XV

Reunión Española sobre Criptología y Seguridad de la Información

Granada, 3-5 octubre 2018

UNIVERSIDAD DE GRANADA

RENIC Red de Excelencia Nacional de Investigación en Ciberseguridad

CCN Centro Criptológico Nacional

indra

Network Engineering & Security Group
https://nesg.ugr.es

Actas XV Reunión Española sobre Criptología y Seguridad de la Información (RECSI)

Granada, 3, 4 y 5 de octubre de 2018

Universidad de Granada

©Universidad de Granada, *Network Engineering & Security Group* (NESG – https://nesg.ugr.es)

Está permitida la descarga y la reproducción total o parcial de esta obra y su difusión siempre y cuando sea para uso personal o académico. Los derechos de autor de cada contribución individual corresponden a sus autores.

Eds.: Pedro García Teodoro, Raúl Barragán Gil y N.Marta Fuentes García.

ISBN: 978-84-09-02463-6

Prefacio

Las *Tecnologías de la Información y de las Comunicaciones* (TIC) son una realidad de uso cotidiano en todos los ámbitos de nuestra sociedad. Sin embargo, no podemos perder de vista los elevados riesgos de seguridad asociados a la adopción generalizada de estas tecnologías. Así, espionaje, robo de información, fraude, ciber-acoso, son nuevas formas de (ciber)crimen de aparición cada vez más amplia e impacto.

Es por ello necesario que la comunidad científico-tecnológica en el ámbito de la seguridad TIC sigamos avanzando en el desarrollo y difusión de propuestas y esquemas que abunden en la provisión de un mayor nivel de seguridad, y con ello de confianza de la sociedad en las tecnologías involucradas.

En este marco, y continuando con la labor de la **Reunión Española sobre Criptología y Seguridad de la Información (RECSI)** de ediciones previas, en esta XV edición pretendemos continuar mejorando con la ayuda de todos de manera que las jornadas resulten provechosas y supongan una nueva aportación en el camino hacia la promoción y consolidación de la cripto y la ciberseguridad en España.

Con este ánimo, desde Granada, sede de esta nueva edición de RECSI, y ciudad cosmopolita y de gran riqueza cultural donde las haya, damos la bienvenida a la comunidad científica en el ámbito de la seguridad TIC, e invitamos a participar activamente a todos a fin de conseguir los mejores resultados posibles.

Equipo RECSI 2018.

Comité de programa

- Almenares Mendoza, Florina - *Universidad Carlos III de Madrid*
- Areitio Bertolín, Javier - *Deusto Unibertsitatea*
- Borrell Viader, Joan - *Universitat Autònoma de Barcelona*
- Bras Amorós, Maria - *Universitat Rovira i Virgili*
- Caballero Gil, Pino - *Universidad de La Laguna*
- Carriegos Vieira, Miguel - *Universidad de León*
- Castellà Roca, Jordi - *Universitat Rovira i Virgili*
- Climent Coloma, Joan Josep - *Universitat d'Alacant*
- Díaz García, Antonio Fco. - *Universidad de Granada*
- Domingo Ferrer, Josep - *Universitat Rovira i Virgili*
- Durán Díaz, Raúl - *Universidad de Alcalá*
- Egea González, M. Soledad - *Minsait by Indra*
- Estévez Tapiador, Juan Manuel - *Universidad Carlos III de Madrid*
- Fernández Arrieta, Miquel - *Mondragon Unibertsitatea*
- Ferrer Gomila, Josep Lluís - *Universitat de les Illes Balears*
- Fúster Sabater, Amparo - *CSIC*
- Gayoso Martínez, Víctor - *CSIC*
- González Vasco, María Isabel - *Universidad Rey Juan Carlos*
- Gutiérrez Gutiérrez, Jaime - *Universidad de Cantabria*
- Hernández Encinas, Luis - *CSIC*
- Hernández Goya, Candelaria - *Universidad de La Laguna*
- Herranz Sotoca, Javier - *Universitat Politècnica de Catalunya*
- Herrera Joancomartí, Jordi - *Universitat Autònoma de Barcelona*
- Jacob, Eduardo - *Euskal Herriko Unibertsitatea*
- Lobillo Borrero, Javier - *Universidad de Granada*
- López Hernández-Ardieta, Jorge - *Minsait by INDRA*
- López Muñoz, Javier - *Universidad de Málaga*
- Martín del Rey, Ángel - *Universidad de Salamanca*
- Martínez López, Consuelo - *Universidad de Oviedo*
- Megías Jiménez, David - *Universitat Oberta de Catalunya*
- Miret Biosca, José María - *Universitat de Lleida*
- Muñoz Muñoz, Alfonso - *Criptored*
- Peinado Domínguez, Alberto - *Universidad de Málaga*
- Ramió Aguirre, Jorge - *Universidad Politécnica de Madrid*
- Sáez Moreno, Germán - *Universitat Politècnica de Catalunya*
- Salazar Riaño, José Luis - *Universidad de Zaragoza*
- Sánchez Ávila, Carmen - *Universidad Politécnica de Madrid*
- Sánchez Crespo, Luis Enrique - *Universidad de Castilla-La Mancha*
- Sebé Feixas, Francesc - *Universitat de Lleida*
- Soriano Ibáñez, Miguel - *Universitat Politècnica de Catalunya*
- Villagrá González, Víctor - *Universidad Politécnica de Madrid*
- Zamora Gómez, Antonio - *Universitat d'Alacant*
- Zurutuza Ortega, Urko - *Mondragon Unibertsitatea*

Comité Organizador

Responsable general:

García Teodoro, Pedro (*Universidad de Granada*)

***Responsable de infraestructuras*:**

Robles Carrilllo, Margarita (*Universidad de Granada*)

Responsable de publicidad:

Maciá Fernández, Gabriel (*Universidad de Granada*)

Responsable de publicación - Easy Chair:

Marta Fuentes García, Noemí (*Universidad de Granada*)

Responsable de publicación - Actas:

Barragán Gil, Raúl (*Universidad de Granada*)

***Webmaster*:**

Gómez Hernández, José Antonio (*Universidad de Granada*)

Lista de patrocinadores

Universidad de Granada

Red de Excelencia Nacional de Investigación en Ciberseguridad

Centro Criptológico Nacional

Indra Sistemas

Sulayr

Índice de contribuciones

Sesión 1: Criptología 12

Cálculo de la complejidad lineal de secuencias criptográficas mediante PN-secuencias ... 12
Amparo Fúster-Sabater and Sara D. Cardell

Steiner Triple Systems and Zero Knowledge Protocols 18
Feliú Sagols, Guillermo Morales-Luna and Edgar González Fernández

Generador de ruido Zener para incrementar la entropía de un TRNG 22
Alfonso Blanco Blanco, Víctor Gayoso Martínez, Luis Hernández Encinas, Agustín Martín Muñoz, Fausto Montoya Vitini and Amalia B. Orúe López

Linear cellular automata, bivariate polynomials and power of 2^r purely periodic binary sequences 26
Domingo Gomez-Perez, Ana Isabel Gomez and Jaime Gutierrez

Sobre secuencias pseudoaleatorias a partir de autómatas celulares y funciones bent 30
Joan-Josep Climent and Verónica Requena

Sesión 2: Forense y Esteganografía 36

Fingerprinting in weighted noisy adder channels 36
Marcel Fernandez, Elena Egorova and Grigory Kabatyanskiy

Diagnóstico de CSM en estegoanálisis 39
Daniel Lerch-Hostalot and David Megías

Técnica de Detección de Manipulaciones de Imágenes Digitales Basada en la Matriz del Filtro de Color 45
Edgar González Fernández, Esteban Alejandro Armas Vega, Ana Lucila Sandoval Orozco and Luis Javier García Villalba

Técnica de Agrupamiento de Vídeos de Dispositivos Móviles 51
Raquel Ramos López, Ana Lucila Sandoval Orozco and Luis Javier García Villalba

Evaluación de un sistema forense para BitTorrent 56
Alberto Caravaca and Gabriel Maciá-Fernandez

Sesión 3: Criptografía y Ciberseguridad 62

Improving a smart metering system using elliptic curves and removing the trusted dealer 62
Ricard Garra, Santi Martinez, Josep M. Miret and Francesc Sebe

Identificación de polinomios primitivos sobre cuerpos extendidos mediante análisis de secuencias entrelazadas 66
Guillermo Cotrina, Andrés Ortiz García and Alberto Peinado

