostración de la pericia para realizarla (ej. reconocer alertas de un sistema de detección de intrusiones). La descripción de una competencia debería incluir qué puede hacer una persona gracias a ella. Hay que señalar que las competencias pueden ser específicas de ciberseguridad o transversales.
- Conocimiento (*Knowledge*) un conjunto de conceptos dentro de la memoria que pueden ser recuperados (ej. conocimiento de fuentes de diseminación de información sobre vulnerabilidades).

El marco de referencia NICE permite asociar competencias a tareas concretas que forman la base de perfiles de trabajo. El uso de perfiles de trabajo facilita a los empleadores la definición de puestos demandados y permite extraer las competencias y conocimientos a incluir en los cursos de formación.

Para nuestro propósito, el nivel de granularidad de NICE resulta muy elevado, pues define 377 competencias (algunas transversales) asociadas a 54 perfiles de trabajo diferentes en ciberseguridad pertenecientes a 32 áreas de especialidad. No obstante, el ámbito de la defensa tiene sus propios roles de trabajo que no están incluidos, por lo que no puede utilizarse este marco directamente para satisfacer nuestro objetivo, pero sí podemos aprovechar las listas de competencias y conocimientos asociados a una tarea concreta.

Aunque en Europa existe una clasificación genérica de competencias, cualificaciones y ocupaciones (ESCO [13]), su empleo en la ciberseguridad resultaría muy difícil ya que se centra básicamente en competencias TIC y transversales.

III. METODOLOGÍA EMPLEADA

Para confeccionar una lista de competencias en ciberseguridad necesarias para el desarrollo de la labor de la DIB se propone una metodología en dos fases secuenciales (Ilustración 1):

1. Identificar una lista de tecnologías de interés del DIB. Para ello, partiendo de una búsqueda documental, se elaborará una taxonomía de tecnologías de ciberseguridad, así como una lista de aplicaciones en defensa de la Ciberseguridad. La relación entre ambas listas y la realización de encuestas a empresas DIB ofrecerán el resultado final de las tecnologías en ciberseguridad de mayor interés en el ámbito de la defensa. Se fija como objetivo un nivel de granularidad intermedio, de entre 20 y 30 tecnologías.
2. Partiendo del resultado anterior, se buscará un conjunto de competencias dentro del marco de NICE relacionadas con dichas tecnologías. El resultado final será un conjunto de competencias agrupadas en: (I) competencias técnicas en ciberseguridad, (II) competencias transversales, aplicables a cualquier campo y (III) competencias en técnicas con potencial ofensivo (defensa) en ciberseguridad.

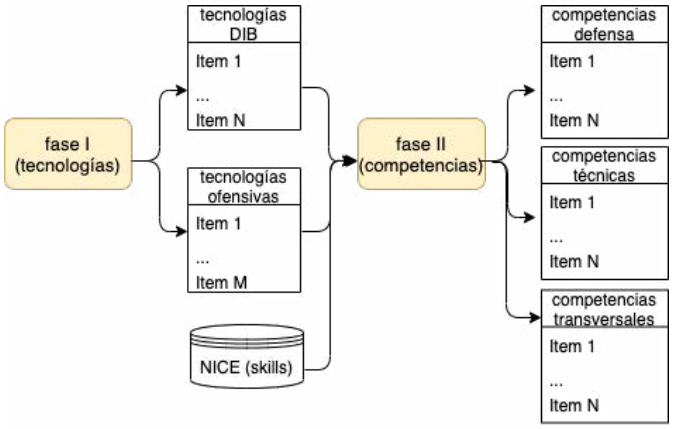


Ilustración 1- Etapas y resultados en la metodología

A continuación, se describe la metodología seguida en cada una de las dos fases anteriores.

A. Fase 1: Identificación de Tecnologías en ciberseguridad de interés en defensa

Las actividades desarrolladas en esta primera fase se descomponen en los siguientes pasos reflejados en la Ilustración 2:

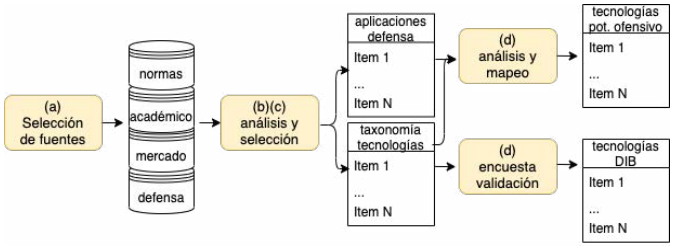


Ilustración 2. Pasos de la fase 1

a) Selección de fuentes de información: se identificaron fuentes para clasificar tecnologías desde distintas perspectivas: industria/mercado, academia/investigación, reguladores, y defensa. La metodología seguida en cada tipo de fuente y los resultados más relevantes han sido:

- Industria y Mercado: búsqueda en Google de los términos: *"market", "cybersecurity technology", "cybersecurity technologies", "cybersecurity technologies classification", "cybersecurity market"*.
- Reguladores: Se han buscado las publicaciones del NIST (principal actor a nivel internacional en ciberseguridad) y en otros organismos similares internacionales como MITRE (en concreto las técnicas y tácticas de la matriz ATT&CK).
- Academia / Investigación: búsquedas en Scopus y Google Scholar con los términos: *"cybersecurity technologies", "cybersecurity classification", "cybertechnologies", "cyber-technologies", "cybersecurity trends", "cybersecurity education"*. También se incluyeron búsquedas de las contribuciones en el *track* de ciberseguridad de la conferencia MILCOM 2018-2019.

• Defensa / Militar: búsquedas en Google, Google Scholar y Scopus sobre los términos: *"cyberwar", "cyberspace", "cyber-warfare", "cyberdefense", "cyber operations", "cybersecurity military", "cybercommand"*. Igualmente se consideró la revista "*Cyberdefense review*" editada por la academia de West Point.

b) Definición de una taxonomía: Tomando como clasificación de partida la única encontrada en el informe sobre ciberseguridad en 18 países europeos [9], se mezcló con la ofrecida por [10] realizada tras el análisis de los productos y servicios de 3.500 empresas de ciberseguridad internacionales. Luego se utilizaron las clasificaciones extraídas de los marcos descritos en la Sección II, de las que se excluyeron las tecnologías ya presentes y se unificaron por analogía de conceptos. Tras ello se utilizaron las fuentes de la academia y defensa para incorporar aquellas tecnologías que no estaban presentes. La clasificación de tecnologías anterior ha sido enlazada con el marco CSF, y a su vez se han buscado empresas europeas que las utilicen gracias a ECSO *market radar.*

c) Búsqueda de aplicaciones de ciberseguridad para defensa. Todos los documentos recopilados relacionados con defensa y Ejército, fueron transformados de PDF a texto plano utilizando la librería *poppler* a fin de automatizar las búsquedas de artículos y párrafos que tuvieran los siguientes términos: *"cyber" and "defense", "war", "warfare", "applications", "military", "technique", "mission", "operations"*. Los párrafos con frases encontradas fueron analizados a fin de encontrar las distintas aplicaciones. Se creó una lista inicial con las aplicaciones encontradas y de ella se filtraron sólo las entradas más referenciadas. A la lista final se añadió una aplicación adicional relacionada con la protección de información sensible en manos de las empresas que forman la DIB.

d) Selección de las tecnologías más apropiadas para la defensa. Para ello se han realizado los siguientes pasos:
- Evaluación de la utilidad de cada una de las tecnologías en cada una de las aplicaciones en defensa, puntuando de 0 a 10 a juicio de experto.
- Validación mediante una encuesta realizada a industrias de defensa españolas del consorcio ASSETs+. En concreto, se ha preguntado: (I) la relevancia de la tecnología para su empresa y para el ámbito de la defensa, (II) la facilidad para encontrar expertos en dicha tecnología y (III) el grado de madurez de la empresa en dicha tecnología. Las posibles respuestas iban de 1 (poco) a 3 puntos (mucho)
- Filtrado de las tecnologías: se han considerado relevantes las tecnologías que cumplan con las siguientes restricciones: (a) que exista alguna empresa europea con productos, (b) que la facilidad para encontrar expertos sea inferior a 2,3 puntos sobre 3 y (c) que sea relevante para la empresa (más de 1,5 puntos sobre 3) o bien que sea aplicable al 50% o más de las aplicaciones encontradas.

B. Selección de competencias

Esta parte se selecciona un subconjunto de competencias de entre las 377 definidas en NICE. Con respecto a la granularidad, se fija como objetivo conseguir una lista final con no más de 40 competencias.

La metodología seguida se representa en la Ilustración 3.

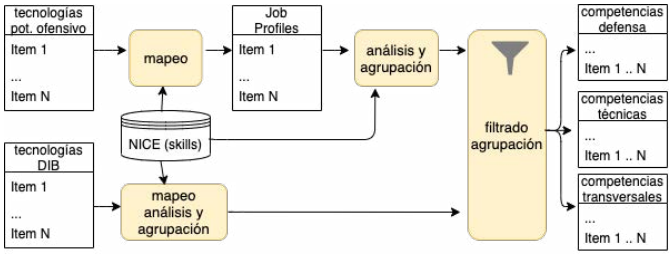


Ilustración 3. Pasos en la fase 2

A continuación, se describen los pasos principales seguidos:

a) Se han relacionado las competencias NICE con su aplicación a las tecnologías con propósito defensivo seleccionadas en el paso anterior. Se eliminaron duplicados y se filtraron las competencias transversales.

b) Se reduce la granularidad de forma iterativa, uniendo competencias similares y seleccionando las competencias con mayor prevalencia (asociadas a más tecnologías). Tras ello se verifica la coherencia de la selección (debe haber al menos una competencia representativa por cada tecnología) y se realiza una nueva iteración, hasta llegar a la lista final de competencias técnicas.

c) Las competencias transversales han sido agrupadas por similitud de conceptos y prevalencia en un proceso similar al anterior. El resultado final ha sido completado con 3 competencias propuestas por el Foro económico mundial (*World Economic Forum)*.

d) Las tecnologías con potencial ofensivo han sido relacionadas con los perfiles de trabajo que presentan mayor afinidad. Para ello, se extraen las competencias asociadas, filtrando las transversales, los duplicados y las ya presentes en las listas anteriores. Las competencias resultantes han sido reducidas en un proceso iterativo similar al descrito para las competencias técnicas hasta llegar a la lista final.

IV. RESULTADOS

En esta sección se presentarán los resultados de aplicar la metodología descrita. A fin de mantener la información lo más fielmente posible a las referencias utilizadas, los resultados procedentes de NICE se presentarán en inglés. Los resultados de la búsqueda de información fueron un total de 135 artículos (50% del campo defensa/Militar y 50% de Academia /investigación), 2 libros y 3 informes de mercado. Estos documentos forman el corpus que sirve de base para realizar los análisis del presente trabajo. Tan sólo en el informe de mercado [10] encontramos una clasificación de tecnologías en ciberseguridad, que, junto con [11] sirvieron de punto de partida.

En la tabla I se especifica la clasificación final realizada para las tecnologías de ciberseguridad, donde en torno al 60% han sido obtenidas de las fuentes de industria/mercado, el

30% conforme a los marcos CMMC y CSF y el 10% de defensa e investigación.

Tabla I
LISTA DE LAS TECNOLOGÍAS EN CIBERSEGURIDAD

ID	Tecnología	Descripción
1	Identificación & Autenticación	Asegurar que una característica declarada por una entidad es correcta. [12]
2	Autorización & Control de acceso	La concesión de derechos, que incluye la concesión de acceso en función de los derechos de acceso. [13]
3	Cortafuegos	Conjunto de técnicas para proteger activamente una red, como cortafuegos o protecciones DDoS
4	*Cyber range*	Un conjunto de activos y capacidades que se pueden integrar en los niveles de clasificación y controles apropiados para realizar investigación, desarrollo, demostración, prueba o evaluación de capacidades militares apoyadas en o para capacitar al personal militar durante las operaciones. [14]
5	Cifrado	Cifrado de datos dentro o en el sistema final de origen, y el descifrado correspondiente se produce dentro o en el sistema final de destino. [15]
6	Certificación	Procesos para certificar la identidad o un atributo de un sistema o servicio.
7	Seguridad en *endpoint*	Los sistemas de seguridad *endpoint* protegen las computadoras y otros dispositivos en una red o en la nube de las amenazas de ciberseguridad. La seguridad de los *endpoints* ha evolucionado desde el software antivirus tradicional hasta proporcionar una protección integral contra malware sofisticado y amenazas de día cero en constante evolución. [16]
8	Seguridad móvil & IoT	Conjunto de técnicas para proteger los dispositivos por pocos recursos.
9	Seguridad en la nube y virtualización	Conjunto de técnicas relacionadas con la protección del entorno de la nube y los entornos virtualizados.
10	Gestión de vulnerabilidades	Proceso de identificación, evaluación y corrección de vulnerabilidades, definido como "Debilidad en un sistema de información, procedimientos de seguridad del sistema, controles internos o implementación que podrían ser explotados o desencadenados por una fuente de amenaza". [17]
11	Análisis de malware	Adquirir datos confidenciales desensamblando y analizando el diseño de un componente del sistema. [18]
12	Seguridad hardware	Conjunto de protecciones físicas aplicadas para asegurar la correcta ejecución y funcionamiento de un dispositivo.
13	Intrusion Prevention/Detection Systems (IDS/IPS)	Productos de hardware o software que recopilan y analizan información de diversas áreas dentro de una computadora o una red para identificar posibles brechas de seguridad, que incluyen tanto intrusiones como mal uso. [19]
14	Inteligencia de ciberamenazas	Actividades que utilizan todas las fuentes de "inteligencia" en apoyo de la ciberseguridad para trazar las ciberamenazas, recopilar las intenciones de ataque y las posibilidades de los adversarios potenciales, para analizar y comunicar, e identificar, localizar y asignar la fuente de los ciberataques.. [23]
15	*Honeypots*	Un sistema (por ejemplo, un servidor web) o un recurso del sistema (por ejemplo, un archivo en un servidor) que está diseñado para ser atractivo para los posibles atacantes e intrusos, como la miel es atractiva para los osos. [43]
16	Operaciones de seguridad	Conjunto de tecnologías disponibles en un Centro de operaciones de seguridad, que según McAfee, "es una función centralizada dentro de una organización que emplea personas, procesos y tecnología para monitorear y mejorar continuamente la postura de seguridad de una organización mientras previene, detecta, analiza y responde a incidentes de ciberseguridad ". [25]
17	Evaluación & Gestión de riesgo	El proceso de identificar, evaluar y responder al riesgo. [24]
18	Aseguramiento y cumplimiento	Cumplimiento con, y capacidad para demostrar el cumplimiento a los requisitos obligatorios definidos por las leyes y reglamentos, así como a los requisitos voluntarios resultantes de las obligaciones contractuales y las políticas internas. [25]
19	Desarrollo seguro de software	Proceso de desarrollo de software que considera los problemas relacionados con la seguridad desde el primer paso.
20	Análisis forense	La práctica de recopilar, retener y analizar datos relacionados con el ordenador con fines de investigación de una manera que mantenga la integridad de los datos.. [26]
21	Protección DoS	Implementaciones para protegerse contra ataques de denegación de servicio.
22	Ciber resiliencia	El proceso para asegurar que la recuperación de las operaciones esté asegurada en caso de que ocurra algún incidente inesperado o no deseado que sea capaz de afectar negativamente la continuidad de las funciones comerciales esenciales y los elementos de apoyo. [27]
23	Inteligencia de fuentes abiertas (OSINT)/ Inteligencia privada (PRIVINT)	Inteligencia a partir de información pública / privada que se recopila, explota y se informa para abordar un requisito de inteligencia específico. [27]
24	Test de penetración / *Red Teaming*	Conjunto de herramientas y técnicas enfocadas en atacar un dispositivo, servicio o host para identificar debilidades y protegerlo posteriormente.

Cada tecnología ha sido clasificada de acuerdo con el propósito o finalidad de la misma conforme al marco CSF. Los resultados se muestran en la tabla II.

Tabla II
AGRUPACIÓN DE LAS TECNOLOGÍAS SEGÚN SU FINALIDAD

Propósito		ID
Defensivo	Proteger	1 -12
	Detectar	13 - 16
	Identificar	17 – 19
	Responder	20 – 21
	Recuperar	22
Ofensivo		23 - 24

Así mismo, se ha buscado en *ECSO market radar* el número de empresas de la Unión Europea que presentan productos o servicios relacionados con dicha tecnología. Los resultados se muestran en la columna "Empresas EU" de la tabla III.

A. Selección de aplicaciones

Continuando con la metodología presentada, se han identificado las siguientes aplicaciones de ciberseguridad para la defensa:

- A1: Colaboración de la Base Industrial de Defensa (DIB): DIB proporciona una mayor seguridad para el Departamento de defensa. Un contratista de DIB puede proteger los contratos y controlar la información a un nivel acorde con el riesgo, contabilizando los flujos de información hacia sus subcontratistas en una cadena de suministro de varios niveles [2].
- A2: Concienciación y formación en ciberseguridad: se corresponde con el uso de las plataformas de formación para formar al eslabón más débil de la cadena de ciberseguridad, el usuario; proporcionando guías para una buena conducta y mejorando el uso de la información por parte del personal militar. Así, se persigue aumentar la eficiencia operacional.
- A3: Seguridad de las operaciones (OPSEC): es el proceso de ayudar en la identificación de acciones que puedan ser observadas y recogidas por los adversarios. También persigue determinar los indicadores que los adversarios podrían obtener e interpretar para adquirir información crítica y, si fuese apropiado, seleccionar y ejecutar medidas OPSEC que eliminen y reduzcan el riesgo a niveles aceptables [20].
- A4: Operaciones ciber (CO): se corresponde con las capacidades del ciberespacio cuyo propósito es la de conseguir los objetivos en o a través del ciberespacio. Se pueden clasificar en Operaciones Ciber (OCO) y Operaciones Defensivas (DCO) [21].
- A5: Operaciones de apoyo a la información militar (MISO): son las acciones especialmente relacionadas con el uso de capacidades asociadas a la información del ciberespacio durante las operaciones militares, para influir, interrumpir, corromper o usurpar la toma de decisiones de adversarios y adversarios potenciales.
- A6: Soporte de comando y control: son las decisiones que ofrecen soporte de comando y control (inteligencia, vigilancia, adquisición de objetivos, reconocimiento).
- A7: Comunicaciones seguras (COMSEC): se utilizan para proteger tráfico clasificado y no clasificado en redes de comunicación militar, incluyendo voz, vídeo y datos [22].
- A8: Actividades de enfrentamiento ciber-electrónicos (CEWA): combinan los enfrentamientos ciber y electrónicos en un mismo contexto para respaldar, habilitar, proteger y recopilar las capacidades que operan dentro del espectro electromagnético (EMS), incluidas las capacidades del ciberespacio.

Finalmente se ha procedido con el filtrado de las tecnologías con mayor interés para el ámbito de la defensa conforme se estableció en la metodología para la Fase 1 apartado d). Las 19 tecnologías seleccionadas (color negro en la tabla III) cumplen las restricciones de filtrado y siempre hay al menos una tecnología por cada clase del CSF.

Tabla III
MÉTRICAS DE LAS TECNOLOGÍAS

ID Tecnología	Empresas EU	Aplicaciones defensa	Encuesta DIB	
			Mercado Laboral	Relevancia Empresa
1	8	7	1,6	1,8
2	8	7	1,6	1,8
3	8	5	2,6	2
4	9	4	1,8	2,2
5	10	7	1,6	1,8
6	5	4	1,6	1,8
7	9	5	2,2	1,6
8	12	2	2,4	2,2
9	8	2	2	1
10	15	4	1,6	1,8
11	5	3	1,8	1,2
12	12	3	1,8	1,8
13	14	5	1,8	1,6
14	17	7	1,4	1,6
15	6	2	1,4	0,8
16	14	7	1,6	1,4
17	10	4	1,8	2
18	13	4	2,4	2,2
19	9	3	1,8	2,2
20	14	2	1,4	1,6
21	6	5	2	1,6
22	9	7	1,6	1,6
23	1	7	1,4	1,4
24	4	4	1,2	1,2

B. Selección de competencias

La vinculación de dichas tecnologías con las competencias de NICE ofreció un promedio de 11 competencias por tecnología. Tras el proceso de reducción de granularidad y filtrado, en la tabla IV se muestra el listado de competencias técnicas (indicando en cada una el código y descripción de dicha competencia según NICE).

Tabla IV
LISTADO DE COMPETENCIAS TÉCNICAS

Código	*Descripción*
S0367	Skill to apply cybersecurity and privacy principles to organizational requirements (relevant to confidentiality, integrity, availability, authentication, non-repudiation).
S0357	Skill to anticipate new security threats.
S0124	Skill in troubleshooting and diagnosing cyber defense infrastructure anomalies and work through resolution.
S0371	Skill to respond and take local actions in response to threat sharing alerts from service providers.