Criptografía Basada en Identidad aplicada a una MANET formada por robots 71
Jonay Suárez-Armas, Alexandra Rivero-García, Pino Caballero-Gil and Cándido Caballero-Gil

Prácticas de desarrollo de extensiones en el mercado de Mozilla tras Firefox Quantum ... 75
Félix Brezo and Yaiza Rubio

Seguridad aplicada a una herramienta para el Trastorno por Déficit de Atención e Hiperactividad 81
Nayra Rodríguez-Pérez, Alexandra Rivero-García, Pino Caballero-Gil and Josué Toledo-Castro

Arquitectura de gestión dinámica de riesgos basada en ontologías y reglas de comportamiento 86
Irene Romero, Raúl Riesco, Francisco Barea and Víctor A. Villagrá

Herramienta para la identificación de requisitos de seguridad en un Modelo de Desarrollo Seguro 92
José Carlos Sancho Núñez, Andrés Caro Lindo, Lucio Fondón Terrón and José Andrés Félix de Sande

Ciberdefensa empresarial: Un marco conceptual y práctico en un entorno digitalmente inestable 96
Jeimy Cano

Sesión 4: Privacidad y Anonimato 102

Anonimización de datos no estructurados a través del reconocimiento de entidades nominadas 102
Fadi Hassan, Josep Domingo-Ferrer and Jordi Soria-Comas

Privacy in Microblogging Online Social Networks: Issues and Metrics 107
Samia Oukemeni, Helena Rifà-Pous and Joan Manuel Marquès Puig

Análisis del protocolo para la anonimización de trayectorias desde el lado del cliente ... 113
Cristina Romero and David Megías

Técnica Anti-Forense para la Anonimización de Vídeos de Dispositivos Móviles 118
Carlos Quinto Huaman, Ana Lucila Sandoval Orozco and Luis Javier García Villalba

Data Utility Evaluation Framework for Graph Anonymization 123
Jordi Casas-Roma

Anonimización de trayectorias por medio de intercambios 129
Julián Salas, David Megias and Vicenc Torra

Sesión 5: IoT y SmartGrid 135

Validation of a SIR Epidemic Model for the Propagation of Jamming Attacks in Wireless Sensor Networks 135
Miguel López, Alberto Peinado and Andrés Ortiz García

Ecosistemas IoT que preserven la privacidad 141
Juan Hernández-Serrano, Jose L. Muñoz, Arne Bröring, Lars Mikkelsen, Wolfgang Schwarzott, Oscar Esparza, Olga Leon, Miguel Soriano and Jan Zibuschka

Pruebas Basadas en el Entorno para la Detección de Ataques de Relay en Accesos a Zonas Restringidas 147
Carles Anglés-Tafalla, Jordi Castellà-Roca, Alexandre Viejo, Magdalena Payeras-Capellà and Macià Mut-Puigserver

A Survey on Incident Reporting and Management Systems 153
Amna Qureshi, David Megías and Helena Rifà-Pous

Sistema de comunicaciones seguras para el personal de servicios de emergencias 159
Alexandra Rivero-García, Candelaria Hernández Goya, Iván Santos-González and Pino Caballero-Gil

Sesión 6: Seguridad y Análisis de Datos 163

Criptografía adversaria usando deep learning. Limitaciones y oportunidades 163
Alfonso Muñoz and José Ignacio Escribano

Mejorando la Detección de Spam Social Utilizando la Subjetividad 169
Enaitz Ezpeleta, Mikel Iturbe, Iñaki Garitano, Iñaki Velez de Mendizabal and Urko Zurutuza

Constrained approximate bit-parallel search with application in cryptanalysis 174
Slobodan Petrovic

Hacia una Arquitectura de Referencia Segura para Big Data 180
Julio Moreno, Manuel A. Serrano, Eduardo Fernandez-Medina and Eduardo B. Fernandez

Sistema Basado en Lógica Difusa para la Detección del Robo de Identidad en Redes Sociales 186
José Á. Concepción-Sánchez, Pino Caballero-Gil and Iván Santos-González

Sesión 7: Detección y Ciberdefensa 190

Marisma-CPS: Una Aproximación al Análisis y Gestión de Riesgos para Sistemas Ciberfísicos 190
David G. Rosado, Julio Moreno, Antonio Santos-Olmo, Luis E. Sánchez, Manuel A. Serrano and Eduardo Fernandez-Medina-Marisma

Aplicación de técnicas de compresión de información a la identificación de anomalías en fuentes de datos heterogéneas: análisis y limitaciones 196
Gonzalo de La Torre, Luis Lago and David Arroyo

Mapeo de Dependencias para el Impacto de Ciberataques en Misiones: Una Visión Global 201
Raúl Barragán Gil, Jorge López Hernández-Ardieta and Pedro García Teodoro

Definición de procedimientos para fabricar honeypots IoT basados en criterios de búsqueda 206
Antonio Acien, Ana Nieto, Gerardo Fernandez and Javier Lopez

MEDEA: Monitorizando el espacio nulo para la detección de anomalías en sistemas industriales 212
Ekhi Zugasti, Mikel Iturbe, Iñaki Garitano and Urko Zurutuza

Sesión 8: Criptografía Aplicada 217

Un esquema de intercambio de clave a tres bandas postcuántico 217
Josep M. Miret, Daniel Sadornil, Juan G. Tena and Javier Valera

Sistema de Comunicaciones Seguras para Posicionamiento en Interiores 221
Iván Santos-González, Alexandra Rivero-García and Pino Caballero-Gil

Mejora de la seguridad en la autentificación basada en contraseñas mediante un cifrador en bloque 226
Rafael Alvarez, Alicia Andrade Bazurto and Antonio Zamora Gómez

Analysis of the SegWit adoption in Bitcoin 230
Cristina Pérez-Solà, Sergi Delgado Segura, Jordi Herrera-Joancomarti and Guillermo Navarro-Arribas

Reconstrucción de la Clave PILAR utilizada por el Gobierno Civil de Málaga en 1940 234
Alberto Peinado

Sesión 9: Seguridad y e-Administración 240

Adecuación al Esquema Nacional de Seguridad: Una Aproximación Metodológica 240
Pedro García Teodoro and José Antonio Gómez Hernández

Esquemas criptográficos post-cuánticos aplicados a la e-democracia: LWE-Helios Voting 246
David Yeregui Marcos Del Blanco, Luis Panizo and JoséÁngel Hermida

Notificaciones Certificadas sobre Blockchain 251
Macià Mut Puigserver, Magdalena Payeras-Capellà and Miquel Àngel Cabot-Nadal

Uso y retos de *blockchain* en plataformas de votación electrónica 257
Victor Garcia-Font and Helena Rifà-Pous

Fidelización mediante Cryptotokens 263
Magdalena Payeras, Macià Mut-Puigserver and Llorenç Huguet-Rotger

Cálculo de la complejidad lineal de secuencias criptográficas mediante PN-secuencias

Amparo Fúster Sabater
Instituto de Tecnologías Físicas y de la Información
Consejo Superior de Investigaciones Científicas (CSIC)
Serrano, 144, 28006, Madrid, España
amparo@iec.csic.es

Sara D. Cardell
Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas (Unicamp)
R. Sérgio Buarque de Holanda, 651, Campinas, Brasil
sdcardell@ime.unicamp.br

***Abstract*—En este trabajo presentamos un método para calcular la complejidad lineal de las secuencias producidas por un generador de secuencia cifrante denominado generador auto-*shrinking* generalizado. Nuestro procedimiento está basado en la comparación de versiones desplazadas de una misma PN-secuencia. El método es sencillo, directo y eficaz, ya que el simple análisis de los bits de estas PN-secuencias nos permite determinar la complejidad lineal de las secuencias generadas. El concepto de relación de recurrencia lineal y las filas del triángulo de Sierpinski juegan un papel primordial en nuestra propuesta.**

***Index Terms*—generador auto-*shrinking* generalizado, complejidad lineal, recurrencia lineal, triángulo de Sierpinski**

I. Introducción

La protección de la información confidencial utiliza un tipo de transformación (función de cifrado) para convertir un mensaje original o *texto claro* en un *mensaje cifrado*. En criptografía simétrica existe un elemento de información llamado *clave* que tiene que permanecer secreto y ha de ser compartido únicamente por ambos comunicantes. Este tipo de criptografía se dividide en dos grandes clases: cifrado en flujo y cifrado en bloque, dependiendo de si la transformación de cifrado se aplica a cada bit individualmente o a un bloque de bits.

A día de hoy, los cifradores en flujo son los más rápidos y simples de entre todos los esquemas de cifrado y, por ello, su uso se ha extendido a todo tipo de aplicaciones tecnológicas, por ejemplo el sistema de cifrado E0 en Bluethooth [1], el algorithm RC4 en el procesador Microsoft Word y las hojas de cálculo de Microsoft Excel [2] o el generador SNOW 3G [3] en comunicaciones inalámbricas de datos a alta velocidad con tecnología 4G/LTE (Long-Term Evolution).

El problema básico en el diseño de cifradores en flujo es generar, a partir de una clave corta, una secuencia pseudo-aleatoria larga denominada *secuencia cifrante*. Para cifrar, el emisor realiza una operación XOR entre los bits del texto claro y los bits de la secuencia cifrante. El resultado es el texto cifrado que se envía al receptor. Para descifrar, el receptor genera la misma secuencia cifrante y realiza la misma operación XOR entre el texto cifrado recibido y dicha secuencia, recuperando así el mensaje original. Nótese que tanto el cifrado como el descifrado usan el mismo procedimiento y la misma operación lógica.

Muchas generadores de secuencias cifrantes están basados en *Linear Feedback Shift Registers* (LFSRs) [4] de longitud máxima cuyas secuencias de salida, las llamadas PN-secuencias, se combinan de forma no lineal para producir secuencias pseudo-aleatorias con aplicación en criptografía. Combinadores, filtros no lineales, generadores controlados por reloj o generadores basados en decimación son algunos de los prototipos más conocidos para generar secuencias cifrantes mediante LFSRs, véanse las referencias [5], [6], [7].

Dentro de la familia de generadores basados en decimación irregular, podemos enumerar: 1) el *generador shrinking* [8] que incluye dos LFSRs, 2) el *generador auto-shrinking* [9] que cuenta con un único LFSR y 3) el elemento más representativo de esta familia, el *generador auto-shrinking generalizado* [10] que genera una familia completa de secuencias entre las que se encuentra la producida por el generador auto-*shrinking*. Los generadores basados en decimación producen secuencias equilibradas (mismo número de ceros que de unos), con periodos largos, buena correlación, simplicidad en la implementación, etc. La idea subyacente de este tipo de generadores es la decimación irregular de una PN-secuencia de acuerdo con los bits de otra. El resultado es una secuencia que puede usarse como secuencia cifrante en diferentes procedimientos criptográficos. Este trabajo se centra en el generador auto-*shrinking* generalizado y sus secuencias de salida, conocidas como *secuencias auto-shrunken generalizadas* o simplemente *secuencias generalizadas*.

Desde un punto de vista criptográfico, la complejidad lineal de una secuencia, LC, es una buena métrica de su seguridad [11]. De hecho, LC mide la cantidad de bits necesarios para poder reconstruir la totalidad de la secuencia. En términos criptográficos, la complejidad lineal debe ser lo más grande posible; el valor recomendado es aproximadamente la mitad del periodo, denotado por T, de la secuencia, $LC \simeq T/2$. Tradicionalmente, la complejidad se puede calcular por medio del algoritmo de Berlekamp-Massey [12] después de haber procesado por lo menos $2 \cdot LC$ de sus bits. Para secuencias en un rango criptográfico ($T = 10^{38}$), la generación y aplicación de este algoritmo puede ser una tarea extremadamente compleja. A pesar de su importancia, la complejidad lineal de las secuencias generalizadas es un tópico que no ha sido analizado en la literatura. En este trabajo, introducimos un método simple para calcular la complejidad lineal de esta familia de secuencias. Cabe destacar que este método no necesita la generación de tales secuencias, basta simplemente considerar diferentes versiones desplazadas de una única PN-secuencia.

El trabajo está organizado de la siguiente manera: en la Sección 2 se presentan los fundamentos y conceptos básicos utilizados a lo largo del trabajo. En la Sección 3, se desarrolla el resultado principal, un método para calcular la complejidad

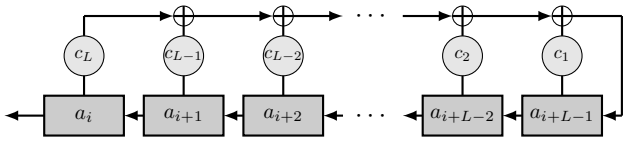


Fig. 1. LFSR de longitud L

lineal de secuencias auto-*shrunken* generalizadas; además de la formulación se presentan ejemplos y una discusión de dicho método. Finalmente, las conclusiones en la Sección 4 completan el trabajo.

II. Fundamentos y conceptos básicos

Para comenzar, introducimos el concepto de decimación de una secuencia binaria ya que se utilizará repetidamente a lo largo de este trabajo. Sea $\{a_i\}$ $(i = 0, 1, 2, \dots)$ una secuencia definida sobre el cuerpo binario de dos elementos, $a_i \in \mathbb{F}_2$. La decimación de la secuencia $\{a_i\}$ con distancia d es una nueva secuencia $\{b_i\}$ $(i = 0, 1, 2, \dots)$ obtenida tomando cada d-ésimo término de $\{a_i\}$, esto es $\{b_i\} = \{a_{d \cdot i}\}$ [13].

Sea L un entero positivo, y sean $c_0, c_1, \dots, c_{L-1}$ elementos del cuerpo binario $\mathbb{F}_2$. Se denomina relación de recurrencia lineal (r.r.l.) de orden L a una expresión del tipo

$$a_{i+L} = c_1 a_{i+L-1} + c_2 a_{i+L-2} + \ldots + c_{L-1} a_{i+1} + c_L a_i,$$

con $i \geq 0$. Una secuencia $\{a_i\}$ que satisface la relación anterior es una secuencia lineal recurrente de orden L en $\mathbb{F}_2$.

El polinomio de grado L

$$P(x) = x^L + c_1 x^{L-1} + c_2 x^{L-2} + \ldots + c_{L-1} x + c_L \in \mathbb{F}_2[x],$$

es el polinomio característico de la relación de recurrencia lineal anterior.

La generación de secuencias recurrentes puede implementarse mediante Linear Feedback Shift Registers (LFSRs) [4]. Estos registros manejan información en forma de bits y están basados en desplazamientos y realimentación lineal. Un LFSR consta de L celdas de memoria interconectadas con contenido binario, el polinomio característico $P(x)$ y su relación de recurrencia lineal, y el estado inicial no nulo (contenido de las celdas en el instante inicial), véase la Figura 1. Si $P(x)$ es un polinomio primitivo, entonces el registro se dice que es un LFSR de longitud máxima y su secuencia de salida $\{a_i\}$ es una PN-secuencia de periodo $T = 2^L - 1$ con 2^{L-1} unos y $(2^{L-1} - 1)$ ceros [4]. Si α es una raíz de $P(x)$, entonces α es un elemento primitivo de $\mathbb{F}_{2^L}$, el cuerpo extendido de $\mathbb{F}_2$, cuyos elementos son el 0 y sucesivas potencias de α [14]. A través del polinomio característico, se puede definir una correspondencia uno a uno:

$$a_i \to \alpha^i \qquad (i = 0, 1, 2, \dots, 2^L - 2) \tag{1}$$

entre el elemento *i*-ésimo, a_i, de la PN-secuencia y la potencia *i*-ésima de α, denotada por α^i. La complejidad lineal es la longitud del mínimo LFSR que genera tal secuencia o, equivalentemente, el menor orden de la relación de recurrencia lineal que genera dicha secuencia.

El elemento más representativo dentro de la clase de generadores basados en decimación irregular es el generador auto-*shrinking* generalizado [10] descrito como sigue:

Tabla I
Secuencias generalizadas del ejemplo 1

p	$\{s(p)_j\}$
0	11111111
1	11100100
2	11011000
3	10101010
4	01010101
5	10110001
6	01101001
7	11000011
8	10010110
9	00100111
10	01001110
11	10001101
12	00011011
13	00111100
14	01110010

Definición 1: Sea $\{a_i\}$ $(i = 0, 1, 2, \dots)$ una secuencia generada por un LFSR de longitud máxima con polinomio característico de grado L. Sea p un entero y $\{v_i\}$ $(i = 0, 1, 2, \dots)$ una versión de $\{a_i\}$ desplazada cíclicamente p posiciones a la izquierda con $(p = 0, 1, 2, \dots, 2^L - 2)$. La regla de decimación es muy simple:

1) Si $a_i = 1$, entonces v_i es el bit de salida, esto es $s_j = v_i$.
2) Si $a_i = 0$, entonces v_i se descarta y no hay bit de salida.