Código	*Descripción*
S0077	Skill in securing network communications.
S0178	Skill in analyzing essential network data (e.g., router configuration files, routing protocols), network traffic capacity and performance characteristics.
S0185	Skill in applying analytical methods typically employed to support planning and to justify recommended strategies and courses of action.
S0221	Skill in extracting information from packet captures.
S0006	Skill in applying confidentiality, integrity, and availability principles.
S0027	Skill in determining how a security system should work (including its resilience and dependability capabilities) and how changes in conditions, operations, or the environment will affect these outcomes.
S0040	Skill in implementing, maintaining, and improving established network security practices.
S0078	Skill in recognizing and categorizing types of vulnerabilities and associated attacks.
S0096	Skill in reading and interpreting signatures (e.g., snort).
S0156	Skill in performing packet-level analysis.
S0173	Skill in using security event correlation tools.
S0202	Skill in data mining techniques (e.g., searching file systems) and analysis.
S0258	Skill in recognizing and interpreting malicious network activity in traffic.
S0269	Skill in researching vulnerabilities and exploits utilized in traffic.
S0288	Skill in using multiple analytic tools, databases, and techniques (e.g., Analyst's Notebook, A-Space, Anchory, M3, divergent/convergent thinking, link charts, matrices, etc.).
S0007	Skill in applying host/network access controls (e.g., access control list).
S0010	Skill in conducting capabilities and requirements analysis.
S0015	Skill in conducting test event and secure test plan design (e. g. unit, integration, system, acceptance)
S0018	Skill in creating policies that reflect system security objectives.
S0020	Skill in developing and deploying signatures.
S0031	Skill in developing and applying security system access controls.
S0032	Skill in developing, testing, and implementing network infrastructure contingency and recovery plans.
S0036	Skill in evaluating the adequacy of security designs.
S0063	Skill in collecting data from a variety of cyber defense resources.
S0084	Skill in configuring and utilizing network protection components (e.g., Firewalls, VPNs, network intrusion detection systems).
S0087	Skill in deep analysis of captured malicious code (e.g., malware forensics).
S0089	Skill in one-way hash functions (e.g., Secure Hash Algorithm [SHA], Message Digest Algorithm [MD5]) and verifying the integrity of all files.
S0093	Skill in interpreting results of debugger to ascertain tactics, techniques, and procedures.
S0120	Skill in reviewing logs to identify evidence of past intrusions.
S0138	Skill in using Public-Key Infrastructure (PKI) encryption and digital signature capabilities into applications (e.g., S/MIME email, SSL traffic).
S0164	Skill in assessing the application of cryptographic standards.
S0195	Skill in conducting research using all available sources (including deep web)
S0197	Skill in conducting social network analysis, buddy list analysis, and/or cookie analysis.
S0270	Skill in reverse engineering (e.g., hex editing, binary packaging utilities, debugging, and strings analysis) to identify function and ownership of remote tools.
S0317	Skill to compare indicators/observables with requirements.
S0064	Skill in developing and executing technical training programs and curricula.

Por su parte, la lista de competencias transversales se muestra la tabla V. Nótese que tres de ellas no tienen identificador en NICE puesto que provienen del Foro económico mundial (señaladas como WEF en dicha tabla).

Tabla V
LISTADO DE COMPETENCIAS TRANSVERSALES

Código	*Descripción*
S0070	Skill in talking to others to convey information effectively
S0356	Skill in communicating with all levels of management including Board members (e.g., interpersonal skills, approachability, effective listening skills, appropriate use of style and language for the audience).
S0344	Skill to prepare and deliver reports, presentations and briefings, to include using visual aids or presentation technology.
S0301	Skill in writing about facts and ideas in a clear, convincing, and organized manner.
S0213	Skill in documenting and communicating complex technical and programmatic information.
S0306	Skill to analyze strategic guidance for issues requiring clarification and/or additional guidance.
S0128	Skill in using manpower and personnel IT systems.
WEF	Skill in Conflict Management
WEF	Skill in Critical Thinking
WEF	Skill in Complex problem solving

Por último, para las competencias en defensa se consideraron tres perfiles profesionales definidos en NICE, a saber: *Exploitation analyst, Cyber Intel Planner, Cyber Ops Planner, Cyber Operator*. De esta forma, la lista de competencias asociadas a defensa se refleja en la tabla VI.

Tabla VI
LISTADO DE COMPETENCIAS DE DEFENSA

Código	*Descripción*
S0182	Skill in analyzing target communications internals and externals collected from wireless LANs.
S0242	Skill in interpreting vulnerability scanner results to identify vulnerabilities.
S0252	Skill in processing collected data for follow-on analysis.
S0255	Skill in providing real-time, actionable geolocation information utilizing target infrastructures.
S0293	Skill in using tools, techniques, and procedures to remotely exploit and establish persistence on a target.
S0295	Skill in using various open-source data collection tools (online trade, DNS, mail, etc.).
S0218	Skill in evaluating information for reliability, validity, and relevance.
S0309	Skill to anticipate key target or threat activities which are likely to prompt a leadership decision.
S0209	Skill in developing and executing comprehensive cyber operations assessment programs for assessing and validating operational performance characteristics.
S0360	Skill to analyze and assess internal and external partner cyber operations capabilities and tools.

V. Conclusiones y líneas de avance

En este trabajo se ha presentado una lista de competencias que puede ser utilizada para la planificación de cursos de ciberseguridad en el ámbito de la defensa. La metodología empleada utiliza una revisión exhaustiva de diversas fuentes documentales y marcos de referencia internacionales. Estos resultados preliminares se corresponden con la ejecución en curso del proyecto europeo ASSETS+, enmarcado en la iniciativa Erasmus+.

Los siguientes pasos consisten en la definición de perfiles de cursos de formación, que serán realizados en consonancia con la industria de defensa. Dichos cursos deberán considerar no sólo las competencias, sino los conocimientos y las dependencias entre ellos.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto ASSETS+ [3] en el ámbito de la iniciativa Erasmus+; por el proyecto CAVTIONS-CM-UC3M, co-financiado por la Comunidad de Madrid (CAM) y la Universidad Carlos III de Madrid; por el MINECO, proyecto ODIO/COW(PID2019-111429RB-C21); por la CAM, proyecto CYNAMON-CM(P2018/TCS-4566), co-financiado con fondos europeos ESF y FEDER; y por el Programa de Excelencia para Investigadores de la Universidad Carlos III de Madrid.

Referencias

[1] Departamento de defensa EE.UU., “Cybersecurity Maturity Model Certification, version 1.02,” Enero 2020. Disponible en: https://www.acq.osd.mil/cmmc/draft.html [Accedido el 5 mayo 2021]

[2] Mouheb, D., Abbas, S., & Merabti, M.. “Cybersecurity curriculum design: A survey” en Transactions on Edutainment XV (pp. 93-107) .Springer, Berlin, Heidelberg. 2020

[3] Proyecto “Alliance for strategic skill addressing emerging technologies in Defense (ASSETs+)”. Disponible en https://assets-plus.eu [Accedido el 5 de mayo 2021]

[4] Švábenský, V., Vykopal, J., & Čeleda, P. “What are cybersecurity education papers about? a systematic literature review of sigcse and iticse conferences” en Proceedings of the 51st ACM Technical Symposium on Computer Science Education (pp. 2-8). 2020

[5] ACM “Cybersecurity curriula guidelines 2017”. Disponible en: https://cybered.hosting.acm.org/wp/wp-content/uploads/2018/02/csec2017_web.pdf [Accedido el 5 de mayo de 2021].

[6] NIST Cybersecurity Framework (CSF). Disponible en: https://www.nist.gov/cyberframework [Accedido el 5 de mayo de 2021]

[7] NIST.SP.181r1. “Workforce Framework for Cybersecurity (NICE Framework)”. Disponible en https://doi.org/10.6028/NIST.SP.800-181r1 2020. [Accedido el 5 de mayo de 2021]

[8] Clasificación europea de capacidades/competencias, cualificaciones y ocupaciones. (ESCO). Disponible en: https://ec.europa.eu/esco/portal/skill [Accedido el 5 de mayo de 2021]

[9] Cybersecurity Technologies & Market - Focus on Europe - 2017-2022. Disponible: https://homelandsecurityresearch.com/reports/cybersecurity-technologies-market-focus-europe/ [Accedido el 5 de mayo de 2021]

[10] Momentum Cybersecurity Group. “Momentum Cyberscape 2021”. Disponible en https://momentumcyber.com/docs/CYBERscape.pdf [Accedido el 5 de mayo 2021].

[11] ISO/IEC.. ISO/IEC 27000: 2014 (E) Information technology—Security techniques—Information security management systems—Overview and vocabulary. 2014.

[12] Reihe, I.. IEC 7498 ISO/IEC 7498-1: 1994-11. Information technology–Open Systems Interconnection–Basic Reference Model: The Basic Model ISO, 7498-2. 1994

[13] Damodaran, S. K., & Smith, K.. CRIS “Cyber Range Lexicon, Version 1.0 (No. MIT-LL-59-0001)”. MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB. 2015

[14] ISO 7498-2:1989. Information processing systems — Open Systems Interconnection — Basic Reference Model. 1989

[15] McAffee, “What Is Endpoint Security? Disponible en https://www.mcafee.com/enterprise/es-es/security-awareness/endpoint.html [Accedido el 5 de mayo de 2021]

[16] NIST, S.. 800-53: 2013. Security and Privacy Controls for Federal Information Systems and Organizations. 2013.

[17] SANS. “Glossary of Security Terms” Disponible en https://www.sans.org/security-resources/glossary-of-terms/ .[Accedido el 5 de mayo de 2021]

[18] US Committee on National Security Systems “National Information Assurance. (IA) glossary”. Tech. Rep. 4009, 2010. Disponible en https://www.dni.gov/files/NCSC/documents/nittf/CNSSI-4009_National_Information_Assurance.pdf [Accedido el 5 de mayo de 2021]

[19] U.S. Department of Energy. Office of Scientific and Technical Information. “Differences Between OPSEC and Security Awareness.” Disponible en https://www.osti.gov/servlets/purl/1367112 .[Accedido el 5 de mayo de 2021].

[20] M. Karamanetal. “Institutional Cybersecurity from Military Perspective“. International Journal of Information Security Science, Vol.5, No.1. 2016.

[21] Azgomi, Mohammad Abdollahi, et al. "Introduction to the special issue on secure communications." Telecommunication Systems vol 69. 2018:

[22] OTAN. Tech Repport ACST–Strategy-CyberSecurity-001 "Cyber Security Strategy for Defence". Editado por COS STRAT. 2014

[23] “What Is a Security Operations Center (SOC)?”. McAfee. https://www.mcafee.com/enterprise/es-es/security-awareness/operations/what-is-soc.html

[24] ISACA, Cybersecurity Glossary, 2014. https://www.isaca.org/resources/glossary

[25] NIST, S. (2004). 800-61. Computer security incident handling guide, 800-61.

[26] ISO-18028-1:2006. Information technology — Security techniques — IT network security. 2006

[27] Josh Huff, OSINT: Open Source Intelligence. Disponible en https://www.coursehero.com/file/71790416/summit-archive-1533737204pdf/ . [Accedido el 5 de mayo de 2021]

.

Sesión de Formación e Innovación Educativa
http://doi.org/10.18239/jornadas_2021.34.64

Mapa Funcional de competencias en seguridad para el personal no TI de las universidades españolas

Josu Mendívil Caldentey
Universidad de Deusto
josu.mendivil@deusto.es
https://orcid.org/0000-0002-4774-8943

Miren Gutierrez Almazor
Universidad de Deusto
m.gutierrez@deusto.es
https://orcid.org/0000-0003-1527-3434

Borja Sanz Urquijo
Universidad de Deusto
borja.sanz@deusto.es
https://orcid.org/0000-0003-2039-7773

***Resumen*—Las empresas y organizaciones requieren de nuevas y eficientes iniciativas en formación y concienciación en ciberseguridad que se sumen a la constante evolución en la tecnología y en los procedimientos. La capacidad de una organización en general y de una universidad en particular para hacer frente a sus amenazas y vulnerabilidades depende en gran medida de la actitud y el conocimiento en ciberseguridad de su personal, y en consecuencia, de la existencia de un marco adecuado de competencias que identifique los ítems y niveles de formación y concienciación necesarios para cada puesto.**

Para contribuir a esta tarea, este trabajo propone, desde un nuevo enfoque metodológico, la construcción de un mapa de competencias que permita identificar las necesidades en el ámbito de la concienciación y formación en ciberseguridad para el personal no TI de las universidades españolas. El interés y relevancia de esta investigación se deriva de utilizar por primera vez un estándar o norma de seguridad como es el Esquema Nacional de Seguridad. Con ello se pretende ampliar la investigación en el campo de la formación en ciberseguridad, proponiendo el uso de estándares y normas de seguridad para configurar, orientar y justificar el diseño de competencias en esta materia.

***Index Terms*—Jornadas, Ciberseguridad, ENS, Mapa Funcional**

Tipo de contribución: *Formación e innovación educativa*

I. Introducción

Las universidades reflejan la creciente complejidad de nuestros sistemas económicos, sociales o culturales. Sin embargo, sus características intrínsecas y diferenciadoras respecto de otras organizaciones, su actividad, objetivos y funciones, unido a su relevancia social, las sitúan en un entorno caracterizado por una especial complejidad y por estar sujetas a condiciones cambiantes y siempre exigentes [1].

La cultura organizacional universitaria es un primer factor diferenciador. Ian Mcnay [2] contempla hasta cuatro culturas coexistiendo dentro de una universidad: el claustro, la administración, la dirección y la empresa. Todas ellas con sus propias características y necesidades, no siempre coincidentes. Un segundo factor diferenciador son los estudiantes. Identificados habitualmente como colectivo, presentan sin embargo necesidades y demandas muy heterogéneas. A estos dos factores debemos añadir la actividad investigadora de las universidades, en ocasiones con proyectos de alto valor estratégico o económico, desarrollada en un escenario en el que la seguridad de la información está supeditada a otros aspectos, como la colaboración y la compartición de recursos, el acceso libre a los servicios o el uso de operativas abiertas. [3]

Estas características hacen que las universidades, además de ser objetivos atractivos para sufrir un ataque, presenten un elevado número de vulnerabilidades y por tanto un alto riesgo de seguridad, siendo uno de los sectores más afectados por las ciberamenazas, tal y como demuestra el último informe llevado a cabo sobre ciberseguridad en el sector universitario español, donde el 80 % de las universidades participantes en el estudio afirmaron haber sufrido algún incidente de seguridad [4]. Ejemplos de esta situación son los ataques sufridos por la Universidad de Burgos, [5] la Universidad de Valladolid [6], la Universidad de Cádiz [7] o la Universidad de Castilla La Mancha [8].

Ante esta realidad, no existe un mapa de competencias que identifique las actitudes ni conocimientos necesarios en materia de seguridad de la información en el ámbito laboral no TI universitario español. En consecuencia, las labores de formación y concienciación en seguridad que se llevan a cabo en las universidades españolas no cuentan con un modelo o estándar de referencia basado en competencias que apoye y facilite la configuración o validación de planes de formación y concienciación. La única iniciativa existente es el convenio marco de colaboración entre CRUE e INCIBE firmado en el año 2016 con el objetivo de desarrollar actuaciones básicas de concienciación sobre ciberseguridad en las universidades [9].

En este escenario se presenta este trabajo, que desarrolla el Mapa Funcional de competencias asociado a las actividades en el área de la seguridad de la información que deben formar parte de las competencias laborales del personal no TI de las universidades españolas.

Desde el punto de vista competencial, el Mapa Funcional construido permitirá incorporar al mapa de competencias laborales de las universidades las competencias en seguridad de la información, dándoles la relevancia que deben tener. También dotará a los responsables de formación en seguridad de las universidades españolas de una herramienta con la que podrán configurar o validar sus propios planes de formación, así como diseñar planes de formación compartidos, al contar con un corpus de competencias estandarizado.

Como segunda contribución significativa al cuerpo de conocimiento, se explora el uso de estándares de la industria en la elaboración de mapas funcionales. Esta línea de investigación se lleva a cabo mediante la incorporación al modelado del Mapa Funcional, además de la participación de expertos en el ámbito analizado, las medidas de seguridad recogidas en el Anexo II del Esquema Nacional de Seguridad [10], norma de obligado cumplimiento para todas las universidades públicas y un referente para el resto de universidades. Este enfoque tiene como objetivo dotar de mayor credibilidad y niveles de

estandarización a los mapas funcionales obtenidos.

Normalizar las actividades en el área de la seguridad de la información como competencias laborales presenta varias ventajas, tanto para el personal laboral como para las organizaciones, entre las que podemos destacar:

- Para las organizaciones:
  - Permite disponer de un marco de referencia o patrón de medición de competencias con el que contrastar el nivel de formación y concienciación en seguridad de la información del personal no TI, tanto a nivel individual como agregado.
  - El marco de referencia ofrece un conjunto de competencias normalizadas, compartidas y conocidas.
  - Posibilita planificar actividades individualizadas de capacitación de las personas en el ámbito de la seguridad de la información.
  - Incrementa la seguridad de la organización.
  - Mejora el desempeño de la organización.
  - En aquellos casos en los que el marco o modelo de referencia sea compartido por un sector o industria, se puede hablar de un marco estandarizado, lo que supone nuevas ventajas:
    - Permite integrar y compartir esfuerzos de formación y concienciación.
    - El lenguaje sobre los contenidos y niveles de competencia es común.
    - Los niveles de competencia de las personas son homogéneos.
- Para las personas:
  - Les permite conocer su nivel competencial respecto a los requerimientos y necesidades de su organización en el ámbito de la seguridad de la información.
  - Tienen la posibilidad de que sus competencias sean objetivables y reconocidas

II. Fundamentos teóricos

II-A. Esquema Nacional de Seguridad

El Real Decreto 3/2010, de 8 de enero regula el Esquema Nacional de Seguridad, ENS, dando de este modo cumplimiento al artículo 42 de la Ley 11/2007, de 22 de junio, de acceso electrónico de los ciudadanos a los Servicios Públicos.

La finalidad del ENS es generar la necesaria confianza en el uso de los sistemas informáticos, las comunicaciones y los datos que manejan los ciudadanos y ciudadanas en su relación con las administraciones públicas. [10]

El ENS es el resultado de un largo trabajo coordinado por el Ministerio de Hacienda y Administraciones Públicas con el apoyo del Centro Criptológico Nacional y la participación de todas las Administraciones Públicas, incluyendo las universidades públicas y órganos colegiados con competencias en materia de administración electrónica. En su elaboración también se contó con los informes preceptivos de diferentes ministerios, con la Agencia Española de Protección de Datos, así como con la opinión de las asociaciones de la Industria del sector TIC.

En su Anexo II establece un conjunto de medidas de seguridad, organizadas en tres grupos: [11]

- Marco organizativo. Constituido por medidas relacionadas con la organización general de la seguridad.
- Marco operacional. Formado por las medidas encaminadas a proteger la operación del sistema de información.
- Medidas de protección. Orientadas a proteger los activos según su naturaleza y el nivel de seguridad requerido.

Estos tres apartados establecen setenta y cinco indicadores, de acuerdo con la siguiente distribución: el marco organizativo establece cuatro indicadores, el marco operacional identifica treinta y un indicadores organizados en seis grupos, y las medidas de protección recogen cuarenta indicadores estructurados en cinco grupos.

El origen, diseño, rigor y alcance de estas medidas las convierten en un excelente punto de partida para la construcción del Mapa Funcional:

- Son medidas de obligado cumplimiento para toda la Administración Pública, incluidas las universidades públicas, así como un referente y estándar de cumplimiento en el ámbito de la seguridad en España para muchas otras organizaciones y empresas.
- Son el resultado de un largo y minucioso proceso de análisis y estudio realizado desde el más alto nivel, por lo que su utilidad y pertinencia están fuera de duda.
- Su uso permite establecer la necesaria relación entre la necesidad de cumplimiento y el modo de cómo hacerlo efectivo en el ámbito de la formación y concienciación, dando de este modo cumplida respuesta a las medidas "5.2.3 Concienciación" y "5.2.4 Formación" del propio Anexo II, en los que se explicita la necesidad de llevar a cabo de manera regular todas las acciones que sean necesarias para formar y concienciar al personal sobre su responsabilidad en la seguridad de los sistemas de información. [10]

II-B. Competencia laboral

El concepto de competencia ha sufrido a lo largo del tiempo una notable evolución y desarrollo en su significado. En la actualidad engloba diferentes enfoques y conceptos, generando en ocasiones cierta confusión a la hora de definir qué se entiende por competencia [12], [13]. Esta situación se complica cuando se habla de competencia laboral, término sobre el que también existe confusión en su significado y uso [14], [15].

El término de competencia como aspecto motivacional, más allá de la mera capacidad de interactuar de manera efectiva con el entorno, fue tratado por primera vez por Robert White [16]. A partir de ese momento, y a medida que se desarrollan diferentes aproximaciones a una realidad tan compleja y ambigua como son las competencias, se diversifican las definiciones y enfoques que intentan categorizarlas. Una buena aproximación a esta diversidad puede ser el artículo "From task-based to competency-based" de Klas Soderquist et al [17], en el que después de llevar a cabo una amplia revisión de la literatura sobre los distintos enfoques en gestión de competencias, presenta una tipología de competencias que integra los enfoques previos.

El concepto de competencia laboral también ha seguido un proceso de redefinición constante, como adaptación natural a la evolución del desempeño laboral en las organizaciones. El concepto de competencia laboral no fue utilizado hasta 1973 por David McClelland en su conocido artículo "Testing for

competence rather than for intelligence", en el que plantea la insuficiencia de los tradicionales test basados en la inteligencia para predecir el éxito en el desempeño laboral. [18]

Andrew Gronzci y James Athanasou, [19] señalan que las competencias pueden tipificarse en tres grupos: como lista de tareas, como conjunto de atributos personales y la tercera como relación holística.

II-B1. Modelo de competencias como lista de tareas: Basado en el modelo Taylorista, este enfoque de las competencias como una relación de tareas a desempeñar en el puesto de trabajo se basa en la definición funcional del puesto de trabajo, explicando las competencias laborales como la relación de tareas y conocimientos asociadas a un puesto de trabajo. Esta relación específica, ordena y define las tareas y conocimientos que debe desarrollar y conocer una persona en su actividad laboral. Se trata un enfoque muy reduccionista, pero sin embargo ampliamente utilizado por su facilidad de elaboración y seguimiento.

En este modelo, las competencias (entendidas como tareas y conocimientos) observadas en los puestos de trabajo permiten establecer un perfil estándar de requisitos para cada competencia. Esto permite que se pueda evaluar al trabajador respecto al perfil estándar, y de acuerdo con los resultados obtenidos establecer el aprendizaje necesario en función de las necesidades de cada puesto.

II-B2. Modelo de competencias como conjunto de atributos personales: El enfoque anterior fue completado por los trabajos de David McClelland [18] con la inclusión de los comportamientos y actitudes, que ayudaban a explicar de una manera más completa el desempeño laboral. La competencia laboral, desde esta perspectiva, se entiende no sólo por lo que la persona sabe y puede hacer; también por lo que quiere hacer. Un ejemplo de este enfoque es la definición de Anne Marelli [20], que define la competencia como "una capacidad laboral, medible, necesaria para realizar un trabajo eficazmente, es decir, para producir los resultados deseados por la organización. Está conformada por conocimientos, habilidades, destrezas y comportamientos que los trabajadores deben demostrar".