De este modo, para cada p, se genera una secuencia de salida $\{s_0\, s_1\, s_2 \dots\}$ denotada por $\{s(p)_j\}$ o simplemente $\{s_j\}$ que es equilibrada. Tal secuencia se denomina secuencia auto-*shrunken* generalizada asociada al desplazamiento p. ■

Recordemos que $\{a_i\}$ permanece fija mientras que $\{v_i\}$ es la versión desplazada p posiciones con relación a $\{a_i\}$. Cuando p recorre el intervalo $p \in [0, 1, 2, \dots, 2^L - 2]$, entonces se obtiene la clase de $2^L - 1$ secuencias auto-*shrunken* generalizadas (o simplemente secuencias generalizadas). Veamos un ejemplo sencillo.

Ejemplo 1: Para un LFSR de longitud $L = 4$, con polinomio característico $P(x) = x^4 + x^3 + 1$ y estado inicial $(1, 1, 1, 1)$, su PN-secuencia correspondiente viene dada por $\{a_i\} = \{111101011001000\}$. Aplicando la regla de decimación anterior, obtenemos $2^L - 1 = 15$ secuencias generalizadas $\{s_j\}$ basadas en $\{a_i\}$ y representadas en la Tabla I. Las $2^L - 1 = 15$ elecciones de p dan origen a 15 desplazamientos distintos de $\{v_i\}$. Para cada secuencia $\{v_i\}$, una nueva secuencia generalizada es generada. ■

El periodo de las secuencias generalizadas es un divisor de 2^{L-1}. Esta familia de secuencias siempre incluye la secuencia $\{111111, \ldots\}$ para $p = 0$ y las secuencias $\{101010, \ldots\}$ y $\{010101, \ldots\}$ para $p = n, n+1$, donde n es el entero positivo tal que $\alpha^n \in \mathbb{F}_{2^L}$ satisface $\alpha^{n+1} = \alpha^n + 1$.

Finalmente, sea $\{u_i\}$ $(i = 0, 1, \dots, 2^{L-1} - 1)$ una secuencia de periodo $T = 2^{L-1}$ cuyos términos u_i son elementos de $\mathbb{F}_{2^L}$. Teniendo en cuenta la correspondencia uno a uno definida en (1), los términos de $\{u_i\}$ son las potencias α asociadas con los unos de $\{a_i\}$. Veamos la secuencia $\{u_i\}$ del Ejemplo 1 (Tabla II).

La secuencia viene dada por $\{u_i\} = \{1, \alpha, \alpha^2, \alpha^3, \alpha^5, \alpha^7, \alpha^8, \alpha^{11}\}$. En [15] se demuestra que la secuencia $\{u_i\}$ tiene complejidad lineal acotada

Tabla II
SECUENCIA $\{u_i\}$ PARA EL EJEMPLO 1

$\mathbb{F}_{2^L}$	1	α	α^2	α^3	α^4	α^5	α^6	α^7	α^8	α^9	α^{10}	α^{11}	α^{12}	α^{13}	α^{14}
$\{a_i\}$	**1**	**1**	**1**	**1**	0	**1**	0	**1**	**1**	0	0	**1**	0	0	0
$\{u_i\}$	1	α	α^2	α^3		α^5		α^7	α^8			α^{11}			

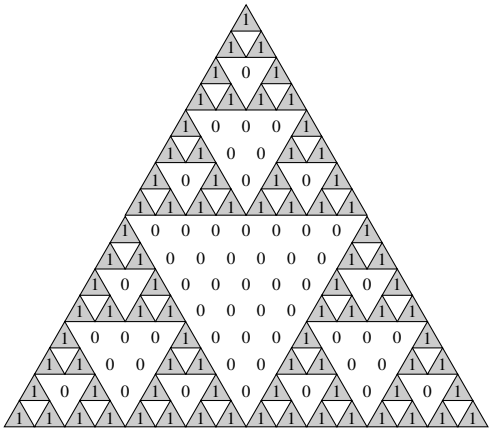


Fig. 2. Triángulo de Sierpinski binario

superiormente por:

$$LC(\{u_i\}) \leq 2^{L-1} - (L-2). \tag{2}$$

III. COMPLEJIDAD LINEAL DE LAS SECUENCIAS GENERALIZADAS

Como las secuencias generalizadas tienen periodo potencia de 2, entonces el polinomio característico de cada secuencia generalizada es de la forma $(x+1)^M$ donde $M = LC$. Así pues, $(x+1)^{M+1}, (x+1)^{M+2}, (x+1)^{M+3}, \ldots$ son polinomios característicos de grado superior que definen relaciones de recurrencia lineal que la correspondiente secuencia generalizada tiene que verificar. Por el contrario, $(x+1)^{M-1}, (x+1)^{M-2}, (x+1)^{M-3}, \ldots$ no son polinomios característicos, y por tanto, la correspondiente secuencia generalizada no verifica las relaciones de recurrencia lineal definidas por estos polinomios. Esta es la idea clave para calcular la LC de la clase de secuencias auto-*shrunken* generalizadas.

Los coeficientes de un polinomio de la forma $(x+1)^M$ son los números combinatorios $\binom{M}{i}$ $(i = 0, 1, \ldots, M)$ de la *M*-ésima fila del triángulo de Pascal [16], [17]. Cuando estos términos se reducen módulo 2, entonces obtenemos el triángulo de Sierpinski, véase la Figura 2. Asimismo los Autómatas Celulares lineales (CA) con leyes 102 y 60, referenciados en [18], [19], [20], definen igualmente los coeficientes de este tipo de polinomio. En la Figura 3 se muestran las CA-imágenes de ambas leyes, donde los cuadrados negros representan los unos y los blancos los ceros.

A continuación, estudiamos las relaciones de recurrencia lineal de las secuencias generalizadas para sucesivos valores decrecientes de M.

PASO 1. Para $M = 2^{L-1} - (L-2)$:

Según la ecuación (2), la secuencia $\{u_i\}$ satisface la r.r.l.

$$\sum_{i=0}^{M} c_i \; u_{i+p} = 0 \qquad (p = 0, 1, 2, \ldots, 2^L - 2), \tag{3}$$

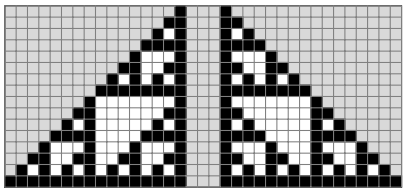


Fig. 3. CA-imágenes: ley 102 (izquierda) y ley 60 (derecha)

donde los c_i son los coeficientes binarios de la M-ésima fila del triángulo de Sierpinski, Figura 2 (ó de las CA-imágenes de la Figura 3).

Si denotamos $u_i = \alpha^{\tau(i)}$ donde $\tau(i)$ representa la posición del i-ésimo 1 en la PN-secuencia $\{a_i\}$, entonces podemos reescribir la correspondencia anterior (1) tal y como sigue:

$$u_i = \alpha^{\tau(i)} \quad \rightarrow \quad a_{\tau(i)} \quad (i = 0, 1, 2, \ldots, 2^L - 2).$$

Por lo tanto,

$$\sum_{i=0}^{M} c_i \; u_{i+p} = \sum_{i=0}^{M} c_i \; \alpha^{\tau(i)+p} = 0 \quad \rightarrow \quad \sum_{i=0}^{M} c_i \; a_{\tau(i)+p} = 0, \tag{4}$$

donde $\{a_{\tau(i)+p}\}$ $(i = 0, 1, 2, \ldots)$ denota la secuencia generalizada $\{s(p)_j\}$ asociada con el desplazamiento p. Así, según la ecuación (4), todas las secuencias generalizadas $\{s(p)_j\}$ satisfacen $\sum_{i=0}^{M} c_i \; a_{\tau(i)+p} = 0$ y sus complejidades lineales están acotadas superiormente por

$$LC(\{s(p)_j\}) \leq 2^{L-1} - (L-2) \qquad (p = 0, 1, 2, \ldots, 2^L - 2).$$

Por tanto hemos determinado una cota superior de LC para todas las secuencias generalizadas.