Esta definición presenta dos aspectos destacados. En primer lugar, la competencia laboral debe ser medible. Esta es una condición que debe satisfacer cualquier competencia laboral y un aspecto común en todos los modelos. Tal y como ya figuraba en el modelo de lista de tareas, medir una competencia significa que puede ser evaluada, y en consecuencia formar parte de programas de formación. [21]

El segundo aspecto es la inclusión del comportamiento del trabajador. La competencia viene definida no sólo por lo que la persona sabe hacer, también por lo que quiere hacer. En el ámbito de la seguridad de la información este es un aspecto muy relevante si consideramos que en España, el 67 % de los usuarios opinan que las herramientas de seguridad son demasiado costosas, y el 80 %, cuatro de cada cinco personas, piensan que el uso de herramientas y procedimientos de seguridad requieren demasiado esfuerzo por su parte. [22]

II-B3. Modelo de competencias como relación holística: Los dos enfoques anteriores, complementarios, derivan en un enfoque holístico, en el que, a los conocimientos y actitudes se añade el contexto en el que se desarrolla la actividad. Un enfoque relevante desde la seguridad de la información, donde los peligros y amenazas son difusos y en permanente cambio. Así pues, a las habilidades, conocimientos y actitud para desempeñar un puesto de trabajo se le añade el contexto en el que se desarrolla la actividad, lo que permite incluir conceptos y valores como la ética, a la que la Comisión Europea considera una meta-competencia, al considerarla una parte integral del conocimiento, las habilidades y las actitudes necesarias para desarrollar una actividad profesional [23].

En este enfoque, las competencias son entendidas como relaciones complejas de habilidades, conocimientos, actitudes y valores, que varían de acuerdo con las necesidades específicas del entorno [24].

Los anteriores modelos comparten una serie de puntos comunes, que podemos considerar como características que debe satisfacer toda competencia [25].

- Cada competencia se identifica con un nombre y cuenta con una definición precisa.
- Cada competencia tiene un determinado número de niveles que reflejan conductas observables, no juicios de valor.
- Las competencias se pueden desarrollar.
- Los puestos de trabajo están asociados con un perfil de competencias, entendido como un inventario de las mismas, incluyendo el nivel requerido para cada una de ellas.
- Todo modelo de gestión por competencias llega hasta la definición de niveles y de indicadores de conductas esperadas.

En la actualidad existen varias metodologías para identificar perfiles de competencias, entre las que se pueden destacar [26]:

- DACUM (Developing a Curriculum). Desarrollado en el Centro de Educación y Formación para el Empleo de la Ohio State University en 1995, es un método globalmente aceptado dada su fiabilidad y eficacia en el análisis y descripción de puestos de trabajo estándar. Se trata de un procedimiento estructurado que tiene como objetivo identificar, de la manera más clara y precisa posible, lo que el trabajador debe conocer y poder hacer para desempeñar adecuadamente su desempeño.
- SCID (Systematic Curriculum and Instructional Development). Es un modelo complementario a DACUM, orientado a la producción de materiales de instrucción relevantes y de calidad, a partir del análisis del puesto de trabajo desarrollado usando el método DACUM.
- AMOD (A Model). Se trata de una variante de la metodología DACUM, orientada a identificar las competencias de una familia de ocupaciones.
- Análisis funcional. Es una metodología que permite identificar las competencias laborales que debe reunir una persona para desempeñarse de manera competente en su puesto de trabajo. Los principios metodológicos de este enfoque, presentados en el siguiente apartado, resultan especialmente aptos para alcanzar los objetivos de este estudio, y por ello será la metodología empleada en el mismo.

II-C. Análisis funcional

El Análisis funcional es un método utilizado para identificar competencias laborales mediante la desagregación de las

funciones de una empresa u organización en subfunciones más específicas, que a su vez son divididas en actividades cada vez más concretas, hasta conseguir identificar las acciones elementales que pueden ser asignadas a un trabajador [27].

A diferencia de otras metodologías que trabajan con un enfoque de competencias entendidas como una relación de tareas, el análisis funcional analiza las relaciones existentes entre la función productiva de un determinado sector, organización o puesto y las habilidades, conocimientos y actitudes necesarias para desempeñarlas con éxito.

Es necesario señalar que esta metodología no describe procesos, sino resultados. No importa en consecuencia conocer cómo obtiene un trabajador el resultado, sino que el resultado se logre. Este es un factor importante, ya que dota al trabajador o trabajadora de la posibilidad de utilizar diferentes métodos o estrategias para conseguir los resultados esperados. Por este motivo, en el Mapa Funcional no se deberán describir tareas, sino identificar resultados [21]. Este es un enfoque muy conveniente para los objetivos de este estudio, ya que nos interesará en efecto los conocimientos y actitudes de los trabajadores universitarios hacia los aspectos de seguridad de la información en su trabajo, y no cómo los lleven a cabo.

II-C1. Principios metodológicos: Existen tres principios metodológicos en los que se basa el Análisis Funcional [28]:

- El Análisis Funcional se aplica de lo general a lo particular:
  La metodología se inicia identificando un propósito principal que defina la finalidad de la actividad productiva. A partir del propósito se realiza una desagregación por funciones, de acuerdo con los siguientes niveles:
  - Funciones clave
  - Funciones principales
  - Funciones básicas o unidades de competencia
  - Elementos de competencia

  El proceso concluye cuando se alcanzan las funciones productivas mínimas que puede desarrollar un trabajador o trabajadora.
- El Análisis Funcional debe identificar funciones delimitadas e independientes de un contexto concreto:
  Las funciones deben tener un comienzo y un fin claramente delimitado, y no estar asociadas a un puesto de trabajo concreto. Esto permite que las funciones identificadas sean válidas en diferentes entornos laborales. Para facilitar este principio, está generalmente aceptado el utilizar la siguiente estructura gramatical para expresar las funciones:

  Verbo + objeto + condición

  En esta estructura, el verbo indica la acción que debe ser ejecutada por la persona, el objeto describe el elemento sobre el que recae la acción, y la condición señala la forma, el criterio o el contexto que debe ser considerado en la realización de la acción [29].
- El desglose en el Análisis Funcional se hace siguiendo la lógica de causa-efecto:
  Tal y como se ha comentado en el primer principio, la metodología trabaja de lo general a lo particular. Para ello, la lógica que se lleva a cabo se basa en la pregunta: "para cumplir con este propósito (o función), ¿qué funciones son necesarias realizar?"

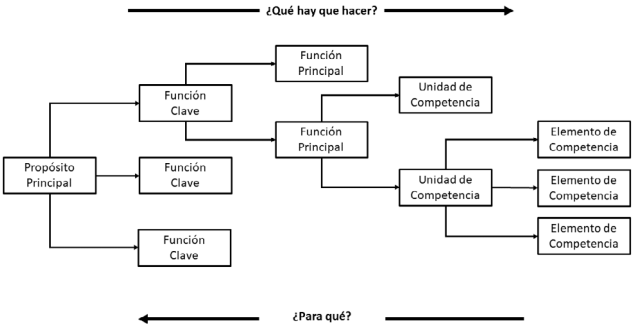


Figura 1. Esquema de un mapa funcional. Adaptado de [21]

II-C2. Procedimiento: Antes de entrar en detalle en describir el procedimiento, es conveniente señalar que el análisis funcional no intenta ser un método exacto o un conjunto de fórmulas que den un resultado exacto [30]. Se trata de una metodología que permite obtener un perfil de competencias laborales de una manera coherente y sistemática, y cuyos resultados garanticen un estándar de competencias consistentes y comparables.

Como ya se ha dicho, el método de trabajo comienza estableciendo el propósito principal de la función productiva y a continuación se pregunta sucesivamente qué funciones hay que llevar a cabo para conseguir la función precedente. De acuerdo con este método, una vez establecido el propósito principal, se inicia el proceso de desagregación respondiendo a la pregunta ya mencionada de qué hay que hacer para lograr el propósito principal. La respuesta será un nivel de desagregación formado por las actividades, conocimientos o actitudes que identificaremos como Funciones Clave. A continuación, se repetirá la pregunta para obtener un segundo nivel de desagregación, que identificaremos como Funciones Principales. Se repite el proceso para un tercer nivel de desagregación, al que identificaremos como Unidades de Competencia. Los siguientes niveles, que identificaremos como Elementos de Competencia, tienen que identificar lo que debe ser capaz de hacer el trabajador o trabajadora.

De acuerdo con el alcance del análisis, la profundidad en la desagregación puede variar. En general, se considera que un mapa de una empresa u organización constará de tres o cuatro niveles. Y el mapa de una ocupación laboral uno o dos. El proceso finaliza cuando la función pueda ser desempeñada por una persona [30]. Dicho de otro modo, se pueda utilizar la expresión "la persona debe ser capaz de…" con la descripción del elemento de competencia [15].

El resultado de este proceso será la obtención del Mapa Funcional, que puede ser representado gráficamente tal y como se muestra en la figura 1, donde leído de izquierda a derecha, responde a qué es necesario hacer, mientras que, leído de derecha a izquierda, responde a porqué se hace esto.

En esta metodología, el proceso se lleva a cabo por un grupo de expertos [27] que identifican el propósito principal y llevan a cabo el proceso recurrente de desagregación. No cabe dudar de la validez de este enfoque, especialmente si se considera que en las funciones productivas que son objeto de estudio no existe ningún tipo de conocimiento estandarizado

previo.

III. Metodología de la investigación

III-A. Objetivos

El objetivo principal de este trabajo es identificar las competencias en el ámbito de la seguridad de la información del personal no TI de las universidades españolas, a través de la construcción del Mapa Funcional, utilizando para ello las medidas de seguridad del Anexo II del ENS.

Como objetivos particulares:

- Determinar qué medidas de seguridad del ENS aplican en un mapa de competencias del personal no TI de las universidades españolas.
- Establecer la relación existente entre las medidas de seguridad del ENS y las funcionas clave y funciones principales del Mapa Funcional definido.
- Identificar los elementos de competencia que el personal no TI de las universidades españolas debe conocer o ser capaces de realizar en el ámbito de la seguridad de la información en su puesto de trabajo.

III-B. Diseño de investigación

La construcción del Mapa Funcional forma parte de un trabajo más amplio. Basado en el modelo de discrepancia de Witkin y Altschuld [32], la investigación tiene como objetivos definir "lo que debería ser", en cuya fase se sitúa el trabajo presentado en este documento, determinar "lo que es", identificando los perfiles laborales y proponiendo un modelo de evaluación respecto al Mapa Funcional construido, y finalmente "cómo alcanzarlo", proponiendo un modelo de formación y concienciación que permita superar las brechas detectadas.

Para definir "lo que debería ser", este trabajo pretende abrir nuevas hipótesis construyendo un mapa funcional basado en un estándar de seguridad, por lo que se trata de un estudio exploratorio y descriptivo, de carácter no experimental y donde dado el objetivo de la investigación, la metodología cualitativa se presenta como la mejor alternativa. No será por tanto un proceso lineal ni estrictamente definido como en los procesos cuantitativos, sino iterativos y recurrente [33], donde las entrevistas en profundidad y la observación participante serán las técnicas de recogida de información y datos.

Para la construcción del Mapa Funcional se partirá de la metodología del Análisis Funcional con el aporte de la utilización del ENS en su construcción.

III-B1. Participantes: El Análisis Funcional presenta como una característica relevante en su método la necesidad de que sea desarrollado por expertos y expertas de la actividad o materia a analizar [15]. De acuerdo con esta premisa, se ha formado un grupo de personas expertas en seguridad de la información pertenecientes a universidades españolas. En su selección se han tenido presentes una serie de características específicas que aseguren su validez:

- Deberán tener una experiencia contrastada en materia de seguridad de la información. Para ello se seleccionarán a personas que ocupen el cargo de Responsables de seguridad, CISO, o sus equivalentes en sus universidades, con una experiencia de al menos cinco años.
- Se buscará la mayor variedad posible respecto al tamaño de universidad, la situación geográfica y la titularidad de la universidad.
- Se garantizará la paridad de género.

En base a estos criterios, se construyen dos equipos. El equipo técnico, formado por la autora y autores de este trabajo que llevarán a cabo las labores de diseño y coordinación de las entrevistas de trabajo, así como la elaboración de los materiales de trabajo y recopilación de resultados. El grupo de expertos y expertas está formado por cinco CISO de universidades españolas.

En total, serán ocho personas las que llevarán a cabo esta tarea, siguiendo las recomendaciones de que el grupo de expertos y expertas no sea numeroso, no excediendo de diez personas [15].

III-B2. Procedimiento de análisis: Las tareas que deberán llevar a cabo los expertos y expertas son las siguientes:

- Transformar los textos de las medidas de seguridad del Anexo II del ENS a la estructura gramatical propuesta por la metodología del Análisis Funcional.
- Identificar las medidas de seguridad que no aplican al ámbito de estudio abordado.
- Señalar los elementos de competencia que permitan dar respuesta a las unidades de competencia identificadas.

En primer lugar, se adaptarán los textos del ENS adoptando la estructura gramatical propuesta en la metodología del análisis funcional, comprobando que se cumple no sólo dicha estructura, sino que el resultado sea el proceso de reflexión propuesto por la misma, esto es, preguntarse qué hay que hacer para lograr ese propósito. A continuación, se organizarán las medidas de seguridad recogidas en el ENS con la estructura de árbol característica de un mapa funcional, asociando cada nivel de agrupación del ENS con el correspondiente del Mapa Funcional. De esta manera se consigue explicitar los primeros niveles del Mapa Funcional, correspondientes al propósito principal, funciones clave, funciones principales y unidades de competencia.

Para llegar a este conocimiento, las cuestiones que al menos deberá responder cada experto o experta para cada una de las medidas serán las siguientes:

- ¿Aplica esta medida para la mejora de la formación y concienciación en materia de seguridad de la información para el PAS/PDI de las universidades españolas?
- Si aplica, ¿cuáles son las acciones concretas en formación que deben de llevarse a cabo para asegurar que el PAS/PDI cumple en el ámbito de su responsabilidad con la medida de seguridad?
- Si aplica, ¿cuáles son las acciones concretas de concienciación que deben llevarse a cabo para asegurar que el PAS/PDI cumple en el ámbito de su responsabilidad con la medida de seguridad?
- Si la medida tiene varios niveles de seguridad (Bajo, medio y Alto) ¿se deben adoptar acciones de formación o concienciación diferentes para alguno de ellos? ¿Si así fuera, qué acciones se deben registrar en cada nivel?

El siguiente paso será identificar los elementos de competencia, donde ya se establece lo que el trabajador o trabajadora debe conocer o ser capaz de hacer.

Para ello, los investigadores se reúnen con el primer entrevistado, a quien presentan y explican los motivos y objetivos del estudio. También se le orienta en la dinámica de uso del Análisis Funcional, y en lo que se espera de él. Confirmada su comprensión, se comienza definiendo el propósito principal del Mapa Funcional a partir de los objetivos del ENS, adaptándolo al Análisis Funcional. A continuación, se repite la misma mecánica con las tres funciones claves, identificadas en la primera clasificación de las medidas de seguridad. El siguiente paso es repasar los diferentes ítems que conforman las medidas de seguridad, ajustándolas a la sintaxis propia del Análisis Funcional, y de esta manera ir obteniendo las funciones principales, las unidades de competencia y los elementos de competencia.

Con la segunda persona entrevistada se repite la parte inicial de explicación y orientación. En este caso, de acuerdo con la Teoría Fundamentada [34], se le muestran las funciones clave, las funciones principales, las unidades de competencia y los elementos de competencia realizados por el primer entrevistado, para que corrija, elimine o añada lo que considere oportuno, incluyendo comentarios o aclaraciones.

El resultado de esta segunda experta se presenta al primer experto para que corrobore o aclare algún aspecto no contemplado en su entrevista inicial.

Al tercer entrevistado se le muestra el Mapa Funcional con las aportaciones anteriores, incluyendo las explicaciones pertinentes sobre los cambios realizados y las correcciones efectuadas. Éste aporta su conocimiento para modificar o ratificar lo realizado hasta el momento, añadiendo a su vez lo que considere oportuno. Esta información se comparte con los expertos y expertas anteriores, para eventualmente validar o corregir lo modificado. En este último caso se comparten las correcciones y motivos con el tercer experto o experta, hasta alcanzar el consenso. Con los siguientes entrevistados y entrevistadas se sigue la misma dinámica de trabajo. En este proceso, el cuarto entrevistado en su primer análisis validó casi en su totalidad el Mapa Funcional obtenido y el quinto lo ratificó con ajustes menores, alcanzando con ello el nivel de saturación. El resultado final es enviado a todos los expertos y expertas para que confirmen una vez más si están de acuerdo con el mapa obtenido.

IV. Resultados

El resultado de este trabajo es el Mapa Funcional sobre la seguridad de la información del personal no TI de las universidades españolas. Este mapa se compone de un propósito principal, tres funciones clave, dieciocho funciones principales, setenta y cinco unidades de competencia y ciento diez elementos de competencia.

Propósito principal:

La metodología del Análisis funcional comienza con la identificación del propósito clave del área objeto de análisis. Este propósito definirá la meta, objetivo o finalidad de la actividad analizada. En esta investigación se ha desarrollado a partir del Preámbulo del ENS.

"Garantizar la seguridad de los sistemas, los datos, las comunicaciones, y los servicios electrónicos, para permitir a los usuarios de la universidad el desarrollo de sus actividades a través de estos medios y fundamentar la confianza en que los sistemas de información prestarán sus servicios y custodiarán la información de acuerdo con sus especificaciones funcionales, sin interrupciones o modificaciones fuera de control, y sin que la información pueda llegar al conocimiento de personas no autorizadas."

Funciones clave, funciones principales, unidades de competencia y elementos de competencia:

Dado el tamaño del Mapa Funcional completo, en la Tabla I se presenta, a modo de ejemplo, las funciones, unidades y elementos de competencia correspondientes a la función principal"Proteger los equipos". Para mayor comodidad a la hora de trabajar los elementos de la tabla, se ha mantenido la numeración que se utiliza en el Anexo II del ENS.

V. Conclusiones

El resultado de este estudio es el Mapa Funcional de competencias en ciberseguridad para el personal no TI de las universidades españolas. Mapa Funcional que parte para su construcción, como elemento novedoso y aporte al cuerpo de conocimiento, de las medidas de seguridad del Esquema Nacional de Seguridad. La metodología empleada permite establecer una relación directa e inequívoca entre las competencias que en el ámbito de la seguridad de la información debe entender, asumir, conocer y ser capaz de realizar el personal universitario no TI y el estándar de seguridad seleccionado. El mapa obtenido permite trabajar la evaluación y formación en este campo sobre criterios conocidos, comunes y contrastados, así como disponer de una poderosa herramienta de aseguramiento del cumplimiento de los requisitos de formación y concienciación del propio estándar.

V-A. Aportación del trabajo

El Mapa Funcional construido bajo este paradigma proporciona a las universidades españolas un instrumento de trabajo de suma utilidad para disponer de un referente común respecto a qué competencias en el ámbito de la seguridad de la información deben ser conocidas por su personal, así como contar con un instrumento que les permita medir y evaluar dichas competencias. Todo ello ayudará a la mejora de los procesos de formación y concienciación en seguridad de la información, y como consecuencia, en un mejor nivel de seguridad en las universidades.

Las competencias definidas permiten contar con un estándar común, lo que permite compartir esfuerzos y costes destinados a la elaboración de procesos de formación y concienciación, así como los de seguimiento y evaluación. En este sentido, se podría plantear un futuro estándar de competencia, entendiendo como tal "un proceso de acuerdo entre empresas, trabajadores e instituciones públicas con el propósito de establecer un estándar sobre las competencias que son representativas de una determinada ocupación o área ocupacional" [21].

Los trabajadores universitarios, en este contexto, tienen un conocimiento claro de qué es lo que deben conocer y qué se espera de ellos.

V-B. Pasos futuros y otras posibilidades que ofrece este enfoque

Dado que el ENS es una ley de obligado cumplimiento para las organizaciones y empresas públicas españolas, para empresas privadas nacionales que trabajen con la administración

Tabla I
MAPA FUNCIONAL (DETALLE)

Funciones clave	Funciones principales	Unidades de competencia	Elementos de competencia
...	...	...	...
5. Proteger activos concretos, según su naturaleza, con el nivel requerido en cada dimensión de seguridad.	5.3. Proteger los equipos.	5.3.1. Exigir que los puestos de trabajo permanezcan despejados.	5.3.1.1. Entender la necesidad e importancia de mantener el puesto de trabajo despejado, sin más material encima de la mesa que el requerido para la actividad que se está realizando en cada momento.
			5.3.1.2. Comprender la importancia de guardar en lugar cerrado la documentación cuando no se esté utilizando.
		5.3.2. Bloquear al cabo de un tiempo prudencial de inactividad los equipos, requiriendo una nueva autenticación para reanudar la actividad.	5.3.2.1. Entender la necesidad de que el puesto de trabajo se bloquee después de un tiempo de inactividad.
		5.3.3. Proteger los equipos que sean susceptibles de salir de las instalaciones y que no puedan beneficiarse de su protección física.	5.3.3.1. Conocer el procedimiento para informar de la pérdida o sustracción de un portátil.
			5.3.3.2. Entender la necesidad de proteger el portátil y la información que contiene, así como tenerlo en todo momento controlado y custodiado.
			...
		5.3.4. Garantizar la existencia y disponibilidad de medios alternativos para el tratamiento de la información para el caso de que fallen los medios habituales.	5.3.4.1. Conocer el procedimiento de acceso a medios alternativos dispuesto por la universidad.
...	...	...	...

española y un referente para empresas privadas, la metodología propuesta en este trabajo puede ser también utilizada para otras organizaciones y sectores empresariales. Bastaría con trabajar con un grupo de responsables de seguridad del sector del que se desee definir el Mapa Funcional.

También podría utilizarse esta metodología sustituyendo las medidas de seguridad del ENS por los controles de seguridad de la ISO 27001 contenidos en el Anexo A de esta norma. Siendo la ISO 27000 una norma internacional de amplia extensión y uso, podrían obtenerse con esta propuesta metodológica mapas funcionales de competencias en ciberseguridad específicos para cualquier sector e industria.

Como se ha indicado, este trabajo forma parte de un proyecto más amplio. Una vez establecidos los elementos de competencia, a continuación se tiene previsto identificar los roles o perfiles laborales existentes en la universidad, para finalmente poder determinar para cada rol o perfil laboral qué elementos de competencia le aplican y en qué grado. Con ello se podrán diseñar mapas específicos de competencias para cada perfil laboral, y de esa manera disponer de un perfil de competencias con el que se podrán establecer planes personalizados de formación y concienciación en un área tan crucial para las universidades españolas.

REFERENCIAS

[1] M. Tomás, D. Castro: "Multidimensional Framework for the Analysis of Innovationsat Universities in Catalonia", en *Education Policy Analysis Archives*, vol. 19, n. 27, 2010.

[2] I. Mcnay: "From Collegial Academy to the Corporate Enterprise: The Changing Cultures of Universities", 1995.

[3] F. J. Sampalo, J. Cortés, J. P. Gumbau, J. Mendívil: "Marco de recomendaciones de Seguridad y Auditoría en Universidades", en *CRUE TIC*, 2018.

[4] Deloitte: "Estudio de Ciberseguridad: Principales Universidades en España". [En línea]. Disponible en https://www2.deloitte.com/content/dam/Deloitte/es/Documents/governance-risk-compliance/Deloitte-ES-GRC-Ciberseguridad-Universidades.pdf. [Accedido: 19-mar-2021]

[5] INCIBE: "Incidente de seguridad en la Universidad de Burgos". [En línea]. Disponible en https://www.incibe-cert.es/alerta-temprana/bitacora-ciberseguridad/incidente-seguridad-universidad-burgos. [Accedido: 19-mar-2021]

[6] INCIBE: "Expuestos datos de alumnos de la Universidad de Valladolid". [En línea]. Disponible en https://www.incibe-cert.es/alerta-temprana/bitacora-ciberseguridad/expuestos-datos-alumnos-universidad-valladolid. [Accedido: 19-mar-2021]

[7] INCIBE: "La Universidad de Cádiz sufre ciberataque de ransomware". [En línea]. Disponible en https://www.incibe-cert.es/alerta-temprana/bitacora-ciberseguridad/universidad-cadiz-sufre-ciberataque-ransomware. [Accedido: 19-mar-2021]

[8] INCIBE: "La UCLM ha sido víctima del ransomware Ryuk". [En línea]. Disponible en https://www.incibe-cert.es/alerta-temprana/bitacora-ciberseguridad/uclm-ha-sido-victima-del-ransomware-ryuk. [Accedido: 8-may-2021]

[9] CRUE-TIC: "Adaptación del Kit de Concienciación de INCIBE a universidades". [En línea]. Disponible en https://tic.crue.org/wp-content/uploads/2017/11/10.30-Seguridad-presentacion_kit_incib_CRUE_20171026_US.pdf. [Accedido: 22-mar-2021]

[10] BOE: "Real Decreto 3/2010, de 8 de enero, por el que se regula el Esquema Nacional de Seguridad en el ámbito de la Administración Electrónica", en *Boletín Oficial del Estado*, n. 25, pp. 8089-8138, 2010.

[11] CCN CERT: "Anexo II. Medidas de seguridad". [En línea]. Disponible en https://www.ccn-cert.cni.es/publico/ens/ens/index.html. [Accedido: 19-mar-2021]

[12] M. Salman, S. A. Ganie, I. Saleem: "The concept of competence: a thematic review and discussion", en *European Journal of Training and Development*, 2020.

[13] C. Levy-Leboyer: *"Gestión de las competencias"*, Barcelona, Gestión 2000, 1997.

[14] M. Alles: *"Dirección estratégica de RRHH. Gestión por competencias"*, Buenos Aires, Granica, 2015.

[15] M. Irigoin, F. Vargas: "Competencia laboral: manual de conceptos, métodos y aplicaciones en el sector salud", en *Organización Internacional del Trabajo. CINTERFOR*, 2002.

[16] Robert White: "Motivation reconsidered: the concept of competence", en *Psychological Review*, vol. 66, n. 5, pp. 297-333, 1959.

[17] K. Soderquist, A. Papalexandris, G. Ioannou, G. Prastacos: "From task-based to competency-based. A typology and process supporting a critical HRM transition", en *Personnel Review*, vol. 39, n. 3, pp. 325-346, 2010.

[18] D. McClelland: "Testing for Competence Rather Than for Intelligence", en *American Psychologist*, 1973.

[19] A. Gonczi, J. Athanasou: *"Instrumentación de la educación basada en competencias. Perspectiva de la teoría y la práctica en Australia"*, Mexico, Limusa, 1996.

[20] A. Marelli: "Introducción al análisis y desarrollo de modelos de competencia", en *Documento de trabajo*, 1999.

[21] F. Vargas, F. Casanova, L. Montanaro: "El enfoque de competencia laboral: manual de formación", en *OIT/Cinterfor*, 2001.

[22] The Cocktail Analysis: "Panorama actual de la Ciberseguridad en España". [En línea]. Disponible en https://drive.google.com/file/d/18TNjaDus-lrSl5gL5Wt-Z4DOsKXtQ46m/view. [Accedido: 20-mar-2021]

[23] The European Commission: "Explaining the European Qualifications Framework for Lifelong Learning". [En línea]. Disponible en https://europa.eu/europass/system/files/2020-05/EQF-Archives-EN.pdf. [Accedido: 20-mar-2021]

[24] D. Guerrero, I. de los Ríos: "Modelos internacionales de competencias profesionales", en *DYNA: Ingeniería e industria*, vol.88, n. 3, pp. 266-270, 2013.

[25] Miriam Escobar: "Las competencias laborales: ¿La estrategia laboral para la competitividad de las organizaciones?", en *Estudios Gerenciales*, vol.21, n. 96, pp. 31-35, 2005.

[26] Sistema Nacional de Certificación de Competencias Laborales: "Mirada comparativa sobre métodos para identificar competencias laborales. Documento de trabajo Nº 3". [En línea]. Disponible en https://www.oitcinterfor.org/sites/default/files/file_publicacion/ documento_de_trabajo_chilevalora_n__3_130111.pdf [Accedido: 20-mar-2021]

[27] P. Martínez, M. Martínez: "La tutoría sanitaria: mapa funcional ¿Amenaza u oportunidad?", en *Revista de Investigación Educativa*, vol. 29, n. 1, pp. 111-135, 2011.

[28] Consejo de Normalización y Certificación de Competencia Laboral de México. CONOCER: "Análisis Ocupacional y Funcional del Trabajo". [En línea]. Disponible en https://www.oei.es/historico/oeivirt/fp/iberfop03.htm, 2000. [Accedido: 19-mar-2021]

[29] Programa Regional de Formación Ocupacional e Inserción Laboral (FOIL): "Metodología para la elaboración de Normas de Competencia Laboral". [En línea]. Disponible en https://www.ilo.org/wcmsp5/groups/public/—americas/—ro-lima/—sro-san_jose/documents/publication/wcms_207580.pdf. [Accedido: 19-mar-2021]

[30] Servicio Nacional de Aprendizaje de Colombia [SENA]: "Guía de Apoyo para la Elaboración del Análisis Funcional". [En línea]. Disponible en https://www.oitcinterfor.org/sites/default/files/certificacion/ChileValora_GuiaApoyoAnalisisFuncional.pdf. [Accedido: 19-mar-2021]

[31] CCN-CNI: "ENS. FAQ". [En línea]. Disponible en https://www.ccn.cni.es/index.php/es/esquema-nacional-de-seguridad-ens/faq-ens. [Accedido: 20-mar-2021].

[32] B. R. Witkin, J. W. Altschuld: *"Planning and Conducting Needs Assessments: A Practical Guide"*, Thousand Oaks, CA, Sage Publications, 1995.

[33] R. Hernández, C. Fernández, P. Baptista: *"Metodología de la investigación"*, Mexico D.F., McGraw-Hill, 2014, pp. 394-467

[34] B. G. Glaser, A. L. Strauss: *"The Discovery of Grounded Theory. Strategies for Qualitative Research"*, USA, Aldine Transaction, 2000.

Sesión de Formación e Innovación Educativa
http://doi.org/10.18239/jornadas_2021.34.65

Análisis de componentes principales (ACP), una aproximación basada en proyectos

Diego Asterio de Zaballa
Research Institute of Applied Sciences in Cybersecurity (RIASC), Universidad de León (MIC), Dirección Miguel Carriegos,
https://orcid.org/0000-0003-3464-4712,
dzabr@unileon.es

Miguel Carriegos
Research Institute of Applied Sciences in Cybersecurity (RIASC), Departamento de Matemáticas, Universidad de León,
https://orcid.org/0000-0002-6850-0277
mcarv@unileon.es

***Resumen*—El impacto de los métodos de reducción de dimensionalidad en el campo del Aprendizaje Automático es notorio. La tendencia actual indica que la Ciberseguridad es un área de conocimento íntimamente ligada a este campo. Esto pone de manifiesto una necesidad de los programas de formación en Ciberseguridad: la de adquirir competencias en dichos métodos.**

En este artículo se presenta una experiencia didáctica centrada en la reducción de dimensionalidad. La metodología docente escogida se basa en la estrategia metodológica de Aprendizaje Basado en Proyectos. Con esta, se ha abordado el análisis de componentes principales. El pilar fundamental desde el que realizar dicho análisis es la Descomposición en Valores Singulares de una matriz compleja. A partir de esta descomposición los alumnos han realizado un trabajo final: el cálculo de las *eigenfaces* usando el lenguaje de programación *Python*. Todo ello con el objetivo de que el estudiante aprendiese la importancia y las limitaciones del método. Los resultados son satisfactorios aunque todavía presentan margen de mejora.

***Palabras clave*—Análisis de componentes principales, ciberseguridad, aprendizaje basado en proyectos, formación**

Tipo de contribución: *Formación e innovación educativa en desarrollo*

I. Introducción

Uno de los retos tecnológicos actuales de mayor embergadura consiste en el tratamiento de la enorme cantidad de datos generados diariamente. Se utilizan técnicas de reducción de la dimensionalidad con el objetivo de dar tratamiento a estas grandes masas de datos. Esto permite utilizar los datos de manera más eficiente y representativa. Asimismo, estas técnicas influyen de manera transversal en la ciberseguridad. En un mundo donde se generan millones de datos de incidencias cada veinticuatro horas, la utilidad de estas técnicas se pone de manifiesto [22]. Por otra parte, la ingente cantidad de datos con la que se debe trabajar dificulta los procesos de toma de decisiones en materia de ciberseguridad. También son de gran ayuda en este sentido las técnicas de reducción de la dimensionalidad veasé [2]. Por este motivo, se requiere que la formación de profesionales de la Ciberseguridad haga hincapié en las técnicas de reducción de dimensionalidad existentes.

Presentamos una experiencia didáctica en el contexto de la asignatura optativa "Matemáticas para Ciberseguridad II Análisis de datos y redes" del "Máster Universitario en Investigación en Ciberseguridad" de la Universidad de León.

Al cursar la asignatura que aquí nos ocupa, se obtienen competencias que ayudan a la correcta adquisición de todos estos conocimientos.

El diseño de la guía docente se ha realizado a partir del modelo pedagógico de **Aprendizaje Basado en Proyectos** a partir del cual se ha enfocado una asignatura esencialmente teórica de forma práctica. Se ha centrado el trabajo de los estudiantes en un objetivo concreto. Esto, sumado al trabajo autónomo de los estudiantes les ha permitido alcanzar los objetivos.

El objetivo principal de la metodología docente, recogida en el siguiente artículo, es aprovechar las condiciones tan especiales del grupo para presentar una posible alternativa a la metodología clásica. Como trabajo futuro, a fin de poder comparar ambas metodologías, se debe escalar la metodología docente que aqui se plantea hasta poder ser aplicada en una clase con un número normal de alumnos.

I-A. Organización del artículo

El artículo se organiza como sigue:

- La sección II recoge el estado del arte del **Análisis de componentes principales** (ACP), la ciberseguridad y la educación.
- La sección III describe una experiencia docente basada en proyectos en un curso de máster.
- La sección IV analiza los resultados obtenidos a partir de esta experiencia
- La sección V recoge conclusiones sobre la enseñanza-aprendizaje y una discusión acerca del trabajo futuro.

II. Trabajo relacionado: Análisis de componentes principales, Ciberseguridad y educación

II-A. El ACP y la Ciberseguridad

El ACP [13] se utiliza para reducir la dimensionalidad de los datos. Escogiendo una pequeña cantidad de componentes principales se explica un porcentaje suficiente de la varianza de los datos. Es de vital importancia en otras disciplinas que se relacionan directamente con la Ciberseguridad: en ciencia de datos se utiliza para reducir [30], tratar [24] y limpiar [11] bases de datos. Algunos autores [3] señalan el ACP como pretratamiento indispensable para aplicar algoritmos de Aprendizaje Automático . Y en otras ocasiones, como en este artículo [14], el ACP se utiliza como herramienta principal

para el control de intrusiones en sistemas industriales de control. En [1], el ACP se utiliza junto a una red neuronal para detectar ciberataques a partir de observar un sistema físico. Por último, en [25], se utiliza ACP y un "Autoencoder" para detectar anomalías en un sistema de control.

II-B. El ACP y la educación

En la literatura que relaciona el ACP y la educación se distinguen dos líneas claramente diferenciadas. La primera utiliza el ACP como una herramienta de los estudios de investigación en educación, mientras que la segunda línea se centra en la investigación de la enseñanza y el aprendizaje del ACP. El presente artículo se enmarca en esta última línea de investigación.

A continuación, se recogen varios artículos que aplican el ACP como herramienta de investigación en educación. En [7] se utiliza ACP como herramienta de validación al evaluar el aprendizaje autónomo de cursos online. Otro caso se encuentra en [16] donde el ACP se usa para elaborar predicciones tempranas del rendimiento académico de los alumnos. En otras ocasiones [15] el ACP se utiliza para medir la carga cognitiva al resolver problemas. Por último, también se ha utilizado este método [26] para validar los marcos de evaluación que surgen a partir de la analítica de aprendizaje. En este caso se utiliza ACP como herramienta de evaluación en la analítica de aprendizaje.

La línea enfocada en la enseñanza y el aprendizaje del ACP recoge estudios relacionados con la metodología docente utilizada. Por ejemplo, en [28] se enseña ACP a través de ejemplos reales utilizando software libre. Concretamente utilizan bibliotecas del lenguaje de programación R para explicar los casos de uso más comunes del ACP a estudiantes de Química. En [12], se elabora una introducción detallada del ACP dirigida a resolver los problemas de comprensión de los meteorólogos en este punto. En [29], se identifican los problemas de comprensión del alumnado al afrontar el ACP. En este último artículo se propone obviar el álgebra matricial y los lemas de maximización a través de la estadística multivariante. Nuestro artículo, se enmarca en esta línea de investigación. Como en [3] nuestra experiencia docente trata de dar una visión global de la lógica del método de ACP, buenas prácticas y errores que se cometen al aplicarlo.

II-C. Educación en Ciberseguridad: marco NICE.

Nuestra experiencia docente se basa en el marco de educación en Ciberseguridad NICE [23]. Este establece como identificar, desarrollar, reclutar y mantener talento en Ciberseguridad. Algunas organizaciones y sectores utilizan este recurso al desarrollar sus herramientas y trabajos. En concreto aquellas que tengan como finalidad guiar el desarrollo, la planificación, el entrenamiento y la educación de sus profesionales.

El objetivo del marco NICE es dar una referencia fundamental para formar trabajadores que hagan frente a las necesidades existentes en materia de ciberseguridad veasé [18]. Esto se consigue mediante el establecimiento de bloques conceptuales mínimos. A partir de estos bloques, las organizaciones son capaces de describir el trabajo en ciberseguridad atendiendo a tres puntos:

1. Categoría,
2. área de especialidad
3. y rol de trabajo.

Gracias a este marco se establece una comunicación consistente entre las organizaciones y el sector de cara a la educación en ciberseguridad, el entrenamiento y el desarrollo de los futuros trabajadores.

El principal promotor de este documentos es el NIST, aunque también han intervenido otros departamentos y agencias estadounidenses. Está dirigido a:

- empresas que quieran contratar profesionales de la Ciberseguridad para identificar el tipo de profesional que necesitan, del que carecen y ofertar puestos que sean consistentes con el lenguaje que se utiliza en Ciberseguridad,
- profesionales presentes y futuros de la Ciberseguridad que quieran conocer los últimos avanzes de conocimiento y habilidades requeridas para la Ciberseguridad,
- organizaciones que formen profesionales de la Ciberseguridad,
- aquellos que diseñen de planes de estudios para desarrollar certificaciones, grados y currículos que cubran los estándares de la ciberseguridad
- y consultoras tecnológicas que quieran ofrecer soluciones de seguridad.

El estándar NICE plantea cómo describir la educación en ciberseguridad. Se describen los objetivos a alcanzar a través de unos Estándares de Tareas, y los conocimientos que se deben obtener para alcanzar dichos objetivos a través de estándares de Conocimientos y Habilidades. Se ha utilizado el estándar NICE en las tablas I, II y III para describir el conocimiento relacionado con el ACP.

Anteriormente el estándar NICE se ha utilizado en los programas[1] *CARE:Cybersecurity in Action, Research & Education, NICCS Education and Training Catalog y Palo Alto Networks Cybersecurity Academy Curriculum* para elaborar planes docentes para los profesionales de ciberseguridad del mañana. Otros autores lo aplican a título individual e.g. recoge una experiencia metodológica construida a partir del marco NICE para una asignatura práctica de ciberseguridad de "ataque/defensa".

Son muchas las publicaciones que señalan el marco como el futuro de la docencia en Ciberseguridad y la necesidad de integrarlo de forma total a la formación en ciberseguridad [4], [17], [19], [21].

II-D. Aprendizaje basado en proyectos

El **Aprendizaje basado en proyectos** (ABP) [10] es una modalidad de enseñanza aprendizaje centrada en tareas. Los participantes de la experiencia negocian un producto final con el docente. El aprendizaje individual y autónomo se promueve mediante objetivos. El estudiante se responsabiliza así de su propio aprendizaje descubriendo estrategias en el proceso. Este protagonismo que adquiere el estudiante trabaja su espíritu crítico y le saca del rol pasivo que en la metodología clásica adquiere el estudiante. Además se trata de una metodología inclusiva ya que es el propio estudiante el que adecua el conocimiento a sus aptitudes.

Los objetivos del ABP son:

[1]Para más ejemplos veasé https://www.nist.gov/itl/applied-cybersecurity/nice/nice-framework-resource-center/education-and-training-provider

Tabla I
ESTÁNDAR DE CONOCIMIENTOS NICE REQUERIDO PARA COMPLETAR LA ASIGNATURA MATEMÁTICAS PARA CIBERSEGURIDAD II - ANÁLISIS DE DATOS Y REDES

Code	Standard Alignment
K0005	Knowledge of cyber threats and vulnerabilities.
K0006	Knowledge of specific operational impacts of cybersecurity lapses.
K0054	Knowledge of current industry methods for evaluating, implementing, and disseminating information technology (IT) security assessment, monitoring, detection, and remediation tools and procedures utilizing standards-based concepts and capabilities.
K0059	Knowledge of new and emerging information technology (IT) and cybersecurity technologies.
K0016	Knowledge of computer programming principles
K0028	Knowledge of organization's evaluation and validation requirements.
K0039	Knowledge of cybersecurity and privacy principles and methods that apply to software development.
K0050	Knowledge of local area and wide area networking principles and concepts including bandwidth management.
K0060	Knowledge of operating systems.
K0068	Knowledge of programming language structures and logic.
K0079	Knowledge of software debugging principles.
K0082	Knowledge of software engineering.
K0139	Knowledge of interpreted and compiled computer languages.

Tabla II
ESTÁNDAR DE TAREAS NICE REQUERIDO PARA COMPLETAR LA ASIGNATURA MATEMÁTICAS PARA CIBERSEGURIDAD II - ANÁLISIS DE DATOS Y REDES

Code	Standard Alignment
T0371	Establish acceptable limits for the software application, network, or system.
T0265	Assure successful implementation and functionality of security requirements and appropriate information technology (IT) policies and procedures that are consistent with the organization's mission and goals.
T0026	Compile and write documentation of program development and subsequent revisions, inserting comments in the coded instructions so others can understand the program.
T0111	Identify basic common coding flaws at a high level.
T0455	Develop software system testing and validation procedures, programming, and documentation.
T0500	Modify and maintain existing software to correct errors, to adapt it to new hardware, or to upgrade interfaces and improve performance.
T0077	Develop secure code and error handling.
T0171	Perform integrated quality assurance testing for security functionality and resiliency attack.
T0176	Perform secure programming and identify potential flaws in codes to mitigate vulnerabilities.
T0324	Direct software programming and development of documentation.
T0303	Identify and leverage the enterprise-wide version control system while designing and developing secure applications.
T0236	Translate security requirements into application design elements including documenting the elements of the software attack surfaces, conducting threat modeling, and defining any specific security criteria.
T0117	Identify security implications and apply methodologies within centralized and decentralized environments across the enterprise's computer systems in software development.
T0228	Store, retrieve, and manipulate data for analysis of system capabilities and requirements.
T0371	Establish acceptable limits for the software application, network, or system.

1. Formar estudiantes capaces de interpretar el entorno en el que se desenvuelven.
2. Motivar la adquisición de conocimientos utilizando ejemplos complejos y concretos.
3. Producir nuevos conocimientos de forma autónoma y no supervisada.
4. Desarrollar nuevas habilidades útiles para el mundo real.
5. Enfocar bien el resultado del aprendizaje al proponer objetivos claros.
6. Trabajar de forma colaborativa.

El aprendizaje se centra en una problemática real que no es sencilla y que involucra varias áreas de conocimiento. A través de la resolución de esta tarea el estudiante saca sus propias conlcusiones. Otra característica fundamental de esta estrategia metodológica es que se favorece la colaboración entre todos los participantes del proceso enseñanza-aprendizaje a fin de conseguir un conocimiento compartido y distribuido entre los miembros. A menudo, se hace uso de la tecnología actual como ordenadores, internet y otras herramientas cognitivas. Los contenidos se alargan más en el tiempo ya que a un mismo proyecto se le suele dedicar desde varias semanas a un mes. Se hace hincapié en la presentación de un producto o resultado final. También es común que al comenzar el proceso el profesorado no facilite ningún tipo de material para que los estudiantes puedan realizar el proyecto.