PASO 2. Para $M = 2^{L-1} - (L-2) - 1$**:**

Comprobamos si la secuencia $\{u_i\}$ satisface la r.r.l. con el nuevo valor de M y la nueva fila de coeficientes c_i en el triángulo de Sierpinski's (ó en las CA-imágenes).

$$\sum_{i=0}^{M} c_i \; u_{i+p} = \alpha^m \neq 0 \qquad (p = 0, 1, 2, \ldots, 2^L - 2),$$

Por lo tanto,

$$\sum_{i=0}^{M} c_i \; \alpha^{\tau(i)+p} = \alpha^m \quad \rightarrow \quad \sum_{i=0}^{M} c_i \; a_{\tau(i)+p} = a_{m+p}.$$

1) Si $a_{m+p} = 0$ para algún p, entonces la r.r.l. se cumple para las secuencias correspondientes $\{s(p)_j\}$ y su $LC(\{s(p)_j\}) \leq 2^{L-1} - (L-2) - 1$.
2) Si $a_{m+p} = 1$ para algún p, entonces la r.r.l. no se cumple para las secuencias correspondientes $\{s(p)_j\}$ y su $LC(\{s(p)_j\}) = 2^{L-1} - (L-2)$, es decir el valor anterior para el cual sí se cumplía la r.r.l.

Como a_{m+p} $(p = 0, 1, 2, \ldots, 2^L - 2)$ es la PN-secuencia $\{a_i\}$ empezando en el término a_m, entonces tendremos $(2^{L-1} - 1)$ términos $a_{m+p} = 0$, esto es $(2^{L-1} - 1)$ secuencias generalizadas para las cuales la r.r.l. se cumple y 2^{L-1} términos $a_{m+p} = 1$, esto es 2^{L-1} secuencias generalizadas para las cuales la r.r.l. no se cumple.

PASO 3. Para $M = 2^{L-1} - (L-2) - 2$**:**

Comprobamos si la secuencia $\{u_i\}$ satisface la r.r.l. con el nuevo valor de M y la nueva fila de coeficientes c_i del triángulo de Sierpinski (ó en las CA-imágenes).

De hecho, para los sucesivos valores de p tenemos dos valores alternativos de esta r.r.l.

$$\sum_{i=0}^{M} c_i\ u_{i+p} = \alpha^{m_1} \neq 0, \qquad \sum_{i=0}^{M} c_i\ u_{i+p} = \alpha^{m_2} \neq 0.$$

Luego,

$$\sum_{i=0}^{M} c_i\ a_{\tau(i)+p} = a_{m_1+p}, \qquad \sum_{i=0}^{M} c_i\ a_{\tau(i)+p} = a_{m_2+p}.$$

Por tanto, obtenemos dos versiones desplazadas de la PN-secuencia $\{a_i\}$, una empezando en el término a_{m_1} y la otra en a_{m_2}.

1) Si $a_{m_1+p} = a_{m_2+p} = 0$ para algún p, entonces la r.r.l se cumple para las secuencias correspondientes $\{s(p)_j\}$ y su $LC(\{s(p)_j\}) \leq 2^{L-1} - (L-2) - 2$.
2) Si $a_{m_1+p} = a_{m_2+p} = 1$ para algún p, entonces la r.r.l. no se cumple para las secuencias correspondientes $\{s(p)_j\}$ y su $LC(\{s(p)_j\}) = 2^{L-1} - (L-2) - 1$.

Así, tendremos $(2^{L-2} - 1)$ términos $a_{m_1+p} = a_{m_2+p} = 0$, esto es, habrá $(2^{L-2} - 1)$ secuencias generalizadas con $LC(\{s(p)_j\}) \leq 2^{L-1} - (L-2) - 2$. De la misma manera, tendremos 2^{L-2} términos $a_{m_1+p} = a_{m_2+p} = 1$, esto es, habrá 2^{L-2} secuencias generalizadas con $LC(\{s(p)_j\}) = 2^{L-1} - (L-2) - 1$.

Para valores sucesivos y decrecientes de M, esto es, $M = (2^{L-1} - (L-2) - 3), (2^{L-1} - (L-2) - 4), (2^{L-1} - (L-2) - 5), \ldots$, obtenemos 4, 8, 16, $\ldots$ versiones desplazadas de la PN-secuencia $\{a_i\}$. El número de secuencias generalizadas que satisfacen (no satisfacen) la r.r.l. en un paso determinado es la mitad del número de secuencias obtenidas en el paso anterior. A continuación, un ejemplo numérico clarifica todo este proceso.

A. Un ejemplo numérico

Consideremos un LFSR de longitud máxima con $L = 5$, polinomio característico $P(x) = x^5 + x^2 + 1$ y estado inicial $(1, 1, 1, 1, 1)$, su PN-secuencia correspondiente es

$$\{a_i\} = \{1111100011011101010000100101100\}$$

con periodo $T = 31$. La secuencia $\{u_i\}$ viene dada por

$$\{u_i\} = \{1, \alpha, \alpha^2, \alpha^3, \alpha^4, \alpha^8, \alpha^9, \alpha^{11}, \alpha^{12}, \alpha^{13}, \alpha^{15}, \alpha^{17}, \alpha^{22}, \alpha^{25}, \alpha^{27}, \alpha^{28}\}$$

y, además, tenemos que $\alpha^{14} = \alpha^{13} + 1$. Computamos la r.r.l. para diferentes valores de M.

PASO 1. Para $M = 2^{L-1} - (L-2) = 13$:

La secuencia $\{u_i\}$ satisface la r.r.l. dada por la ecuación (3), donde los coeficientes c_i son $(1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1)$, véase la decimotercera fila de la Figura 2 (ó de la Figura 3). Como $\sum_{i=0}^{13} c_i\ a_{\tau(i)+p} = 0$ $(p = 0, 1, 2, \ldots, 30)$, sabemos que la complejidad lineal de todas las secuencias generalizadas es $LC(\{s(p)_j\}) \leq 13$.

PASO 2. Para $M = 2^{L-1} - (L-2) - 1 = 12$**:**

Comprobamos si la secuencia $\{u_i\}$ satisface la r.r.l. dada en la ecuación (3) para $M = 12$ y para los nuevos coeficientes c_i $(1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1)$, véase la duodécima fila de las Figura 2 (ó de la Figura 3). De hecho, $\sum_{i=0}^{12} c_i\ \alpha_{\tau(i)+p} = \alpha^6 \neq 0$, entonces $\{a_{6+p}\}$ $(p = 0, 1, 2, \ldots, 30)$ es la PN-secuencia generalizada $\{a_i\}$ empezando en el término a_6. Así pues, según la Tabla III, las 15 secuencias generalizadas correspondientes a $p = 0, 1, 4, 8, 10, 12, 13, 14, 15, 17, 18, 20, 23, 24, 30$ (los ceros de $\{a_{6+p}\}$ en negrita) satisfarán $LC(\{s(p)_j\}) \leq 12$; mientras que las 16 secuencias generalizas correspondientes al resto de valores de p (los unos de $\{a_{6+p}\}$) tendrán $LC(\{s(p)_j\}) = 13$. Por tanto, hemos computado LC para algunas secuencias generalizadas simplemente analizando dígitos binarios de la PN-secuencia $\{a_{6+p}\}$.

PASO 3. Para $M = 2^{L-1} - (L-2) - 2 = 11$**:**

Comprobamos si la secuencia $\{u_i\}$ satisface la r.r.l. dada en la ecuación (3) para $M = 11$ y para los nuevos coeficientes c_i $(1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1)$, véase la undécima fila de la Figura 2 (ó de la Figura 3). Ahora, $\sum_{i=0}^{11} c_i\ \alpha_{\tau(i)+p} = \alpha^{10}, \alpha^{16} \neq 0$, alternativamente para sucesivos valores de p. Así pues, según la Tabla IV, las 7 secuencias generalizadas correspondientes a $p = 0, 4, 8, 10, 13, 14, 20$ satisfarán $LC(\{s(p)_j\}) \leq 11$. Recordemos que los valores previos de p corresponden a los ceros coincidentes en las 3 PN-secuencias desplazadas $\{a_{6+p}\} = \{a_{10+p}\} = \{a_{16+p}\} = 0$ (véase las columnas en negrita de la Tabla IV).