El ABP es un enfoque útil para la formación en Ciberseguridad. El aprendizaje autónomo que se realiza en el marco de los retos de la tecnología actual lo acerca mucho a la experiencia real de un técnico o investigador de ciberseguridad. Esta inmersión en las técnicas que se utilizan para resolver los problemas otorga un conocimiento más profundo a los extudiantes y mantiene su posterior interés.

Existen numerosos ejemplos de experiencias didácticas de formación en ciberseguridad que se basan en ABP. En [27] se recoge una experiencia que tuvo lugar en la que los estudiantes analizaron parte de la red de su campus tecnológico descubriendo numerosas fallas y exploits. Otra experiencia parecida es [20] centrada en el análisis de los estándares actuales de conexiones y redes. El artículo [5] describe una asignatura de asesoría de vulnerabilidades de seguridad que sigue la metodología de ABP. El proyecto final consistía en que los alumnos presentasen un producto comercial a una empresa externa. Otras autoras [6] utilizan el ABP para desarrollar propuestas metódologicas orientadas a solventar los problemas de comprensión del estudiante en la asignatura

Tabla III
ESTÁNDAR DE HABILIDADES Y COMPETENCIAS NICE REQUERIDO PARA COMPLETAR LA ASIGNATURA MATEMÁTICAS PARA CIBERSEGURIDAD II - ANÁLISIS DE DATOS Y REDES

Code	Standard Alignment
S0034	Skill in discerning the protection needs (i.e., security controls) of information systems and networks.
S0367	Skill to apply cybersecurity and privacy principles to organizational requirements (relevant to confidentiality, integrity, availability, authentication, non-repudiation).
A0090	Ability to identify external partners with common cyber operations interests.
A0170	Ability to identify critical infrastructure systems with information communication technology that were designed without system security considerations.
A0118	Ability to understand technology, management, and leadership issues related to organization processes and problem solving.
A0101	Ability to recognize and mitigate cognitive biases which may affect analysis.
A0106	Ability to think critically.
A0108	Ability to understand objectives and effects.
A0116	Ability to prioritize and allocate cybersecurity resources correctly and efficiently.
S0014	Skill in conducting software debugging.
S0017	Skill in creating and utilizing mathematical or statistical models.
S0060	Skill in writing code in a currently supported programming language (e.g., Java, C++).
A0021	Ability to use and understand complex mathematical concepts (e.g., discrete math).

de criptografía.

III. SECCIÓN EXPERIMENTAL

III-A. Participantes de la experiencia

La experiencia se ha implementado sobre cuatro estudiantes de género másculino del segundo curso del *Máster de Investigación en Ciberseguridad* de la *Universidad de León*. De esos cuatro estudiantes tres cursaban la modalidad presencial, y sólo uno la modalidad online. La asignatura de *Matemáticas para Ciberseguridad II - Análisis de datos y redes* tiene lugar durante el primer semestre del curso escolar 2020/2021. Notesé que uno de los alumnos no podía atender a las sesiones debido a incompatibilidades con su horario laboral. Este alumno tenía acceso a los vídeos de las sesiones en el Moodle de la asignatura.

III-B. Descripción de la asignatura.

La asignatura en la que se ha llevado a cabo la experiencia es *Matemáticas para Ciberseguridad II - Análisis de datos y redes* del *Máster de Investigación en Ciberseguridad* . Se trata de una asignatura optativa del primer semestre de segundo de máster. Es una asignatura de de 6 créditos ECTS. Este programa tiene un cáracter mixto: se desarrollan competencias profesionales que son útiles para la industria y otras, de tipo académico, que permiten poder llevar a cabo la labor investigadora. El plan de estudios recoge la fuerte naturaleza interdisciplinar de la ciberseguridad. En este máster, se obtienen conocimientos sobre ciberseguridad en sistemas operativos, ciberseguridad en redes, seguridad en el software, seguridad de sistemas ciber-físicos, aspectos humanos de la ciberseguridad, implicaciones jurídicas de la ciberseguridad, auditorías de seguridad, análisis forense y criptografía.

Los contenidos estan separados en tres bloques. El primero trata sobre Álgebra Lineal en Análisis de Datos, el segundo atiende a cuestiones de Teoría de Grafos y Redes y el tercero se centra en las aplicaciones de los anteriores. Estas aplicaciones estan relacionadas con: las redes Sociales, la computación en la nube, las infraestructuras críticas y los sistemas confiables controlados por el usuario.

Las competencias de la asignatura se recogen en la tabla IV.

Los resultados de aprendizaje se encuentran en la tabla V.

III-C. Evaluación de la propuesta

Para evaluar la experiencia se han valorado la satisfacción de los estudiantes con la metodología docente y su proceso de aprendizaje.

Los estudiantes han rellenado una encuesta de calidad basada en SEEQ con la cual se valora la calidad docente y el grado de satisfacción de los estudiantes con el profesorado.

Para evaluar el proceso de aprendizaje se han propuesto trabajos a lo largo del curso que posteriormente han sido evaluados de acuerdo a cuatro métricas. Resultado correcto, claridad del código, claridad en la expresión del estudiante y presentación del documento.

IV. RESULTADOS DE LA EXPERIENCIA

IV-A. Metodología propuesta: proyectos detallados

Nuestra experiencia docente se basa en ABP. En la presentación de la asignatura los alumnos nos transmitieron su dificultad para comprender el álgebra lineal. Motivados por este hecho, propusimos un proyecto final donde cobrara importancia el ACP. El proyecto final consistiría en el cálculo de las *eigenfaces* para aplicaciones de reconocimiento facial. Además, a lo largo del curso se propusieron otros trabajos que complementasen el final atendiendo a los objetivos, contenidos y competencias de la asignatura.

IV-A1. Proyecto I: Obtener el mejor modelo lineal que explicase un conjunto de datos: El primer proyecto consistía en explicar una variable Z a través de otras dos variables X e Y a partir de unos datos que venian dados en formato ".xls". El modelo debía ser de la forma $Z = aX + bY$ y debía ser óptimo en el sentido de los mínimos cuadrados.

Para resolver este proyecto los alumnos tuvieron libertad de elección. Un alumno utilizo *Mathlab*, otro *Mathematica* y otros dos *Python*. El tiempo que se proporcionó para resolver el enunciado fue de dos semanas.

IV-A2. Proyecto II: Obtener las direcciones principales de una nube pseudoaleatoria de puntos: En el segundo proyecto se trataba de explicar a partir de sus direcciones principales y posteriormente discutir cual era el porcentaje explicado de la varianza de los datos. En la figura 1 se puede observar el resultado de dicho análisis siendo la flecha roja la primera dirección principal.

No había restricciones a la hora de que tecnología utilizar para completar esta tarea. Se facilitó a los alumnos enlaces a la libreria *Numpy* de *Python* con las funciones más útiles

Tabla IV
COMPETENCIAS DE LA ASIGNATURA MATEMÁTICAS PARA CIBERSEGURIDAD II - ANÁLISIS DE DATOS Y REDES

Tipo A	Código	Competencias Específicas
	A17082	1733E6 Entender, aplicar e investigar Matemáticas aplicadas a la ciberseguridad.
	A17095	1733EOPT3 Conocer las líneas de investigación multidisciplinar más novedosas en ciberseguridad. Estudiar problemas en ciberseguridad con las adecuadas técnicas matemáticas.
Tipo B	**Código**	**Competencias Generales y Transversales**
	B5220	1733G1 Elaborar y defender argumentos y resolver problemas dentro del área de seguridad informática y de las comunicaciones.
	B5221	1733G2 Reunir e interpretar datos relevantes dentro del área de seguridad informática y de las comunicaciones.
	B5222	1733G3 Emitir juicios sobre temas relevantes de índole social, científica o ética desde la perspectiva de la ciberseguridad.
	B5223	1733G4 Transmitir soluciones al entorno industrial y empresarial en el campo de la ciberseguridad.
	B5224	1733G5 Aprender de forma autónoma.
	B5225	1733G6 Ser capaz de desarrollar proyectos de seguridad informática y de las comunicaciones.
	A17095	1733EOPT3 Conocer las líneas de investigación multidisciplinar más novedosas en ciberseguridad. Estudiar problemas en ciberseguridad con las adecuadas técnicas matemáticas.
Tipo C	**Código**	**Competencias Básicas**
	C1	Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
	C2	Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios.
	C3	Que los estudiantes sepan comunicar sus conclusiones (y los conocimientos y razones últimas que las sustentan) a públicos especializados y no especializados de un modo claro y sin ambigüedades.
	C4	Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.
	C5	Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación

Tabla V
RESULTADOS DE APRENDIZAJE DE LA ASIGNATURA MATEMÁTICAS PARA CIBERSEGURIDAD II - ANÁLISIS DE DATOS Y REDES

Código	Resultados	Competencias
O1	Integrar, formular y comunicar juicios y conclusiones.	C2, C3
O2	Planteamiento autónomo y resolución de problemas relacionados con el análisis de datos.	C1, C4
O3	Desarrollo de ideas originales aplicando investigación basada en datos. Elaborar y defender argumentos y resolver problemas dentro del área de seguridad informática y de las comunicaciones.	A17082, A17095
O4	Reunir e interpretar datos relevantes dentro del área de seguridad informática y de las comunicaciones. Emitir juicios sobre temas relevantes de índole social, científica o ética desde la perspectiva de la ciberseguridad. Transmitir soluciones al entorno industrial y empresarial en el campo de la ciberseguridad. Aprender de forma autónoma. Ser capaz de desarrollar proyectos de seguridad informática y de las comunicaciones.	B5220, B5221, B5222, B5223, B5444, B5225

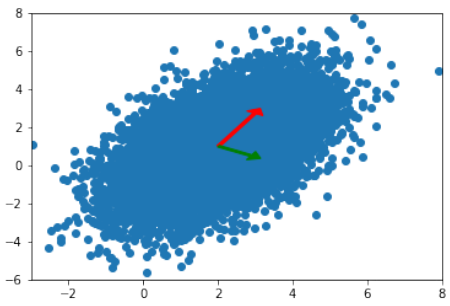


Figura 1. Primer trabajo de ACP.

para la práctica . Los 3 alumnos que completaron la tarea lo hicieron en *Python*. Trabajaron en esta tarea durante una semana.

IV-A3. Proyecto III: Comprimir una imágen con un factor de compresión p: Este tercer proyecto consistía en elaborar un algoritmo de compresión de imágenes a partir del Teorema de Eckart-Young [9]. En la figura 2 se observan distintas tasas de compresión para la misma imágen [2]. No había restricciones a la hora de que tecnología utilizar para completar esta tarea. Se facilitó a los alumnos enlaces a las libreria *Numpy* y *OpenCV* de *Python* con las funciones más útiles para la práctica. Sólo tres alumnos completaron la tarea utilizando *Python*. Este proyecto transcurrió durante tres semanas.

IV-A4. Proyecto IV: Cálculo de las eigenfaces del dataset INRIA: En este caso se les pedía a los alumnos replicar el modelo *eigenfaces*. Se trata de un método desarrollado por Sirovich y Kirby en 1987 utilizado en el problema de reconocimiento facial. Se persigue encontrar una representación de la imágen de una cara con una dimensión inferior a la de la imágen inicial. En la figura vemos en cada fila una suma ponderada entre una *eigenface* a la izqda. y la cara media de todas las caras del dataset a la derecha. Cada una de las seis filas corresponde a una de las seis primeras *eigenfaces* obtenidas a partir del dataset INRIA [3]. No había restricciones a la hora de que tecnología utilizar para completar esta tarea. Se facilitó a los alumnos enlaces a las libreria *Numpy* y *OpenCV* de *Python* con las funciones más útiles para la

[2]https://en.wikipedia.org/wiki/Lenna
[3]http://pascal.inrialpes.fr/data/human/

Figura 2. Imágenes comprimidas utilizando el Teorema de Eckart-Young

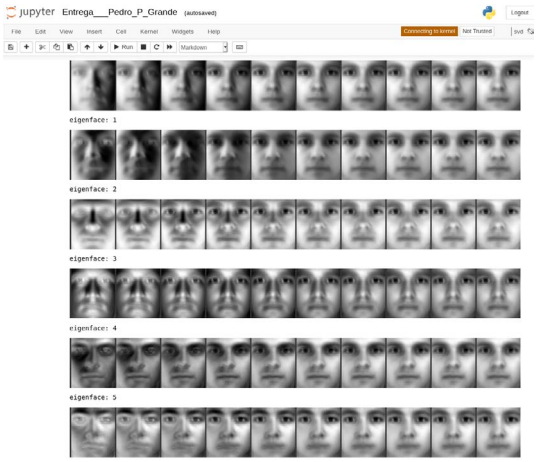


Figura 3. Suma ponderada entre cada *eigenface* (izqda.) y la cara media (dcha.)

práctica. Solo tres alumnos completaron la tarea utilizando *Python*. Este proyecto transcurrió durante un mes.

IV-A5. Proyecto V: Reconocimiento facial a partir de eigenfaces: Para obtener puntación extra se propuso elaborar un sistema de reconocimiento facial basado en las *eigenfaces* calculadas con anterioridad.

No había restricciones a la hora de que tecnología utilizar para completar esta tarea. Se facilitó a los alumnos enlaces a las libreria *Numpy* y *OpenCV* de *Python* con las funciones más útiles para la práctica. Ningún alumno completó esta tarea por ser opcional.

Para observar el grafo de dependencia entre los conocimientos de los diversos proyectos veasé la figura 4.

Los contenidos relacionados, objetivos perseguidos y competencias adquiridas en cada proyecto se pueden encontrar en la tabla VI.

Para evaluar el aprendizaje hemos calificado cada proyecto de forma independiente y uniforme. En la tabla VII se recoge la temporalización del curso y el peso de cada trabajo en la nota final.

IV-B. Resultados del aprendizaje

Los resultados de la encuesta se pueden encontrar en la tabla de la figura 5.

Las calificaciones en los proyectos se recogen en la tabla VIII.

Las calificaciones medias de la materia se encuentran en la figura 6.

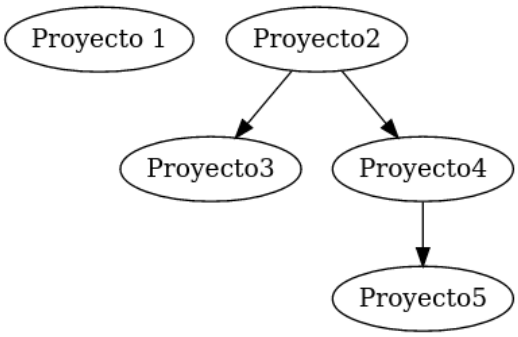


Figura 4. Dependencia entre conocimientos

Tabla VI
CONTENIDOS, OBJETIVOS Y COMPETENCIAS POR PROYECTO

Proyecto	Contenidos	Competencias	Objetivos
Proyecto I	Álgebra lineal en análisis de datos y teoría de grafos y redes	C1, C2, C3, C4, B5224	O2
Proyecto II	Álgebra lineal en análisis de datos y aplicaciones de redes sociales, infraestructuras críticas y sistemas confiables controlados por el usuario	C1, C2, C3, C4, B5220, B5221, B5222, B5223, B5224, B5225	O1, O2, O4
Proyecto III	Álgebra lineal en análisis de datos, teoría de redes y aplicaciones de computación en la nube, infraestructuras críticas	A17082, A17095, B5220, B5221, B5222, B5223, B5224, B5225, C1, C2, C3, C4	O1, O2, O4
Proyecto IV	Álgebra lineal en análisis de datos, teoría de redes y aplicaciones de redes sociales, infraestructuras críticas y sistemas confiables controlados por el usuario	A17082, A17095, B5220, B5221, B5222, B5223, B5224, B5225, C1, C2, C3, C4, C5	O1, O2, O3, O4
Proyecto V	Álgebra lineal en análisis de datos, teoría de redes y aplicaciones de redes sociales, infraestructuras críticas y sistemas confiables controlados por el usuario	A17082, A17095, B5220, B5221, B5222, B5223, B5224, B5225, C1, C2, C3, C4, C5	O1, O2, O3, O4

Los resultados de la experiencia fueron satisfactorios, trabajar el ACP mediante proyectos de programación facilita la asimilación de conceptos. En concreto, permite explicar la variabilidad de los datos a partir de las direcciones principales. Todo esto, sin tener que pagar el coste teórico que supone explicarlo de forma abstracta. También nos permite atajar los errores que más comete el alumnado al aplicar las técnicas de reducción de la dimensionalidad. Se ha conseguido motivar

Tabla VII
TEMPORALIZACIÓN DE LA ASIGNATURA MATEMÁTICAS PARA CIBERSEGURIDAD II - ANÁLISIS DE DATOS Y REDES

Proyecto	Peso en la calificación	Deadline
Proyecto I	25 %	2020/10/28
Proyecto II	25 %	2020/11/11
Proyecto III	25 %	2020/12/07
Proyecto IV	25 %	2021/01/07
Proyecto V	20 % (extra)	2021/01/12

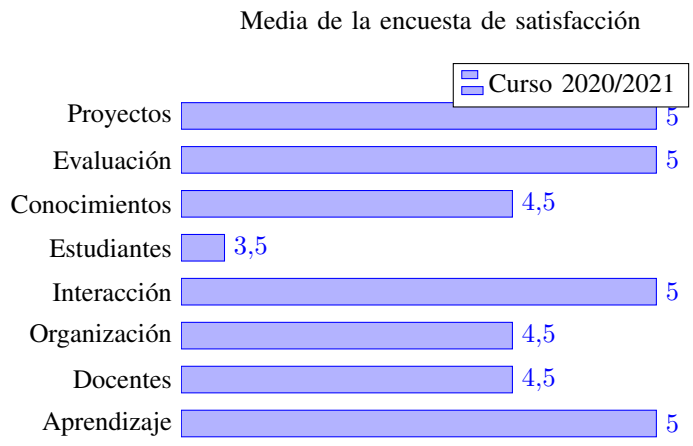


Figura 5. Encuesta de satisfacción de la asignatura Matemáticas para Ciberseguridad II - Análisis de datos y redes

Tabla VIII
CALIFICACIONES DE LA ASIGNATURA MATEMÁTICAS PARA CIBERSEGURIDAD II - ANÁLISIS DE DATOS Y REDES

Participante	Tarea 1	Tarea 2	Tarea 3	Tarea 4
P1	8	10	10	9.5
P2	10	10	10	10
P3	9	10	10	9
P4	7	0	0	0

a unos estudiantes que desde el primer día expusieron su rechazo a la teoría matemática subyacente y ha sido a través de proyectos que los estudiantes han trabajado con estas herramientas.El resultado es un éxito ya que una amplia mayoria de los estudiantes ha conseguido resumir las caras de un dataset en sólo 6 direcciones principales. Los estudiantes completaron el objetivo. Sin embargo, donde mejor resultados se obtuvieron fue en los proyectos 2 y 3. Esto se debe a que las entregas se planificaron durante el transcurso del curso y, por lo tanto, los estudiantes expresaron su dificultades durante las sesiones. Por otra parte el trabajo final se debía entregar una vez acabadas las sesiones. Otro aspecto a tener en cuenta

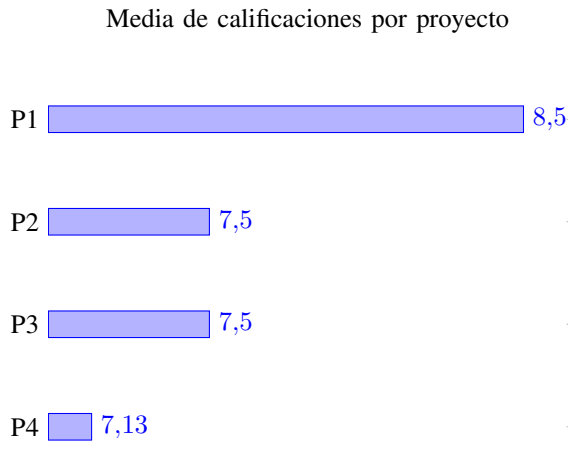


Figura 6. Calificaciones medias de la asignatura Matemáticas para Ciberseguridad II - Análisis de datos y redes

es que los trabajos voluntarios no surten efecto. Se puede observar que el último proyecto no lo realizó nadie, así que , en futuras implementaciones de la experiencia, el proyecto se hará obligatorio.

V. CONCLUSIONES.

Constatamos que otra forma de explicar matemáticas en un curso de informática de máster es posible. El desarrollo teórico debe estar siempre acompañado de ejemplos que sean cercanos a la especialización del máster. A pesar de que dichos ejemplos aparezcan en las bibliotecas más populares en ocasiones es necesario implementarlos uno mismo.

Finalmente, remarcamos el carácter transversal de las herramientas matemáticas como ACP. Esto permite transmigrar este tipo de experiencias a estudios especializados de Máster en otras ramas de Ingeniería o Ciencias.

En cuanto al trabajo futuro hace falta estudiar la escalabilidad de esta experiencia. El reto consiste en conseguir que una clase de grado lleve a cabo este tipo de metodología de forma satisfactoria. Para ello proponemos una experiencia híbrida entre la metodología clásica y el ABP.

REFERENCIAS

[1] Abokifa, Ahmed A., et al: "Detection of cyber physical attacks on water distribution systems via principal component analysis and artificial neural networks." en *World Environmental and Water Resources Congress 2017*, pp. 676-691, 2017.

[2] Bajaj, R. K., y Guleria, A.: "Dimensionality reduction technique in decision making using Pythagorean fuzzy soft matrices." en *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, vol. 13, n. 3, pp. 406-413, 2020.

[3] Brems, Matt: "A one-stop shop for principal component analysis." en *Medium towards Data Science*, vol. 17, (2017).

[4] Caulkins, Bruce D., et al. "Cyber workforce development using a behavioral cybersecurity paradigm." en *2016 International Conference on Cyber Conflict (CyCon US)*, pp. 1-6, 2016.

[5] Conklin, A. y White, G.: "A graduate level assessment course: A model for safe vulnerability assessment." en *Proc. 9th Colloquium for Information Systems Security Education (CISSE)*, vol. 9, pp. 109–114, July 2005.

[6] DeCastro-García, N. y Suárez Corona, A.: "Portafolio como herramienta educativa en Ciberseguridad: Aprendizaje por aproximación en Criptografía," en *Actas JNIC*, 2017.

[7] Dongho, Kim, et al: "Learning analytics to support self-regulated learning in asynchronous online courses: A case study at a women's university in South Korea." en *Computers & Education 127*, vol. 127, pp. 233-251, 2018.

[8] Dorr, B. J. : "The NIST data science initiative," en *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1-10, 2015.

[9] Eckart C. y Young G.: "The approximation of one matrix by another of lower rank" en *Psychometrika*, vol. 1, n. 3, pp. 211-218, 1936.

[10] Gary, Kevin. "Project-based learning." en *Computer*, vol. 48, n. 9, pp. 98-100, 2015.

[11] Gudivada, Venkat, Amy Apon, y Junhua Ding: "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations." en *International Journal on Advances in Software* vol. 10, n. 1, pp. 1-20, 2017.

[12] Jolliffe, I. T.: "Principal component analysis: a beginner's guide—I. Introduction and application." en *Weather*, vol. 45, n. 10, pp. 375-382, 1990.

[13] Jolliffe, I. T.:"Principal component analysis", en *series: Springer series in statistics*, vol. 1, n. 2, 2002.

[14] Kravchik, Moshe y Asaf Shabtai: "Efficient cyber attack detection in industrial control systems using lightweight neural networks and pca." en *IEEE Transactions on Dependable and Secure Computing*, pp. 1-1, 2021.

[15] Larmuseau, Charlotte, et al: "Multimodal learning analytics to investigate cognitive load during online problem solving." en *British Journal of Educational Technology* vol. 51, n. 5, pp. 1548-1562, 2020.

[16] Lu, Owen HT, et al: "Applying learning analytics for the early prediction of Students' academic performance in blended learning." en *Journal of Educational Technology & Society* vol. 21, n. 2, pp. 220-232, 2018.

[17] Martini, Ben, and Kim-Kwang Raymond Choo. "Building the next generation of cyber security professionals." en *SSRN*, 2014.
[18] McGettrick, Andrew. "Toward effective cybersecurity education." en *IEEE Security & Privacy*, vol. 11, n. 6,pp. 66-68, 2013.
[19] Ministr, Jan, Tomáš Pitner, y Nikola Šimková. "Cybersecurity Qualifications." en *IT for Practice*, p. 141, 2018.
[20] New Jersey Institute of Technology. "Cybersecurity Real World Connections Summer Boot Camp at NJIT." en *Online* 2016.
[21] Paulsen, Celia, et al. "NICE: Creating a cybersecurity workforce and aware public." en *IEEE Security & Privacy*, vol. 10, n. 3, pp. 76-79, 2012.
[22] Petrenko, S. A., y Makoveichuk, K. A.: "Big data technologies for cybersecurity." en *CEUR workshop*, pp. 107-111, 2017.
[23] Petersen, Rodney, et al: "Workforce Framework for Cybersecurity (NICE Framework)". en *NIST Special Publication (SP) 800-181 National Institute of Standards and Technology*, 2020.
[24] Reid, M. K. y K. L. Spencer: "Use of principal components analysis (PCA) on estuarine sediment datasets: the effect of data pre-treatment." en *Environmental pollution*, vol. 157, n. 8-9, pp. 2275-2281, 2009.
[25] Sakurada, Mayu y Takehisa Yairi: "Anomaly detection using autoencoders with nonlinear dimensionality reduction." en *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, pp. 4-11, 2014.
[26] Scheffel, Maren, et al: "The proof of the pudding: examining validity and reliability of the evaluation framework for learning analytics." en *European conference on technology enhanced learning*, pp. 194-208, 2017.
[27] Sherman, Alan T., et al. "Project-Based Learning Inspires Cybersecurity Students: A Scholarship-for-Service Research Study." en *IEEE Security & Privacy*, vol. 17, n. 3, pp. 82-88, 2019.
[28] Sidou, Laís Feltrin y Endler Marcel Borges: "Teaching Principal Component Analysis Using a Free and Open Source Software Program and Exercises Applying PCA to Real-World Examples." en *Journal of Chemical Education*, vol. 97, n. 6, pp. 1666-1676, 2020.
[29] Westfall, Peter H., Andrea L. Arias, y Lawrence V. Fulton: "Teaching principal components using correlations." en *Multivariate behavioral research*, vol. 52, n. 5, pp. 648-660, 2017.
[30] Zhang, Tonglin, y Baijian Yang: "Big data dimension reduction using PCA." en *2016 IEEE international conference on smart cloud (SmartCloud). IEEE*, vol. 1, n. 1, pp. 152-157, 2016.

Sesión de Formación e Innovación Educativa
http://doi.org/10.18239/jornadas_2021.34.66

Curso de Especialización en Ciberseguridad, ¿están preparados nuestros docentes?

Francisco José de Haro Olmo
Dpto. Informática.
Universidad de Almería
04120 Almería
Orcid:0000-0003-3130-0877
fdo730@inlumine.ual.es

Ángel Jesús Varela-Vaca
Dpto. Lenguajes y Sistemas Informáticos.
Universidad de Sevilla
ETS Ingeniría Informática
Orcid:0000-0001-9953-6005
ajvarela@us.es

José Antonio Álvarez Bermejo
Dpto. Informática
Universidad de Almeria
04120 Almería
Orcid:0000-0002-5815-7858
jaberme@ual.es

***Resumen*—La aparición del nuevo título de Formación Profesional, Curso de Especialización en *"Ciberseguridad en entornos de tecnologías de la información"*, establece el punto de partida sobre los conocimientos del profesorado en cuestiones de Ciberseguridad. Para ello hemos realizado un estudio sobre los conocimientos del profesorado que imparte docencia en la formación profesional, basado en la propuesta curricular en ciberseguridad de la guías *ACM/IEEE/AIS SIGSEC/IFIP Cybersecurity*, con el objetivo de proponer un itinerario formativo en Ciberseguridad para el profesorado, de forma que estén en disposición de ofrecer una mejor respuesta y mayor calidad ante el proceso de formación de los profesionales del futuro en dicha materia. En primer lugar hemos desarrollado un estudio basado cuestionarios, a través del cual se han presentado las unidades de conocimiento en materia de Ciberseguridad al profesorado de Andalucía y sobre las que han realizado la valoración de su conocimiento en dicha materia. Se presenta un análisis cuantitativo de los resultados obtenidos priorizando las necesidades formativas. Como conclusión de nuestro estudio hemos propuesto la elaboración de un itinerario formativo para el profesorado basado en las diez unidades de conocimiento.**

***Index Terms*—Ciberseguridad, Profesorado, Formación Profesional, Currícula**

Tipo de contribución: *Formación e innovación educativa*

I. Introducción

La necesidad de profesionales expertos (técnicos especialistas, ingenieros, etc.) en materia de Ciberseguridad se acrecienta día tras en el mercado laboral, donde las empresas están tomando concienciando de la necesidad de contar con personal cualificado ante los nuevos retos, problemas y amenazas emergentes ante los sistemas informáticos [2], [3], [4].

Las Cyber kill chains (cadenas de ataque) [1][5] utilizadas por los atacantes para penetrar en los sistemas son cada vez más sofisticadas, demostrando la necesidad de organizaciones y administraciones de contar con expertos en Ciberseguridad. Además, en la nueva era de la de digitalización, las nuevas tecnologías como el Robotic Process Automation (RPA), el procesamiento en la nube, Big Data, y entornos como la Industria 4.0 hace que la Ciberseguridad adquiera mayor relevancia.

A nivel universitario ya existen titulaciones en Ciberseguridad que abarcan un completo abanico de contenido para formar a los profesionales de la Ciberseguridad que el mercado laboral está demandando. Sin embargo, no se consigue cubrir todos los perfiles requeridos en materia de Ciberseguridad [6]. Estos nuevos perfiles profesiones relacionadas con la Ciberseguridad, se encuentran a todos los niveles de formación y especialización, con distintos niveles de cualificación entre los que los profesionales (técnicos superiores) procedentes de la Formación Profesional (FP) son un referente más.

Estas nuevas necesidades de especialización que aparecen en el mercado laboral y han dado lugar a la publicación de nuevas titulaciones [7]. El objetivo de estas titulaciones es cubrir las deficiencias del mercado dotando de profesionales técnicos especialistas con alta especialización, concretamente en materia de Ciberseguridad. Por tanto, es de vital importancia considerar la actualización formativa de los docentes y la concienciación de la importancia de las competencias digitales [9] de quienes imparten docencia en estas titulaciones.

En el contexto de las 6 titulaciones de la familia "Informática y Comunicaciones", tan sólo existen un par de módulos profesionales dedicados a la seguridad informática: (a) **Seguridad Informática** en el Título de Técnico en Sistemas Microinformáticos y Redes; y el otro es (2) **Seguridad y Alta Disponibilidad** en el Título de Técnico Superior en Administración de Sistemas Informáticos en Red. Si bien es cierto que los contenidos necesitan una revisión para adaptarse a la realidad del mercado laboral, cabe destacar que el resto de titulaciones carecen de la formación necesaria en materia de Ciberseguridad, aspecto que hace más que necesaria la aparición de la nueva titulación [7] como forma de especialización en esta materia y que se accede desde cualquiera de los títulos de grado superior de la misma familia profesional.

Para abordar estos nuevos retos en la Formación Profesional, desde el Ministerio de Educación y Formación Profesional se ha publicado el *I Plan Estratégico de Formación Profesional del Sistema Educativo 2019-2022* [9] donde en el *Eje 7* se hace referencia a la mejora de la actualización y formación permanente del profesorado. Resaltamos los Objetivos 12 y 13:

- **Objetivo 12**. Promover la formación especializada del profesorado.
- **Objetivo 14.** Impulsar la implantación de metodologías didácticas novedosas que supongan innovación educativa.

Estos antecedentes unidos a la existencia de un título denominado **Curso de Especialización en Ciberseguridad en entornos de las tecnologías de la información** [7] y que

pretende dar respuesta rápida a las innovaciones producidas en el sector productivo y a la necesidad de formar a los futuros profesionales de la Ciberseguridad, profundizando en los conocimientos o bien ampliando las competencias propias de cada título de referencia. Por todo ello, en primer lugar, debemos indagar el nivel inicial y la necesidad de formar al profesorado en esta disciplina, Ciberseguridad. Para realizar este estudio, haremos una evaluación inicial de conocimientos en materia de Ciberseguridad a través de un cuestionario de autovaloración sobre las diferentes unidades de conocimiento integradas en el modelo ACM/IEEE/AIS SIGSEC/IFIP CYBERSECURITY CURRICULA 2017 (CSEC 2017) [10] y que consideramos que están relacionadas con este nuevo título. El objetivo final de dicho estudio será plantear un itinerario formativo para la formación de los docentes con el nuevo título en Ciberseguridad de manera que de respuesta a las necesidades reales y que permita adecuar los conocimientos y habilidades técnicas del profesorado a los retos planteados en la disciplina de la Ciberseguridad.

El articulo se ha estructurado en los siguientes apartados: en la Sección II daremos un vistazo al modelo de curriculum expuesto en CSEC 2017; a continuación, en la Sección III presentaremos el nuevo título en ciberseguridad; en la Sección IV describimos la metodología empleada para llevar a cabo el estudio; ya en la Sección V describiremos y analizaremos los datos obtenidos en el estudio; como consecuencia de los resultados obtenidos proponemos un itinerario formativo en la Sección VI, exponiendo a continuación una discusión en la Sección VII y finalmente en la Sección VIII presentaremos las conclusiones derivadas de este trabajo.

II. El modelo *CSEC 2017*

El Joint Task Force on Cybersecurity Education, conocido como CSEC 2017 JTF, surge de la necesidad de aunar fuerzas entre profesionales del entorno productivo y la sociedad científica en materia de computación para llevar a cabo el desarrollo exhaustivo de una guía curricular para la formación en materia de ciberseguridad. Esta iniciativa es el resultado de la colaboración entre las mayores sociedades internacionales de ciencias de la computación: Association for Computing Machinery (ACM), IEEE Computer Society (IEEECS), Association for Information Systems Special Interest Group on Information Security and Privacy (AIS SIGSEC), y la International Federation for Information Processing Technical Committee on Information Security Education.

Entre los destinatarios del modelo curricular propuesto se encuentran por una parte el profesorado de disciplinas basadas en la informática interesados en desarrollar o incorporar programas de Ciberseguridad, y por otra parte, la administración educativa con competencias para el desarrollo y revisión de programas formativos y cursos.

El marco curricular de ciberseguridad propuesto por CSEC 2017 JTF supone una guía para el desarrollo que incluye recomendaciones en la formación en Ciberseguridad y un esfuerzo por cualificar profesionales en ciberseguridad. Este modelo trata de alinear los programas académicos con las necesidades del mercado.

Este modelo incorpora dos tipos de habilidades a desarrollar:

- Por una parte tenemos las habilidades técnicas (hard-skills). En el modelo curricular propone tres dimensiones:
  1. Áreas de conocimiento y desglosadas en unidades de conocimiento que desprenden un total de 142 resultados de aprendizaje.
  2. Conceptos transversales que suponen los principios de la ciberseguridad: confidencialidad, integridad, disponibilidad, riesgo, pensamiento adversario y pensamiento sistémico.
  3. El desarrollo de las actuales disciplinas de computación: ciencias de la computación, ingeniería de computadores, sistemas de información, tecnología de la información e ingeniería del software.
- Por otra parte, tenemos las habilidades no-técnicas (soft-skills). Se trata de habilidades necesarias para el éxito de los profesionales de la ciberseguridad. Estaríamos hablando del trabajo en equipo, asignación adecuada de recursos, conciencia de la situación, trabajar con culturas heterogéneas. Podríamos añadir la capacidad de contabilizar, atención a los detalles, resiliencia, gestión de conflictos razonadamente, comunicación.

Este modelo trata de analizar la relación entre las demandas específicas en materia de ciberseguridad y la relación entre curriculum y el marco de referencia para el personal de ciberseguridad. La visión de este proyecto es aportar un recurso de contenido curricular en Ciberseguridad de forma integral para las instituciones académicas que pretendan ofrecer formación en ciberseguridad a nivel post-secundaria. De esta forma, CSEC 2017 JTF avanza la ciberseguridad como una nueva disciplina de computación, que se añade a las otras cinco ya existentes: Ingeniería Informática, Ciencias de la Computación, Sistemas de Información, Tecnologías de la Información, Ingeniería del Software.

III. Curso de Especialización: Ciberseguridad en entornos de las tecnologías de la información.

Este Real Decreto [7] está fundamentado en la necesidad de incluir en la FP del sistema educativo una profundización en el campo del conocimiento de los títulos de referencia o que suponen una ampliación de las competencias que se incluyen en los mismos. Estos nuevos títulos de especialización [7] pretenden *"dar una respuesta de forma rápida a las innovaciones que se produzcan en el sistema productivo así como a los ámbitos emergentes que complementen la formación incluida en los títulos de referencia"*.

En este caso en concreto, estamos ante un curso de especialización denominado *"Ciberseguridad en entornos de las tecnologías de la información"*, considerado en el nivel de FP de Grado Superior y con un total de 720 horas (43 ECTS). Pertenece a la familia profesional de Informática y Comunicaciones y se encuadra en la rama de conocimiento de Ingeniería y Arquitectura. Se accede desde los títulos de referencia de grado superior:

- Técnico Superior en Administración de Sistemas Informáticos en Red.
- Técnico Superior en Desarrollo de Aplicaciones Multiplataforma.
- Técnico Superior en Desarrollo de Aplicaciones Web.

- Técnico Superior en Sistemas de Telecomunicaciones e Informáticos.
- Técnico Superior en Mantenimiento Electrónico.

El objetivo general del título, recogido en [7]: *"consiste en definir e implementar estrategias de seguridad en los sistemas de información realizando diagnósticos de ciberseguridad, identificando vulnerabilidades e implementando las medidas necesarias para mitigarlas aplicando la normativa vigente y estándares del sector, siguiendo los protocolos de calidad, de prevención de riesgos laborales y respeto ambiental."*

Para el desarrollo de los contenidos, el curso se estructura en 6 módulos profesionales, cada uno con atribución para profesorado de los dos cuerpos docentes, tanto Profesores Técnicos de Formación Profesional (PTFP), correspondiente a la especialidad Sistemas y Aplicaciones Informáticas, como el cuerpo de Profesores de Enseñanza Secundaria (PES), de la especialidad de Informática. Las atribuciones del profesorado puede consultarse en la Tabla I.

Tabla I
ATRIBUCIÓN DE CUERPOS DOCENTES A MÓDULOS PROFESIONALES.

Cuerpo	Módulos profesionales
P.T.F.P.	Bastionado de redes y sistemas
	Puesta en producción segura
	Análisis forense informático
P.E.S.	Incidentes de ciberseguridad
	Hacking ético
	Normativa de ciberseguridad

IV. METODOLOGÍA

La metodología usada para recabar información relativa a las necesidades del profesorado está fundamentada en la realización de encuestas de opinión (Personal Opinion Surveys) [11][12]. El objetivo principal de encuesta es arrojar información de la situación actual respecto al nivel de conocimientos en materia de Ciberseguridad que presenta el profesorado de FP involucrado en los ciclos formativos correspondientes a la familia profesional de Informática y Comnunicaciones.

Definimos los pasos propuestos para llevar a cabo el estudio:

1. Seleccionar los objetivos.
2. Diseño del cuestionario.
3. Confección del cuestionario con la herramienta seleccionada.
4. Evaluar el instrumento empleado para el cuestionario.
5. Obtención de datos válidos.
6. Análisis de datos.

Selección del objetivo.

El estudio de situación respecto a las necesidades de formación que el profesorado que impartía docencia en los ciclos formativos de formación profesional de la familia profesional de Informática y Comunicaciones, profesorado objetivo de nuestro estudio.

Diseño del cuestionario.

Se diseñó un cuestionario basado en las unidades de conocimiento de CSEC 2017 JTF [10] que tenían relación con la nueva titulación en Ciberseguridad [7]. Este cuestionario está formado por 31 preguntas, cada una correspondía a una unidad de conocimiento con 5 posibles respuestas. Las respuestas están valoradas usando una escala de Likert, desde 1 (muy bajo) a 5 (muy alto). En el cuestionario se incluyó también el cuerpo docente al que pertenece el participante y la provincia en la que trabaja.

Confección del cuestionario con la herramienta seleccionada.

El instrumento seleccionado para llevar a cabo la recogida de datos fue un formulario electrónico (Google Forms).

Evaluación del cuestionario confeccionado mediante la herramienta seleccionada.

Previamente a lanzar dicho cuestionario, se llevaron a cabo varias pruebas por parte de un grupo cerrado de participantes para comprobar y verificar que la recepción de datos a través del cuestionario confeccionado era correcta.

Obtención de datos válidos.

Se envió el cuestionario mediante un enlace a través del correo electrónico, al profesorado objeto de este estudio en las 8 provincias de Andalucía. Se dejó un mes de plazo para cumplimentar el cuestionario.

Análisis de datos.

Sobre los resultados obtenidos y volcados a una hoja de cálculo, se realizó una valoración global de cada unidad de conocimiento para posteriormente centrar el análisis de resultados en la distribución de las 5 posibles respuestas (autovaloración de los niveles de conocimiento) relativas a cada cuestión planteada en el cuestionario.

En una primera fase de análisis, se determinaron aquellas unidades de conocimiento que suponían que más del 50 % de la muestra había reportado entre "1-muy bajo" y "2-bajo".

En la segunda fase, de valores totales de cada unidad de conocimiento se seleccionó por orden de menor a mayor puntuación las 10 unidades de conocimiento sobre las que el profesorado había reportado mayor desconocimiento.

V. ANÁLISIS DE RESULTADOS

Una vez obtenidos los resultados, recogidos en una hoja de cálculo, se procesan un total de 273 respuestas a cuestionarios (n=273) (ver Tabla III) procedentes de las ocho provincias de Andalucía. En el curso 2019/2020 hay un total de 1174 docentes del cuerpo de Profesores de Enseñanza Secundaria y 560 pertenecientes al cuerpo de Profesores Técnicos de Formación Profesional, lo que supone una población objetivo de 1734 [8]. Las respuestas obtenidas suponen un 15,75 %, teniendo en cuenta que no todos los docentes imparten clase en formación profesional, la tasa podría ser algo superior. Durante el presente curso 2020-2021, en los 11 centros educativos de Andalucía en los que se imparte la nueva titulación hay 56 docentes dedicados a estas enseñanzas, lo que indica que están dedicados al nuevo título un 3,23 %. Con estos datos se puede considerar la muestra como representativa y proporcionar validez a la muestra final.

La participación en el estudio ha sido heterogénea, ya que aparecen diferencias significativas en el número de cuestionarios respondidos en cada una de las diferentes provincias de Andalucía:

Referente a la distribución de la participación por cuerpos y especialidad (figura 1), los datos obtenidos son compatibles

Tabla II
PARTICIPACIÓN POR PROVINCIAS.

Provincia	% respuestas
Almería	15,9 %
Granada	4,8 %
Jaén	10,3 %
Málaga	19,6 %
Córdoba	3 %
Cádiz	4,1 %
Sevilla	38, 4 %
Huelva	4,1 %

con la dotación de personal por ciclo formativo, en la que los Profesores de Educación Secundaria es mayor a la de Profesores Técnicos de Formación Profesional. En los ciclos de grado superior la proporción es de 1PTFP/3PES mientras que en ciclos de grado medio es de 2PTFP/2PTFP.

Esta distinción del profesorado por cuerpos docentes no ha sido de mucha utilidad al no apreciar diferencias significativas entre los resultados obtenidos.

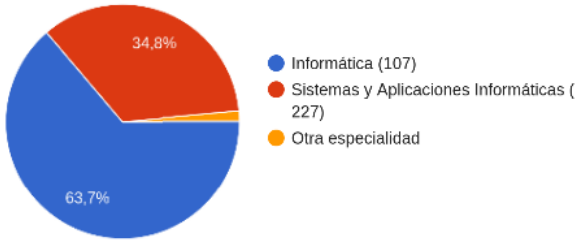


Figura 1. Participación del profesorado por cuerpo docente.