Por otro lado, las 8 secuencias generalizadas correspondientes a $p = 1, 12, 15, 17, 18, 23, 24, 30$ tendrán $LC(\{s(p)_j\}) = 12$. Recordemos que los valores previos de p corresponden a los unos coincidentes en las PN-secuencias desplazadas $\{a_{10+p}\} = \{a_{16+p}\} = 1$ (en los rectángulos grises) pero con $\{a_{6+p}\} = 0$ (véase las columnas de la Tabla IV). Por tanto, la comparación de dígitos binarios en tres versiones desplazadas de una sola PN-secuencia nos permite computar la LC de diversas secuencias generalizadas.

PASO 4. Para $M = 2^{L-1} - (L-2) - 2 = 10$**:**

Comprobamos si la secuencia $\{u_i\}$ satisface la r.r.l dada en la ecuación (3) para $M = 10$ y para los nuevos coeficientes c_i $(1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1)$, véase la décima fila de la Figura 2 (ó de la Figura 3). Ahora, $\sum_{i=0}^{10} c_i\ \alpha_{\tau(i)+p} = \alpha^7, \alpha^5, \alpha^{24}, \alpha^{23} \neq 0$, alternativamente para los sucesivos valores de p. Así, según la Tabla V, las 3 secuencias generalizadas correspondientes a $p = 0, 13, 14$ satisfarán $LC(\{s(p)_j\}) \leq 10$. Recordemos que los valores previos de p corresponden a los ceros coincidentes en las siete PN-secuencias desplazadas $\{a_{6+p}\} = \{a_{10+p}\} = \{a_{16+p}\} = \{a_{7+p}\} = \{a_{5+p}\} = \{a_{24+p}\} = \{a_{23+p}\} = 0$ (véase las columnas en negrita de la Tabla V).

Por otro lado, las 4 secuencias generalizadas correspondientes a $p = 4, 8, 10, 20$ tendrán $LC(\{s(p)_j\}) = 11$. Recordemos que los valores previos de p corresponden a los unos coincidentes en las 4 PN-secuencias desplazadas $\{a_{7+p}\} = \{a_{5+p}\} = \{a_{24+p}\} = \{a_{23+p}\} = 1$ (los rectángulos en gris) pero con $\{a_{6+p}\} = \{a_{10+p}\} = \{a_{16+p}\} = 0$ (véase las columnas en la Tabla V). Ahora, la comparación de los dígitos binarios en siete versiones desplazadas de una misma PN-secuencia nos permite computar LC para más secuencias generalizadas.

PASO 5. Para $M = 9, 8, 7, 6, \ldots, 2$**:**

Seguimos el mismo procedimiento, pero no obtenemos valores nuevos en la expresión $\sum_{i=0}^{M} c_i\ \alpha_{\tau(i)+p}$.

PASO 6. Para $M = 1$**:**

Comprobamos si la secuencia $\{u_i\}$ satisface la r.r.l dada en la ecuación (3) para $M = 1$ y para

Tabla III
COMPLEJIDAD LINEAL DE SECUENCIAS GSS

p	**0**				**4**				**8**				**12**				**16**				**20**				**24**				**28**			**30**
$\{a_{6+p}\}$	**0**	**0**	1	1	**0**	1	1	1	**0**	1	**0**	1	**0**	**0**	**0**	**0**	1	**0**	**0**	1	**0**	1	1	**0**	**0**	1	1	1	1	1	**0**

Tabla IV
COMPLEJIDAD LINEAL DE SECUENCIAS GSS

p =	**0**				**4**				**8**				**12**				**16**				**20**				**24**				**28**			**30**
$\{a_{6+p}\}$	**0**	0	1	1	**0**	1	1	1	**0**	1	**0**	1	0	**0**	**0**	0	1	0	0	1	**0**	1	1	0	0	1	1	1	1	1	0
$\{a_{10+p}\}$	**0**	1	1	1	**0**	1	0	1	**0**	0	**0**	0	1	**0**	**0**	1	0	1	1	0	**0**	1	1	1	1	1	0	0	0	1	1
$\{a_{16+p}\}$	**0**	1	0	0	**0**	0	1	0	**0**	1	**0**	1	1	**0**	**0**	1	1	1	1	1	**0**	0	0	1	1	0	1	1	1	0	1

nuevos coeficientes c_i $(1, 1)$, véase la fila primera en la Figura 2 (ó la Figura 3). Ahora, $\sum_{i=0}^{1} c_i \ \alpha_{\tau(i)+p} = \alpha^{18}, \alpha^{19}, \alpha^{20}, \alpha^{21}, \alpha^{14}, \alpha^{26}, \alpha^{29}, \alpha^{30} \neq 0$, alternativamente para valores sucesivos de p. Así, según la Tabla VI, la secuencia generalizada correspondiente a $p = 0$ satisfará $LC(\{s(0)_j\}) = 1$. Ésta es la secuencia de unos. Recordemos que el valor $p = 0$ corresponde a lo ceros coincidentes en las 15 PN-secuencias desplazadas $\{a_{6+p}\} = \{a_{10+p}\} = \{a_{16+p}\} = \{a_{7+p}\} = \{a_{5+p}\} = \{a_{24+p}\} = \{a_{23+p}\} = \{a_{18+p}\} = \{a_{19+p}\} = \{a_{20+p}\} = \{a_{21+p}\} = \{a_{14+p}\} = \{a_{26+p}\} = \{a_{29+p}\} = \{a_{30+p}\} = 0$ (véase la columna en negrita en la Tabla VI).

Por otro lado, las 2 secuencias generalizadas correspondientes a $p = 13, 14$ satisfarán $LC(\{s(p)_j\}) = 2$. Éstas corresponden a las secuencias generalizadas $\{1010\ldots\}$ y $\{0101\ldots\}$. Recordemos que los valores previos de p corresponden a los unos coincidentes en las ocho PN-secuencias desplazadas $\{a_{18+p}\} = \{a_{19+p}\} = \{a_{20+p}\} = \{a_{21+p}\} = \{a_{14+p}\} = \{a_{26+p}\} = \{a_{29+p}\} = \{a_{30+p}\} = 1$ (rectángulos en gris) pero con $\{a_{6+p}\} = \{a_{10+p}\} = \{a_{16+p}\} = \{a_{7+p}\} = \{a_{5+p}\} = \{a_{24+p}\} = \{a_{23+p}\} = 0$ (véase las columnas en la Tabla VI). De esta forma, hemos computado LC para la familia completa de secuencias generalizadas de este ejemplo.

En un caso general, habrá 2^{L-1} secuencias generalizadas con $LC = 2^{L-1} - (L-2)$, 2^{L-2} secuencias con $LC_1 < 2^{L-1} - (L-2)$, 2^{L-3} secuencias con $LC_2 < LC_1$ y así sucesivamente, hasta que obtenemos las tres secuencias generalizadas $\{101010, \ldots\}$ y $\{010101, \ldots\}$ con $LC = 2$ y la secuencia de unos $\{111111, \ldots\}$ con $LC = 1$. Los valores intermedios LC_i están en el intervalo $2^{L-2} < LC_i \leq 2^{L-1} - (L-2)$ aunque no son necesariamente consecutivos.

B. Discusión del método

Nótese que el método para computar la LC de las secuencias auto-shunken generalizadas implica operaciones muy simples. De hecho, el método está basado en la comparación de una única PN-secuencia con versiones desplazadas de ella misma. Solamente una secuencia en cada grupo de PN-secuencias desplazadas es necesaria para determinar las posiciones de los unos y los ceros. Así, para un $M < 2^{L-1} - (L-2)$ dado, sólo tenemos que comprobar una relación de recurrencia lineal $\sum_{i=0}^{M} c_i \ \alpha_{\tau(i)+p}$. Esto quiere decir que, en total, tenemos que analizar $L-2$ grupos. Para L en rango criptográfico $(L \simeq 128)$ la eficiencia del método computacional es bastante evidente.

Desde el punto de vista del diseño del generador de secuencia cifrante, el procedimiento aquí presentado permite diseñar un generador auto-*shrinking* generalizado con una complejidad lineal máxima garantizada. En efecto, con $M = 2^{L-1} - (L-2) - 1$ y la computación de $\sum_{i=0}^{M} c_i \ \alpha_{\tau(i)+p} = \alpha^m$, los unos de la PN-secuencia $\{a_{m+p}\}$ determinan las secuencias auto-*shrunken* generalizadas $\{s(p)_j\}$ con complejidad lineal máxima de valor $LC = 2^{L-1} - (L-2)$. En resumen, garantizamos un diseño simple de secuencias criptográficas al precio de un número mínimo de operaciones computacionales.