En un primer análisis de sobre cada una de las áreas de conocimiento, y con el objetivo de identificar aquellas unidades de conocimiento sobre las que existe un mayor desconocimiento, ponemos el foco de atención sobre aquellas en las que más del 50 % de la muestra había respondido con "muy bajo" o "bajo". Este dato supone que más de la mitad del profesorado reporta un nivel de muy bajo o bajo para esa unidad de conocimiento en cuestión. Esta primera identificación ya presenta la necesidad de incidir en la mejora a través de la actualización del profesorado en estos aspectos.

En la segunda fase de análisis de los resultados, se pretende determinar cuales son las 10 unidades de conocimiento, detalladas en la tabla IV, de entre las seleccionadas en la primera fase del análisis (tabla III) que obtienen menor puntuación absoluta lo que nos dará una priorización sobre las unidades de conocimiento reportadas como de mayor desconocimiento por parte de los docentes. Esta información obtenida nos servirá de ayuda para organizar las distintas necesidades formativas y el posterior diseño de un itinerario formativo en materia de ciberseguridad, de manera que se aumente el nivel de competencia del profesorado.

VI. PROPUESTA DE ITINERARIO FORMATIVO

Una vez obtenidos resultados cuantitativos sobre el problema planteado inicialmente en referencia a la necesidad de actualización del profesorado de Formación Profesional en materia de ciberseguridad, queremos proponer un itinerario formativo con estructura modular con la finalidad de dotar al profesorado de los conocimientos y habilidades necesarias para impartir docencia en esta materia. Los contenidos estarán estrechamente relacionados con los descritos en el nuevo Título [7].

Entendemos como itinerario formativo en ciberseguridad una serie de cursos o módulos formativos sobre unos temas concretos encaminados a mejorar la competencia necesaria en materia de ciberseguridad, haciendo especial hincapié en aquellos aspectos en los que se ha detectado, durante la realización de este estudio, una mayor necesidad de conocer, ya que el nivel de conocimiento reportado por los participantes en la investigación a través de sus valoraciones así lo refleja. Nuestra propuesta consta de un total de cuatro cursos y aunque recomendamos la realización secuencial y distinguiendo por cuerpos docentes (ver Figura 2). En el caso de los cursos de Hacking ético e Informática forense, en realidad no habría mayor problema en participar en todos ellos. El hecho de haber hecho distinción en estos dos cursos es por la atribución docente que se hace en cada módulo profesional a los cuerpos y especialidades recogidos en el título [7]. En este caso los PTFP de la especialidad de Sistemas y Aplicaciones Informáticas se le atribuye la docencia del módulo profesional de Informática forense, y a los PES de la especialidad de Informática el de Hacking ético. Los otros dos cursos siguientes, serían comunes a ambos cuerpos y especialidades.

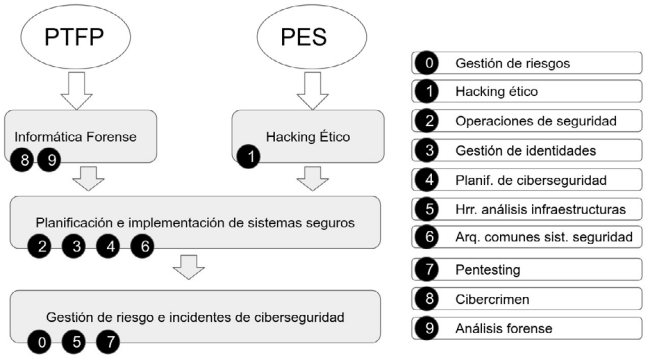


Figura 2. Itinerario formativo relacionado con unidades de conocimiento.

VI-A. Hacking ético

Objetivo: Aprender a utilizar herramientas de análisis y monitorización para detectar vulnerabilidades de sistemas aplicando técnicas de hacking ético.
Duración: 20 horas.
Contenidos:

- Fundamentos del hacking ético.
- Monitorización de sistemas informáticos.
- Ataques a sistemas informáticos y redes de comunicaciones. Escaneo y enumeración.
- Análisis de vulnerabilidades.
- Consolidación y uso de sistemas comprometidos.
- Auditoría de sistemas informáticos y redes.
- Ataques a sitos y aplicaciones web.

Recursos: Máquinas virtuales, software de monitorización de sistemas, sistemas operativos específicos, herramientas de detección y análisis de vulnerabilidades, conexión a Internet.

Tabla III
AUTOINFORME DEL PROFESORADO SOBRE ÁREAS Y UNIDADES DE CONOCIMIENTO.

Área de conoc.	Unidad de conocimiento	Muy bajo 1	Bajo 2	Inter 3	Alto 4	Muy alto 5
Seguridad de la información						
	Criptografía	0,17	0,30	0,26	0,22	0,06
	Informática forense	**0,41**	**0,32**	**0,17**	**0,07**	**0,03**
	Integridad de datos y Autenticación	0,19	0,30	0,32	0,16	0,04
	Control de acceso	0,20	0,31	0,29	0,16	0,04
	Protocolos de comunicación segura	0,16	0,29	0,31	0,16	0,07
	Privacidad de la información	0,14	0,25	0,34	0,22	0,06
	Seguridad en el almacenamiento de la información	0,14	0,27	0,28	0,25	0,07
Seguridad del software						
	Seguridad en el desarrollo del software	0,26	0,30	0,25	0,15	0,03
	Documentación y manuales	0,15	0,22	0,33	0,21	0,08
Seguridad de conexión						
	Arquitectura de sistemas distribuidos	0,30	0,31	0,24	0,11	0,04
	Arquitectura de red	0,12	0,23	0,32	0,21	0,12
	Servicios de red	0,11	0,19	0,29	0,27	0,13
	Defensa de red	0,25	0,31	0,26	0,13	0,04
Seguridad del sistema						
	Pensamiento sistémico	0,21	0,32	0,29	0,12	0,05
	Testeo de sistemas	0,25	0,33	0,26	0,13	0,03
	Arquitecturas comunes en sist. de seguridad	**0,35**	**0,35**	**0,19**	**0,08**	**0,02**
	Pentesting - Test de intrusión	**0,43**	**0,29**	**0,16**	**0,06**	**0,05**
Seguridad personal						
	Gestión de identidades	**0,32**	**0,33**	**0,21**	**0,11**	**0,02**
	Ingeniería social	0,28	0,30	0,27	0,11	0,03
	Implementación de medidas, reglas y políticas de seguridad en la organización	0,27	0,32	0,23	0,14	0,03
	Privacidad y seguridad de los datos personales	0,16	0,27	0,34	0,17	0,05
Seguridad organizacional						
	Gestión de riesgos	**0,28**	**0,33**	**0,24**	**0,12**	**0,03**
	Herramientas de análisis de infraestructuras	**0,36**	**0,33**	**0,19**	**0,09**	**0,03**
	Administración de sistemas	0,14	0,22	0,26	0,27	0,09
	Planificación de la ciberseguridad	**0,38**	**0,27**	**0,21**	**0,11**	**0,03**
	Continuidad de negocio, desastres y recuperación en la gestión de incidentes	0,28	0,33	0,22	0,14	0,03
	Operaciones de seguridad	**0,29**	**0,33**	**0,26**	**0,09**	**0,02**
Seguridad de la sociedad						
	Cibercrimen	**0,38**	**0,32**	**0,20**	**0,08**	**0,01**
	Legislación normativa sobre seguridad informática	0,27	0,33	0,27	0,09	0,03
	Privacidad de la información	0,19	0,32	0,30	0,14	0,04
	Hacking ético	**0,36**	**0,29**	**0,19**	**0,11**	**0,04**

Tabla IV
UNIDADES DE CONOCIMIENTO QUE OBTIENEN MENOR VALORACIÓN ABSOLUTA

Unidad de conocimiento	Valoración
Informática forense	544
Cibercrimen	546
Pentesting - Test de intrusión	548
Arquitecturas comunes en sistemas de seguridad	562
Herramientas de análisis de infraestructuras	570
Planificación de ciberseguridad	576
Gestión de identidades	586
Operaciones de seguridad	592
Hacking ético	596
Gestión de riesgos	617

VI-B. Informática forense

Objetivos: Aplicar metodologías de análisis forense distinguiendo las fases que intervienen en el proceso y elaboración de documentación de reporte.
Duración: 20 horas.
Contenidos:

- Fundamentos legales.
- Cibercrimen.
- Estándares.
- Fases del análisis forense. Metodología.
- Proceso de investigación. Adquisición, conservación y custodia de evidencias.
- Análisis de evidencias en sistemas Windows y Linux.
- Análisis forense de Navegadores de Internet.
- Análisis forense de correo electrónico.
- Análisis forense de dispositivos móviles.
- Análisis forense en cloud.
- Análisis forense en entornos IoT.
- Elaboración de informes periciales.

Recursos: Máquinas virtuales, software específico de investigación y análisis forense (Kits de análisis forense y herramientas de fines específicos), conexión a Internet.

VI-C. Planificación e implementación de sistemas seguros

Objetivos: Planificar y diseñar la securización de sistemas informáticos considerando los dispositivos y entornos disponibles.
Duración: 30 horas.
Contenidos:

- Diseño de plan de seguridad.
- Sistemas de acceso y credenciales.
- Diseño de redes seguras.
- Configuración de dispositivos y sistemas.

- Análisis de infraestructuras.
- Implementación de hardening en aplicaciones web.

Recursos: Máquinas virtuales, software específico de análisis y monitorización de sistemas y redes, conexión a Internet.

VI-D. Gestión de riesgos e incidentes de ciberseguridad

Objetivos: Desarrollar planes de prevención y concienciación en ciberseguridad estableciendo normas y medidas de protección.
Duración: 30 horas.
Contenidos:

- Planes de prevención y concienciación.
- Auditoría de seguridad.
- Investigación de incidentes.
- Monitorización. Detección de incidentes.
- Implementación de medidas.
- Documentación e informes.
- Gestión de conformidad (compliance).

Recursos: Máquinas virtuales, software de auditoría de seguridad informática y de investigación de evidencias así como monitorización de sistemas, documentación específica sobre compliance y conexión a Internet.

VII. Discusión

Aunque se esperaba una formación específica para el profesorado de Andalucía con atribución docente en esta nueva titulación sobre ciberseguridad por parte de la administración educativa, finalmente no fue posible llevarla a cabo antes de que iniciara el curso académico con unos resultados adecuados. Se organizó una actividad formativa de 40 horas de duración entre los meses de octubre y noviembre, la cual no consiguió dar respuesta a las necesidades formativas de los docentes debido a la poca profundidad en el tratamiento de los contenidos y la falta de especialización en cada uno de los contenidos propuestos y relacionados con la nueva titulación. Posteriormente se pusieron en marcha dos cursos, a pesar de que el profesorado manifestaba preferencia por realizar un curso por módulo profesional, paralelos en el tiempo y de 48 horas de duración cada uno, agrupando los contenidos de los seis módulos profesionales que componen la nueva titulación en ciberseguridad. Por una parte - CIBERSEGURIDAD: PUESTA EN PRODUCCIÓN SEGURA. INCIDENTES.BASTIONADO - y por otra - CIBERSEGURIDAD: FORENSE. HACKING ÉTICO. NORMATIVA - ambos cursos en modalidad online y comprendidos entre los meses de noviembre de 2020 y marzo de 2021. Pero la abundancia y complejidad de los contenidos en todo su conjunto, junto con la dificultad de encontrar empresas con dilatada experiencia en el sector de la ciberseguridad, que a su vez estuvieran dispuestas a realizar la formación del profesorado en estos plazos, dio lugar al retraso en el desarrollo del mismo. El hecho de no conseguir dar solución eficaz a las necesidades formativas del profesorado a nivel regional en toda Andalucía, ha llevado a planificar más acciones formativas para tratar de dar una respuesta adicional y conseguir dotar al profesorado de los recursos necesarios en materia de ciberseguridad, esta iniciativa destinada al profesorado de la provincia de Cádiz, ha llevado a organizar dos actividades formativas sobre ciberseguridad, una de ellas de HACKING ÉTICO, de 20 horas de duración, y una segunda actividad sobre BASTIONADO DE REDES Y SISTEMAS, de 25 horas. Una alternativa posible, pero no contemplada hasta la fecha, habría sido acudir a las universidades donde se imparten algunos de los cursos de máster en ciberseguridad, proporcionando una formación validada. Todo el profesorado implicado en la docencia de esta nueva titulación está haciendo un gran esfuerzo por conseguir la formación que consideran que necesitan (mediante sus propios medios y recursos en muchos casos) para estar a la altura de las circunstancias y del nivel que el mercado laboral espera de los futuros profesionales. Cabe realizar algunas preguntas que nos ayuden a la reflexión. ¿Cuál es el coste que está dispuesto a asumir una organización educativa por no dotar de la formación adecuada a sus profesionales docentes? ¿Repercute de algún modo en el aula los éxitos y fracasos en la formación del profesorado? ¿Es eficaz y eficiente la Red de Formación del Profesorado en Andalucía ante los retos que supone implantar una nueva titulación de Formación Profesional? ¿Existe la colaboración necesaria dentro de la propia Administración para materializar estos proyectos?

VIII. Conclusiones

A través de este estudio se ha indagado en el nivel de conocimientos sobre Ciberseguridad entre el profesorado de FP de de la familia profesional de Informática y Comunicaciones en Andalucía, relacionando la propuesta de diseño curricular en ciberseguridad CSEC 2017 JTF, con la necesidad de actualización y formación del profesorado susceptible de estar implicado en la nueva titulación en ciberseguridad: Curso de Especialización en Ciberseguridad en entornos de las tecnologías de la información.

Se ha obtenido la relación de unidades de conocimiento, según el modelo CSEC2017 JTF sobre las que se recomienda implementar un itinerario formativo de forma prioritaria, a ser posible antes de la implantación del nuevo título formativo. Esto implica que el profesorado relacionado con estas nuevas titulaciones debe llevar a cabo un esfuerzo adicional para la actualización de sus competencias profesionales específicas, con el objeto de desarrollar planes de formación relacionados con la ciberseguridad.

Se propone un itinerario formativo para ambos cuerpos docentes, Profesores Técnicos de Formación Profesional (PTFP) de la especialidad de Sistemas y Aplicaciones Informáticas así como de Profesores de Enseñanza Secundaria (PES) de la especialidad de Informática, acorde con las módulos profesionales, incluidos en la nueva titulación, sobre la que tienen competencia docente.

De los resultados desprendidos en este estudio, estaríamos en disposición de planificar la formación necesaria de los profesionales docentes, de forma que garantice una enseñanza de calidad para los futuros profesionales en el campo de la ciberseguridad que el sector productivo está demandando y que, a fecha de hoy, no consigue dar solución a todas sus necesidades de personal especializado.

Agradecimientos

A todos los profesionales docentes dedicados a la Formación Profesional y que han contribuido a la realización de este trabajo. Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia y Tecnología de España a

través del project ECLIPSE (RTI2018-094283-B-C33), y la Junta de Andalucia mediante el projecto the METAMORFOSIS, los fondos European Regional Development Fund (ERDF/FEDER).

Referencias

[1] Yadav, T., Rao, A.M.: Technical aspects of cyber kill chain. In: J.H. Abawajy,1086S. Mukherjea, S.M. Thampi, A. Ruiz-Martínez (eds.) Security in Computing1087and Communications, pp. 438–452. Springer International Publishing, Cham1088 (2015).

[2] B. Ramos, "Se necesitan urgentemente expertos en ciberseguridad: ¿Qué estudiar para ser uno de ellos?", en *EL PAÍS, 2020*. [Online]. Available: https://elpais.com/economia/2019/01/14/actualidad/1547486152_048652.html. [Accedido: 03-Mar-2020].

[3] Diario Siglo XXI, "España necesita profesionales de la ciberseguridad", 2020. [Online]. Available: http://www.diariosigloxxi.com/texto-diario/mostrar/1629553/espana-necesita-profesionales-ciberseguridad. [Accedido: 03-Mar-2020].

[4] Wyser Spain. (2018, September 28). Demanda de expertos en ciberseguridad. Wyser Spain. Retrieved from https://bit.ly/2JQSTr0

[5] Ramírez, V. (2019, January 4). Alemania: El "Bundestag"sufre el mayor hackeo de su historia. CyberSecurity News. Retrieved from https://bit.ly/3aZr1g9

[6] Basallo, A. (2018, July 18). Existen más puestos de trabajo en el sector de Ciberseguridad que profesionales formados. *UNIR Revista*. Retrieved from https://bit.ly/3eaRB8f

[7] Ministerio Educación y Formación Profesional (Ed.)(2020). Real Decreto Curso de Especialización: Ciberseguridad en entornos de tecnologías de la información. TodoFP. Retrieved from https://www.boe.es/diario_boe/txt.php?id=BOE-A-2020-4963

[8] Junta de Andalucía - Recursos humanos del sistema educativo en Andalucía. (2021). [Online]. Available: https://www.juntadeandalucia.es/organismos/educacionydeporte/servicios/estadistica-cartografia/actividad/detalle/175115/175501.html

[9] Ministerio de Educación y Formación Profesional (Ed)(2019). I Plan Estratégico de Formación Profesional del Sistema Educativo 2019-2020. Ministerio de Educación y Formación Profesional. [Online] Available: https://www.todofp.es/dam/jcr:163978c0-a214-471e-868d-82862b5a3aa3/plan-estrategico--enero-2020.pdf

[10] CSEC2017-JTF. (Ed.)(2017). Cybersecurity Curricula 2017. *A Report in the Computing Curricula Series Joint Task Force on Cybersecurity Education*. https://cybered.hosting.acm.org/wp-content/uploads/2018/02/newcover_csec2017.pdf

[11] Shull, F., Singer, J. and Sjoberg, D. (2008). *Guide to advanced empirical software engineering*. London: Springer London. https://doi.org/10.1007/978-1-84800-044-5

[12] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M., Regnell, B. and Wesslén, A. (2012). *Experimentation in software engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-29044-2

[13] Bailetti, T., & Craigen, D. (2020). Examining the relationship between cybersecurity and scaling value for new companies. *Technology Innovation Management Review*, 10(2), 62-70. https://doi:10.22215/timreview/1329

[14] Soblechero, M. V. L., Gaya, C. G., & Ramírez, J. J. H. (2014). A comparative study of classroom and online distance modes of official vocational education and training. *PLoS ONE*, 9(5). https://doi:10.1371/journal.pone.0096052

[15] González-Manzano, L., & de Fuentes, J. M. (2019). Design recommendations for online cybersecurity courses. Computers & Security, 80, 238–256. https://doi.org/10.1016/j.cose.2018.09.009

[16] Markopoulou, D., Papakonstantinou, V., & de Hert, P. (2019). The new EU cybersecurity framework: The NIS Directive, ENISA's role and the General Data Protection Regulation. Computer Law & Security Review, 35(6), 105336. https://doi.org/10.1016/j.clsr.2019.06.007

[17] Muller, J. (2015). The future of knowledge and skills in science and technology higher education. Higher Education, 70(3), 409–416. https://doi.org/10.1007/s10734-014-9842-x

[18] Gallego-Arrufat, M., Torres-Hernández, N., & Pessoa, T. (2019). Competence of future teachers in the digital security area. [Competencia de futuros docentes en el área de seguridad digital]. *Comunicar*, 61, 57-67. https://doi.org/10.3916/C61-2019-05

[19] Engen, B.K. (2019). Understanding social and cultural aspects of teachers' digital competencies. [Comprendiendo los aspectos culturales y sociales de las competencias digitales docentes]. *Comunicar*, 61, 9-19. https://doi.org/10.3916/C61-2019-01

[20] Fang, B., Ren, K., Jia, Y. (2018). The New Frontiers of Cybersecurity. *Engineering*, 4(1), 1-2. https://doi:10.1016/j.eng.2018.02.007

[21] Ávila, J.A. & Tello, J. (2004). Reflections on curricular integration of new communication technologies. [Reflexiones sobre la integración curricular de las tecnologías de la comunicación]. *Comunicar*, 22, 177-182. https://doi.org/10.3916/C22-2004-27

[22] Cabaj, K., Domingos, D., Kotulski, Z., & Respício, A. (2018). Cybersecurity education: Evolution of the discipline and analysis of master programs. Computers & Security, 75, 24–35. https://doi.org/10.1016/j.cose.2018.01.015

[23] Colás-Bravo, P., Conde-Jiménez, J., & Reyes-de-Cózar, S. (2019). The development of the digital teaching competence from a sociocultural approach. [El desarrollo de la competencia digital docente desde un enfoque sociocultural]. *Comunicar*, 61, 21-32. https://doi.org/10.3916/C61-2019-02

[24] Hodhod, R., Khan, S., & Wang, S. (2019). CyberMaster: An expert system to guide the development of cybersecurity curricula. *International Journal of Online and Biomedical Engineering*, 15(3), 70-81. https://doi:10.3991/ijoe.v15i03.9890

[25] Buckley, I. A., & Zalewski, J. (2019). Course development in the cybersecurity curriculum. Paper presented at *the Proceedings of the LACCEI International Multi-Conference for Engineering, Education and Technology*. https://doi:10.18687/LACCEI2019.1.1.238

[26] Duffany, J. L. (2019). Developing cybersecurity skills in intermediate programming courses. Paper presented at the *Proceedings of the LACCEI International Multi-Conference for Engineering, Education and Technology*. https://doi:10.18687/LACCEI2019.1.1.414

[27] Imbernón, F., Silva, P., & Guzmán, C. (2011). Teaching skills in virtual and blended learning environments. [Competencias en los procesos de enseñanza-aprendizaje virtual y semipresencial]. *Comunicar*, 36, 107-114. https://doi.org/10.3916/C36-2011-03-01

[28] Gordillo, A., López-Pernas, S., & Barra, E. (2019). Effectiveness of MOOCs for teachers in safe ICT use training. [Efectividad de los MOOC para docentes en el uso seguro de las TIC]. *Comunicar*, 61, 103-112. https://doi.org/10.3916/C61-2019-09

[29] Velandia-Mesa, C., Serrano-Pastor, F., & Martínez-Segura, M. (2017). Formative research in ubiquitous and virtual environments in Higher Education. [La investigación formativa en ambientes ubicuos y virtuales en educación superior]. *Comunicar*, 51, 09-18. https://doi.org/10.3916/C51-2017-01

[30] Rego-Agraso, L. (2018). Vocational training schools and its relationship with the community: Perspective of trainers and trainees. [Los centros de formación profesional y su vinculación con el entorno: La perspectiva de alumnado y profesorado] .*Revista Complutense De Educación*, 29(3), 683-697. https://doi:10.5209/RCED.53622

[31] Dameff, C. J., Selzer, J. A., Fisher, J., Killeen, J. P., & Tully, J. L. (2019). Clinical Cybersecurity Training Through Novel High-Fidelity Simulations. The Journal of Emergency Medicine, 56(2), 233–238. https://doi.org/10.1016/j.jemermed.2018.10.029

[32] Tynjälä, P., Välimaa, J., & Sarja, A. (2003). Pedagogical perspectives on the relationships between higher education and working life. Higher Education, 46(2), 147–166. https://doi.org/10.1023/A:1024761820500

PREMIOS RENIC

Definición de una Metodología para la Evaluación de Seguridad de Dispositivos del Internet de las Cosas

Sara Nieves Matheu García
Universidad de Murcia
Departamento de Ingeniería de la Información y las Comunicaciones
saranieves.matheu@um.es

***Resumen*—El desarrollo de un framework de certificación es una iniciativa ambiciosa que ha generado un gran interés en todo el mundo, tanto en industria y en investigación, como en organizaciones estandarizadoras y reguladoras. Mientras que en Estados Unidos esta iniciativa está liderada por el NIST, en Europa, tras la aprobación del Cybersecurity Act, ENISA ha adoptado el rol de liderar la creación de dicho framework. Diferentes retos alientan y, a la vez, frenan el desarrollo del framework de certificación, especialmente en el contexto del Internet de las Cosas (IoT). Esta tesis propone una metodología de evaluación de la seguridad enfocada a dispositivos IoT lidiando con muchos de dichos retos, como son la objetividad de la evaluación, los continuos cambios de seguridad y la creación de una etiqueta visual que muestre el resultado de la evaluación.**

***Index Terms*—Seguridad, Evaluación, Internet de las Cosas**

Tipo de contribución: *Premio doctoral RENIC*

I. Introducción

En los últimos años, la tecnología ha avanzado a pasos agigantados cambiando nuestra percepción del mundo y la manera en la que realizamos acciones cotidianas. Uno de los paradigmas que más impacto ha tenido en nuestro día a día es el Internet de las Cosas (IoT), el cual ha permitido conectar a Internet dispositivos cotidianos con el objetivo de recopilar y compartir información. Mientras que en 2019 el número de dispositivos IoT era de aproximadamente 26.66 billones, en los próximos años esta tendencia seguirá al alza, con una estimación de 74.44 billones de dispositivos en 2025 según Gartner[1]. Uno de los principales problemas de este tipo de dispositivos es la poca seguridad que implementan, debido a los bajos costes de producción y a la baja capacidad de cómputo. El hecho es que un atacante dispone de una enorme red de dispositivos interconectados, muchas veces escasamente desprotegidos. Solo hace falta revisar las noticias actuales para ver la cantidad de ataques solamente basados en redes de bots IoT, derivadas de Mirai [1], uno de los ataques con mayor impacto (2016). Juguetes, cámaras, televisores, grabadoras de vídeo, etc., millones de dispositivos quedan a merced del atacante. En este contexto y tratando de lidiar con los problemas de seguridad a los que se enfrenta no solo IoT, sino también otros contextos como 5G o la nube, Europa aprobó en 2019 el Cybersecurity Act [2]. Esta regulación sienta las bases para la construcción de un esquema Europeo de certificación de la ciberseguridad. Actualmente, el foco de su implementación está puesto en los tres dominios mencionados.