Por otro lado, sabemos que el generador auto-*shrinking* es un elemento incluido en la clase del generador auto-*shrinking* generalizado. Es más, la secuencia auto-*shrunken* es la secuencia auto-*shrunken* generalizada correspondiente al desplazamiento $p = 2^{L-1}$. De este modo, la aplicación del método anterior permite determinar fácilmente la complejidad lineal de esta secuencia en un rango criptográfico. De hecho, esta complejidad será $LC = M_0$, donde M_0 es el primer valor del parámetro M para el cual la secuencia desplazada correspondiente $\{a_{s+p}\}$ satisface $a_{s+2^{L-1}} = 1$.

IV. CONCLUSIONES

La clase de secuencias auto-*shrunken* generalizadas exhibe buenas propiedades criptográficas: secuencias equilibradas, periodos largos, excelente distribución de rachas, buena correlación, generación rápida de bits, etc. Sin embargo, una medida fundamental de su seguridad, tal y como es la complejidad lineal, no había sido considerada anteriormente. En este trabajo, presentamos un método para computar la complejidad lineal de secuencias auto-*shrunken* generalizadas. El procedimiento es simple y está basado en la comparación de diferentes versiones desplazadas de una misma PN-secuencia. De hecho, los conceptos de relación de recurrencia lineal, triángulo de Sierpinski y automata celular lineal facilitan el cómputo del punto inicial de tales PN-secuencias desplazadas. La comparación de sus dígitos binarios determina el valor exacto de la complejidad lineal para esta clase de secuencias generalizadas.

El método aquí desarrollado garantiza una fácil elección de secuencias generalizadas con complejidad lineal máxima.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el Ministerio de Economía, Industria y Competitividad (MINECO), la

Tabla V
COMPLEJIDAD LINEAL DE SECUENCIAS GSS

p =	**0**				**4**				**8**				**12**				**16**				**20**				**24**				**28**		**30**
$\{a_{6+p}\}$	**0**	0	1	1	0	1	1	1	0	1	0	1	0	**0**	**0**	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1	0
$\{a_{10+p}\}$	**0**	1	1	1	0	1	0	1	0	0	0	0	1	**0**	**0**	1	0	1	1	0	0	1	1	1	1	1	0	0	0	1	1
$\{a_{16+p}\}$	**0**	1	0	0	0	0	1	0	0	1	0	1	1	**0**	**0**	1	1	1	1	1	0	0	0	1	1	0	1	1	1	0	1
$\{a_{7+p}\}$	**0**	1	1	0	1	1	1	0	1	0	1	0	0	**0**	**0**	1	0	0	1	0	1	1	0	0	1	1	1	1	1	0	0
$\{a_{5+p}\}$	**0**	0	0	1	1	0	1	1	1	0	1	0	1	**0**	**0**	0	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1
$\{a_{24+p}\}$	**0**	1	0	1	1	0	0	1	1	1	1	1	0	**0**	**0**	1	1	0	1	1	1	0	1	0	1	0	0	0	0	1	0
$\{a_{23+p}\}$	**0**	0	1	0	1	1	0	0	1	1	1	1	1	**0**	**0**	0	1	1	0	1	1	1	0	1	0	1	0	1	0	0	1

Tabla VI
COMPLEJIDAD LINEAL DE SECUENCIAS GSS

p =	**0**				**4**				**8**				**12**				**16**				**20**				**24**				**28**		**30**
$\{a_{6+p}\}$	**0**	0	1	1	0	1	1	1	0	1	0	1	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1	0
$\{a_{10+p}\}$	**0**	1	1	1	0	1	0	1	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1	0	0	0	1	1
$\{a_{16+p}\}$	**0**	1	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	1	0	1
$\{a_{7+p}\}$	**0**	1	1	0	1	1	1	0	1	0	1	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1	0	0
$\{a_{5+p}\}$	**0**	0	0	1	1	0	1	1	1	0	1	0	1	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1
$\{a_{24+p}\}$	**0**	1	0	1	1	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	1	0	1	0	1	0	0	0	0	1	0
$\{a_{23+p}\}$	**0**	0	1	0	1	1	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	1	0	1	0	1	0	0	0	0	1
$\{a_{18+p}\}$	**0**	0	0	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	1	0	1	0	1
$\{a_{19+p}\}$	**0**	0	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	1	0	1	0	1	0
$\{a_{20+p}\}$	**0**	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	1	0	1	0	1	0	0
$\{a_{21+p}\}$	**0**	1	0	0	1	0	1	1	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	1	0	1	0	1	0	0	0
$\{a_{14+p}\}$	**0**	1	0	1	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	1
$\{a_{26+p}\}$	**0**	1	1	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	1	0	1	0	0	0	0	0	1	0	0	1	1
$\{a_{29+p}\}$	**0**	0	1	1	1	1	1	0	0	0	1	1	0	1	1	1	0	1	0	1	0	0	0	0	1	0	0	1	0	1	1
$\{a_{30+p}\}$	**0**	1	1	1	1	1	0	0	0	1	1	0	1	1	1	0	1	0	1	0	0	0	0	1	0	0	1	0	1	1	0

Agencia Estatal de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER, UE), a través del proyecto COPCIS, referencia TIN2017-84844-C2-1-R, y por la Comunidad de Madrid (Spain) a través del proyecto CIBERDINE, referencia S2013/ICE-3095-CIBERDINE-CM, también cofinanciado con fondos FEDER de la Unión Europea. El segundo autor ha sido financiado por CAPES y por FAPESP (número de proceso 2015/07246-0).

REFERENCES

[1] S. Fluhrer and S. Lucks, "Analysis of the E_0 Encryption System," in *Selected Areas in Cryptography SAC 2001*, ser. Lecture Notes in Computer Science, S. Vaudenay and A. Youssef, Eds. Berlin Heidelberg: Springer-Verlag, 2001, vol. 2259, pp. 38–48.

[2] G. Paul and S. Maitra, *RC4 Stream Cipher and its Variants*. Boca Raton: CRC Press, Taylor and Francis Group, 2012.

[3] C. Jenkins, M. Schulte, and J. Glossner, "Instructions and hardware designs for accelerating SNOW 3G on a software-defined radio platform," *Analog Integrated Circuits and Signal Processing*, vol. 69, no. 2–3, pp. 207–218, 2011.

[4] S. W. Golomb, *Shift Register-Sequences*. Laguna Hill, California: Aegean Park Press, 1982.

[5] A. J. Menezes and Y.-H. Wu, "The discrete logarithm problem in $GL(n, q)$," *Ars Combinatoria*, vol. 47, pp. 23–32, 1997.

[6] A. Fúster-Sabater, "Computing classes of cryptographic sequence generators," in *Proceedings of International Conference on Computational Science (ICCS2013), Procedia Computer Science 18*, 2013, pp. 2440–2443.

[7] A. Fúster-Sabater, "Generation of cryptographic sequences by means of difference equations," *Applied Mathematics & Information Sciences*, vol. 8, no. 2, pp. 475–484, 2014.

[8] D. Coppersmith, H. Krawczyk, and Y. Mansour, "The shrinking generator," in *Advances in Cryptology – CRYPTO '93*, ser. Lecture Notes in Computer Science, D. Stinson, Ed. Springer-Verlag, 1994, vol. 773, pp. 22–39.

[9] W. Meier and O. Staffelbach, "The self-shrinking generator," in *Advances in Cryptology – EUROCRYPT 1994*, ser. Lecture Notes in Computer Science, A. De Santis, Ed. Berlin, Heidelberg: Springer-Verlag, 1995, vol. 950, pp. 205–214.

[10] Y. Hu and G. Xiao, "Generalized self-shrinking generator," *IEEE Transactions on Information Theory*, vol. 50, no. 4, pp. 714–719, 2004.

[11] C. Paar and J. Pelzl, *Understanding Cryptography*. Berlin: Springer, 2010.

[12] J. L. Massey, "Shift-register synthesis and BCH decoding," *IEEE Transactions on Information Theory*, vol. 15, no. 1, pp. 122–127, 1969.

[13] P. F. Duvall and J. C. Mortick, "Decimation of periodic sequences," *SIAM Journal on Applied Mathematics*, vol. 21, no. 3, pp. 367–372, 1971.