La tesis se enmarca en este contexto, analizando para ello los principales problemas a los que el Cybersecurity Act se va a tener que enfrentar. Este análisis no sólo tiene en cuenta las propiedades analizadas y los esquemas de certificación actuales, sino que también recoge las inquietudes de industria, organismos reguladores y organismos estandarizadores, ya que durante la tesis se ha podido tener contacto con varias de estas entidades, como por ejemplo la Organización Europea de Ciberseguridad (ECSO).

II. Desafíos de la certificación de la seguridad en dispositivos IoT

Entre los desafíos de la certificación, destaca la alta variedad de estándares de seguridad y esquemas [3], que hacen difícil la homogeneización de criterios de evaluación y la comparación entre ellos, así como el uso de métricas que pueden depender del criterio del evaluador. Sin embargo, a pesar de dichas debilidades, el framework debería hacer uso de los puntos fuertes de los estándares existentes y aprovecharse de su aceptación en la comunidad [4]. El proceso de certificación debería ser sencillo, requiriendo la mínima documentación formal posible para entenderlo y aplicarlo y no demasiado caro. Esto es especialmente importante en el entorno IoT, donde el coste de los dispositivos es tan bajo que una certificación que incremente demasiado su coste sería inviable, y donde un retraso excesivo en el lanzamiento del producto al mercado puede derivar en pérdidas económicas para el fabricante [5]. Aunque para evaluar la seguridad se suele tener en cuenta las vulnerabilidades ya conocidas del dispositivo que se está evaluando, en el contexto IoT no hay ninguna base de datos de vulnerabilidades específica. No obstante, cada vez más vulnerabilidades de estos dispositivos son añadidas a la base de vulnerabilidades de referencia, la base de datos nacional de vulnerabilidades (NVD) de Estados Unidos. El contexto donde opere el dispositivo, así como sus componentes y dependencias, y las vulnerabilidades en las diferentes capas de la pila de protocolos, deberían tenerse en cuenta para una correcta evaluación de la seguridad.

Sin embargo, hoy por hoy, el mayor problema es la dinamicidad, ya que aunque un dispositivo haya sido certificado como seguro, esto puede cambiar rápidamente, no sólo por una nueva vulnerabilidad, sino por una actualización o un parche. Este hecho, unido a que la cantidad de dispositivos

[1]https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/

IoT no deja de crecer, hace necesaria una solución escalable y automatizada que permita una recertificación rápida y a bajo coste [5]. Por último, como resultado del proceso de certificación, es necesaria la creación de una etiqueta que aporte información suficiente de una manera visual, sencilla y clara, para que pueda ser interpretada por una persona no experta [6].

III. Contenido de la tesis

Como una propuesta para lidiar con dichos retos, la tesis diseña una metodología de evaluación de seguridad, con el objetivo de que pueda servir como base para la certificación. En particular, dicha metodología se basa en el estándar ETSI EG 203 251 [7], en el cual se plantean dos visiones diferentes: una en la que la evaluación del riesgo es asistida por tests, y otra en la que los tests son dirigidos por la evaluación del riesgo. Dichas visiones fueron combinadas y se añadieron aspectos adicionales inherentes a la certificación, como es el concepto de etiquetado. El esquema de la metodología puede verse en la Figura 1.

Como resultado, la tesis presenta una metodología de evaluación de la seguridad en la que la evaluación del riesgo se hace de forma objetiva y empírica, combinado los resultados de los tests con el estándar Common Vulnerability Scoring System (CVSS) [8] de estimación del riesgo. Además la metodología de testing se basa en modelos, donde el sistema se modela a alto nivel y los tests son generados automáticamente a partir de ese modelo, lo que agiliza el proceso de certificación y de recertificación, ahorrando costos y tiempo. El contexto y la regulación también cobran importancia, y se toman como base para la evaluación de la seguridad. Finalmente, la metodología propone y diseña una etiqueta dinámica (a través de un QR) y multidimensional, reflejando cinco objetivos de seguridad: autenticación, autorización, confidencialidad, integridad y disponibilidad. Además, la etiqueta se diseña de una manera visual, de manera que más área significa más riesgo, lo que es muy intuitivo para consumidores no expertos.

Finalmente, la tesis aborda la pregunta de ¿qué hago con la información obtenida durante el proceso de certificación? Normalmente dicha información simplemente se utiliza para asignar un nivel de seguridad al dispositivo, pero la tesis propone enlazar ese proceso de certificación que suele desarrollarse en la fase de fabricación, con la fase de instalación y de operación del dispositivo. Sabemos qué cosas están minando la seguridad del dispositivo, así que podemos plantear ciertas mitigaciones o recomendaciones para que se utilicen cuando el dispositivo se instale en la red en la que va estar trabajando, de manera que la superficie de ataque se reduzca significativamente. Para ello se plantea el uso del estándar Manufacturer Usage Description (MUD) [9], de manera que se genera fichero con políticas de seguridad que ayudan a mitigar los fallos de seguridad encontrados. Además, la tesis aborda la problemática de obtener e instalar las políticas del fichero MUD durante la fase de instalación del dispositivo, protegiendo la red y el propio dispositivo desde el principio.

IV. Conclusiones

Por un lado, la tesis presenta una propuesta innovadora que sirve como primera piedra hacia ese framework común Europeo de certificación, que es el objetivo del Cybersecurity Act, y por otro lado, la metodología se diseña teniendo en cuenta las necesidades de las diferentes partes implicadas en el proceso. Mientras que se proponen mecanismos dinámicos que ahorren costes y tiempo al fabricante, el consumidor podrá elegir qué producto comprar en base a la seguridad ofrecida solo mirando la etiqueta, de manera similar a la etiqueta de eficiencia energética. La colaboración continua con iniciativas europeas relacionadas con el Cybersecurity Act y el desarrollo de un marco de certificación europeo, ha alineado la tesis con las necesidades reales actuales para un mercado digital Europeo seguro. De esta manera, la tesis pretende representar un esfuerzo inicial para aumentar la conciencia sobre la necesidad de definir enfoques alineados que orienten la creación de dicho esquema unificado de certificación de ciberseguridad para IoT en los próximos años.

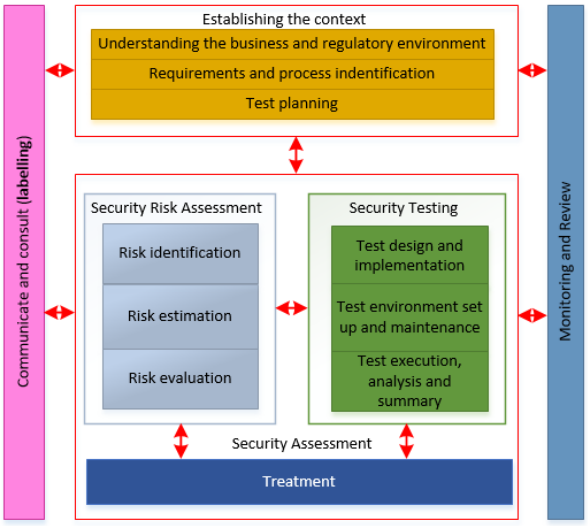


Figura 1. Metodología de evaluación de la seguridad propuesta

V. Enlace a la memoria

La tesis doctoral, defendida el 28 de Julio de 2020, puede ser descargada desde el siguiente enlace: https://digitum.um.es/digitum/handle/10201/95698.

VI. Agradecimientos

A mis directores de tesis, D. Antonio Skarmeta Gómez y D. José Luis Hernández-Ramos.

Referencias

[1] C. Kolias, G. Kambourakis, A. Stavrou, and J. Voas, "DDoS in the IoT: Mirai and other botnets," *Computer*, vol. 50, no. 7, pp. 80–84, 2017.
[2] European Union and European Commission, "Regulation (EU) 2019/881 of the European Parliament and of the Council on ENISA and on Information and Communications Technology Cybersecurity Certification and Repealing Regulation," 2019.
[3] ECSO, "State of the Art Syllabus V2," 2017.
[4] AIOTI, *Report on Workshop on Security and Privacy in the Hyper-Connected World*, 2016.
[5] ENISA, "Considerations on ICT Security Certification in EU - Survey Report," 2017.
[6] ECSO, "European Cyber Security Certification A Meta-Scheme Approach v1.0," 2017.
[7] ETSI, *Methods for Testing & Specification; Risk-based Security Assessment and Testing Methodologies*, 2015.
[8] FIRST, *Common Vulnerabilities Scoring System (CVSS)*, 2014. [Online]. Available: https://www.first.org/cvss
[9] E. Lear, D. Romascanu, and R. Droms, "Manufacturer Usage Description Specification (RFC 8520)," 2019.

Towards Privacy–Preserving Sensor–Based Continuous Authentication

Luis Hernández–Álvarez
Institute of Physical and Information Technologies (ITEFI)
Spanish National Research Council (CSIC)
C/ Serrano 144, Madrid, Spain
luis.hdez.alvarez@iec.csic.es

***Abstract*—Ensuring the privacy of the personal data stored in our technological devices is a fundamental aspect for protecting our personal and professional information. Authentication procedures are among the main methods used to achieve this protection and, typically, are implemented only when accessing to the device. Nevertheless, in many occasions it is necessary to carry out this authentication in a continuous manner to guarantee an allowed use of the device while protecting the authentication data. In this work, we first review the state of the art of Continuous Authentication (CA) and related privacy–preserving methods and, secondly, propose a CA scheme using sensor–based data and machine learning algorithms, that ensures the protection of the information via Format–Preserving Encryption with a unique and secret key per user. Our experimental results on a real–world dataset show the suitability of the proposed scheme, featuring 76.85% of accuracy while respecting users privacy.**

***Index Terms*—Continuous Authentication, Format Preserving Encryption, Machine Learning, Privacy–Preserving, Sensor–Based Data.**

Tipo de contribución: TFM Award RENIC

I. Introduction

Nowadays, smartphones, tablets and some resource constrained devices are commonly being used to store private information such as financial data, personal or professional documents and social communications. With the advent of wearable or implantable medical devices, even medical signals as the heart rate or the blood sugar level can be recorded. This way of storing information is effective and comfortable, but it also makes the data be potentially vulnerable to cyberattacks. They may occur because of software infection (e.g. Android malware [1]) or lack of user's diligence. Therefore, it is essential to implement a minimal access policy. One of the mechanisms for this policy is authentication, ensuring that the porting user is the legitimate one. Traditionally, it has been achieved by means of passwords, PINs or patterns that must be typed in by the user. However, these techniques face two main issues. On the one hand, complex passwords are rare to find as they are difficult to remember, thus leading to guessable authentication codes [2]. On the other hand, they provide with permanent access to the device, so all data would get compromised if the device is stolen at any time after authentication. To address this issue, Continuous Authentication (CA) approaches have been proposed [2]. Thanks to CA, the identity of the porting user is periodically verified, typically by relying on biometric factors such as heart rate, fingerprints, and body– and health–related signals. However, the privacy of the information used to conduct CA is threatened, specially in the cases in which the authentication is carried out by an external server. The leakage of this information could lead to the theft of the user's identity by the attacker. As a consequence, the importance of privacy–preserving techniques that ensure both the security of the user's biometric information and its utility in CA applications, has increased considerably. Some approaches have already been proposed considering two biometric or behavioral traces (namely, iris and touch dynamics), but to the best of author's knowledge, it has not been addressed when smartphones come into play.

II. Objectives

In this work we aim to achieve a twofold objective. The first one is to provide a complete overview of the current research status of biometric CA, privacy–preserving techniques, and their combination [5]. This includes the review of recent publications focused on these fields and the biometric information used, algorithms employed, and results obtained. Secondly, the development of a privacy-preserving CA mechanism suitable for smartphones. It will be based on processing biometric data collected from smartphone accelerometers and gyroscopes, after a cryptographic transformation (particularly, Format-Preserving Encryption (FPE) [3], [4])through a Support Vector Machine (SVM) in a one–vs-all configuration [6]. The most novel and remarkable aspect is the use of cryptographic tools for CA with the objective of preserving the privacy of the users to be authenticated when the authentication decision is outsourced to an untrusted third party.

III. Work Content

The bibliographic review revealed the necessity of adding privacy–preserving techniques to CA protocols. Those are based on behavioral or biological biometric information and make use of artificial intelligence tools (e.g. SVM, Random Forest (RF), Logistic Regressor (LR)). However, only few works focus on protecting the authentication information, and it is typically done by relaying in invertible transformation and homomorphic encryption.

The proposed scheme uses FPE, an encryption procedure characterized by maintaining the format of the encrypted data, to securely outsource the information acquired from accelerometers and gyroscopes to a server (considered as non-trusted) and authenticate an specific user with a SVM. The whole procedure can be divided in two parts, both executed by a smartphone app (named CAwPP): an *Enrollment Protocol*, in which the user's biometric information will be sent to the

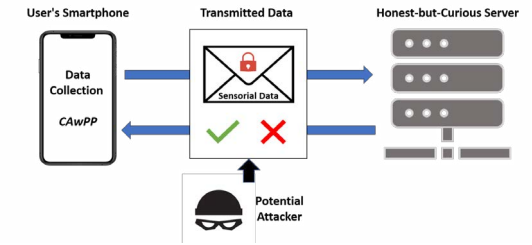


Fig. 1. Model Parties Representation.

server in order to train a SVM; and a *Verification Protocol*, where the biometric data will be recursively collected and sent to conduct the CA. The parties present in the model are showed in Figure 1.

We consider that the attacker could act once the information has arrived to the server or during its transmission and, hence, suggest the implementation of a double encryption procedure. As soon as the data is acquired, it will be encrypted via FPE and input to the SVM. To avoid the leak of information during data transmission from the smartphone to the server, we also propose the use of a hybrid cryptosystem. As a result, the server will receive the information of each of the users for the process of CA. However, this information will not contain the original values, but twice encrypted data, in order to guarantee the anonymity of these data and the subject.

The cryptographic tools needed for the protocols proposed in this study are: 1) A symmetric cryptosystem with secret key, k, functions of encryption $\mathcal{C}_k$ and decryption $\mathcal{D}_k$ (e.g., AES with keys of 256 bits). 2) An asymmetric cryptosystem with public and private keys, (pk, sk), functions of encryption $\mathfrak{C}_k$ and decryption $\mathfrak{D}_k$ (e.g., RSA with public key of 2048 bits). 3) A hash function, $\mathfrak{h}$ (e.g. SHA-256). 4) Digital certificates with public a private keys, (pk, sk), for a user with pseudonym (e.g. X059, v3). Apart from this, a FPE cryptosystem, $\mathcal{F}$, is also needed. In this work, the keys of this cryptosystem will be composed by 20 characters randomly generated, considering the 26 letters of the alphabet and the 10 digits, which could be repeated. This is equivalent to use keys of 103 bits.

The experimental process was divided in two parts: firstly, different configurations of a SVM, RF and LR were explored with the original (non–encrypted) data to establish a baseline performance and crossvalidate parameters (Table I). This experience allowed to reduce the number of features from 90 (originally taken from the Sherlock Database [1]) to 15 and define a train-test distribution of 15:85%. The results also suggested that a SVM might be the best model for this application.

TABLE I
AUTHENTICATION MODELS' PERFORMANCE

Model	SVM		RF		LR	
	Avg. Acc (%)	St. D	Avg. Acc (%)	St. D	Avg. Acc (%)	St. D
1^{st}	82.24	11.75	80.99	13.00	73.92	8.86
2^{nd}	82.25	11.74	80.99	12.98	73.94	8.84
3^{rd}	82.30	11.72	80.88	12.97	73.88	8.86
Overall	82.26	0.04	80.95	0.06	73.91	0.03
Time/user (min)	8.09		6.88		0.10	

Then, the data was encrypted via FPE and the authentication protocol was tested with a SVM. In this case, two options were considered: creating one FPE key shared by all the users (Single–Key) or generating one specific key for each one (Multi–Key). The results (Table II) indicate that the Multi–Key SVM performs better and, therefore, is the preferred option not only from the security perspective, but also from the authentication point of view.

TABLE II
ENCRYPTED CA RESULTS

Model	Single-Key SVM		Multi-Key SVM	
	Avg. Acc (%)	St. D	Avg. Acc (%)	St. D
1^{st}	71.72	8.97	76.78	10.89
2^{nd}	71.61	8.96	76.65	11.05
3^{rd}	71.68	9.06	77.12	10.66
Overall	71.67	0.05	76.85	0.25
Time/user (min)	26.82		26.19	

IV. CONCLUSIONS

With this work, we offer two useful instruments to promote the study of CA methods and privacy–preserving techniques. The first is a review of the state of the art of these two investigation areas, that includes publications focused on each of them individually and combined. Moreover, a novel CA scheme, focused on preserving the privacy of the outsourced information related to the user's data, is proposed. Its novelty is determined by three aspects: 1) the use of data acquired by the accelerometer and gyroscope of a user's smartphone, 2) the protection of the data in the verification process by an external entity via a FPE cryptosystem, and 3) the employment of a SVM to conduct the user authentication with such encrypted data. The outcomes obtained show a good average accuracy of 76.85%, taking into account that the authentication performance with the encrypted data is only reduced a 5.41%, in comparison with the baseline performance with the original data. In addition, we verified that establishing a personal secret key to each user provide better results than using the same secret key for all of them.

ACKNOWLEDGEMENTS

The author would like to thank to J.M. de Fuentes, L. González Manzano, L. Hernández Encinas for their valuable help, and to CSIC Project 202050E304 (CASDiM).

REFERENCES

[1] Mirsky, A. Shabtai, L. Rokach, B. Shapira, and Y. Elovici: "Sher-Lockvs Moriarty: A smartphone dataset for cybersecurity research", in *Proc.2016 ACM Workshop on Artificial Intelligence and Security (AISec'16)*, pp. 1-–12, 2016.

[2] M. Obaidat, I. Traore, and I. Woungang: "Biometric-Based Physical and Cybersecurity Systems", in *Springer, Cham*, 2019.

[3] M. Bellare, P. Rogaway, and T. Spies: "The FFX mode of opera-tion for format-preserving encryption", in *NIST, Tech. Rep.*, 2010.

[4] V. Gayoso Martínez, L. Hernández Encinas, A. Martín Muñoz, J. M. deFuentes, and L. González Manzano: "Cifrado de datos con preservacióndel formato", in *Primeras Jornadas Nacionales de Investigación en Ciberseguridad (JNIC)*, pp. 110—115, 2016.

[5] L. Hernández–Álvarez, J.M. de Fuentes, Lorena González–Manzano and L. Hernández Encinas: "Privacy–Preserving Sensor–Based Continuous Authentication and User Profiling: A Review", in *Sensors*, vol. 21(1), 92, 23 pp., 2021.

[6] L. Hernández-Álvarez, J.M. de Fuentes, L. González-Manzano and L. Hernández Encinas: "SmartCAMPP - Smartphone–based Continuous Authentication leveraging Motion sensors with Privacy Preservation", in *Pattern Recognition Letters* 147, pp. 189-–196, 2021.

PATROCINADORES