[14] R. Lidl and H. Niederreiter, *Introduction to Finite Fields and Their Applications*. New York, NY: Cambridge University Press, 1986.

[15] S. R. Blackburn, "The linear complexity of the self-shrinking generator," *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 2073–2077, 1999.

[16] S. D. Cardell and J.-J. Climent, "An approach to the performance of SPC product codes on the erasure channel," *Advances in Mathematics of Communications*, vol. 10, no. 1, pp. 11–28, 2016.

[17] S. D. Cardell and A. Fúster-Sabater, "Linear models for high-complexity sequences," in *Computational Science and Its Applications.*, ser. Lecture Notes in Computer Science, G. O. et al., Ed., vol. 10404. Springer, 2017, pp. 314–324.

[18] S. Wolfram, "Cellular automata as simple self-organizing system," *Caltrech preprint CALT 68–938*, 1982.

[19] S. D. Cardell and J.-J. Climent, "Recovering erasures by using MDS codes over extension alphabets," *SeMA Journal*, vol. 73, pp. 85–95, 2016.

[20] S. D. Cardell and A. Fúster-Sabater, "Modelling the shrinking generator in terms of linear CA," *Advances in Mathematics of Communications*, vol. 10, no. 4, pp. 797–809, 2016.

Steiner Triple Systems and Zero Knowledge Protocols

Feliú Sagols
Mathematics Department
Cinvestav-IPN
Mexico City, Mexico
Email: fsagols@math.cinvestav.mx

Guillermo Morales-Luna
and Edgar González Fernández
Computer Science Department
Cinvestav-IPN
Mexico City, Mexico
Email: gmorales@cs.cinvestav.mx

***Abstract*—Cycle-switching is a technique to transform in linear time any Steiner Triple System (STS) of a given order into another STS of the same order. Cycle-switching may produce both isomorphic and non isomorphic STS. For orders lower than 19 each isomorphism class can be reached from any STS by successive cycle-switching transformations, and it is conjectured that it holds for any possible order. We propose a zero-knowledge identification protocol based on cycle-switchings and the difficulty to identifying isomorphic Steiner Triple Systems.**

Keywords Null Knowledge, Steiner Triple Systems, Identification Protocols

I. Introduction

Identification protocols are widely used for access control, credit card operations validation, and many others. In an *identification protocol* [1], [2] a *prover* claims to know the *private key* corresponding to his/her *public key* and by showing some evidence, the prover shows his/her identity to a *verifier*. Typically, a private key is the solution of a a computationally difficult problem instance, which is the corresponding public key. Both parts know the problem instance, and through a challenge-response interaction the prover should convince the verifier that he/she holds the private key. Thus, depending on the right answers the verifier accepts or rejects the prover's identity. The protocol is of *zero-knowledge* if no further information about the private key can be obtained, neither by the verifier nor by any intruder listening the conversation.

A *zero-knowledge protocol* (ZKP) should convince the verifier about the identity of the prover, the verifier should not learn anything about the prover's private key, no intruder, not even the verifier, may impersonate the genuine prover. A ZKP is a "proof of knowledge" without providing the knowledge.

In a rather general way, a ZKP is based on the *Isomorphism Problem*. Given a category in which it is computationally hard to recover the isomorphism between two already known isomorphic structures, the prover's public key is a pair of isomorphic structures, say (S_0, S_1), with isomorphism $\phi_{01} : S_0 \to S_1$ which is the prover's private key. The ZKP consists in a series of rounds:

- *Commitment.* The prover builds a third structure S_2 isomorphic to S_1, thus the prover knows the isomorphism $\phi_{21} : S_2 \to S_1$, as well as the isomorphism $\phi_{20} = \phi_{01}^{-1} \circ \phi_{21} : S_2 \to S_0$. The prover sends S_2 to the verifier as a commitment.
- *Challenge.* The verifier selects a bit $b \in \{0, 1\}$:
  $b = 0$: "Which is the isomorphism $S_2 \to S_0$?"
  $b = 1$: "Which is the isomorphism $S_2 \to S_1$?"
  and sends b to the prover as a challenge.
- *Response.* The prover replies with ϕ_{2b}.

The Steiner Triple Systems are combinatorial designs in which the corresponding Isomorphism Problem is not computationally tractable. In [3] it was proved that the isomorphism problem between combinatorial designs is polynomially equivalent to the graph isomorphism problem. For STS, the isomorphism problem can be solved in time $n^{\log n}$ [4], where n is the order of the STS. As far as we know, STS has not been used before for identification purposes. The main practical advantage of our approach is the space economy to specify STS.

The paper is organized as follows. In Section II the basic notions of Steiner Triple Systems are recalled. In Section III the switching cycles of Steiner Triple Systems are introduced as well as the switching transformation. In Section IV our identification protocol is presented. And finally, in Section V some issues about the implementation are discused.

II. Steiner Triple Systems

Steiner Triple Systems (STS), more generally Combinatorial Designs (see [5]), are families of *blocks* of a given base set satisfying certain conditions on smaller blocks. For example, given a base set B, a STS on B is a family of triples (its blocks are 3-sets) such that every pair of elements in B belongs to exactly one 3-set. Combinatorial designs are used to construct experiments in which some balance is required. E. g., in a political debate, the organizers want to confront a group of politicians. Suppose that the candidates are A, B, C, D, E, F, and G. Since seven candidates are too many to debate together, the company decides to form panels with exactly three candidates. Panels should be conformed such that every pair of candidates be confronted exactly in one of them. A solution is given by the triples (A, B, F), (B, C, D), (A, E, D), (A, C, G), (D, F, G), (B, E, G), (C, E, F), which certainly correspond to a STS of order 7. Building combinatorial designs is a hard problem since it is equivalent

to find independent sets [6] in huge graphs, as *Johnson Graphs*. The design of STS is done through standard methods as those of Bose and Skolem [7].

The number of different Combinatorial Designs of a given type explodes in a supra-exponential way as the cardinality of the base set increases. For instance, for orders 7, 9, 13, 15, 19, (namely, congruent with 1, 3 $\bmod 6$) the number of non isomorphic STS are 1, 1, 2, 80, 11 084 874 829. For orders 21 and 25 these numbers are known to be on the order of 10^{13} and 10^{15}, respectively. In 1974 it was conjectured [8] that the number of STS of order n is $\left((1+o(1))\frac{n}{e^2}\right)^{\frac{n^2}{6}}$ by providing as a lower bound $(e^{-5}n)^{n^2/6}$. In 2013 it was proved [9] that $\left((1+o(1))\frac{n}{e^2}\right)^{\frac{n^2}{6}}$ is just an upper bound.

Theorem 2.1 (Wilson): The number of non-isomorphic STS, with respect to n, satisfies [8], [10]:

$$\left((1+o(1))\,\frac{1}{3\sqrt{3}}\,\frac{n}{e^2}\right)^{\frac{n^2}{6}} \leq STS(n) \leq \left((1+o(1))\,\frac{n}{e^2}\right)^{\frac{n^2}{6}}.$$

The 80 non-isomorphic STS of order 15 were found more than 100 years ago. However, the 11 084 874 829 of order 19 were found in 2004 [11] with the help of modern computers, and in this paper it is introduced the cycle graph concept and the switching transformation to convert a STS into another one probably not isomorphic to the original. They used these concepts to generate the non-isomorphic STS of order 19 and on the order of 10^{28} *labeled* STS of the same order.

In [12] it was defined a graph $G(n)$ having the non-isomorphic STS of order n as vertices, two vertices u, v being joined by an edge if an only if a switching transformation on u produces v, or viceversa. It was proved directly that $G(n)$ is connected for $n \leq 19$ and it has been conjectured that it is connected for any possible order. Today this conjecture remains open.

III. Cycle-switching and the switching transformation on STS

As said before, a *Steiner Triple System* (STS) of order n, briefly STS(n), is a collection of 3-element subsets called *triples* or *blocks* of a n-set of points such that every pair of points occurs in exactly one block. Given an STS(n), standard counting arguments prove that each point must occur in exactly $r = \frac{n-1}{2}$ blocks, and that the triple system consists of exactly $b = \frac{n(n-1)}{6}$ blocks. Since both r and b are integers, we get necessary conditions for the existence of an STS(n), which, in fact, turn out to be sufficient. Two STS(n) are *isomorphic* if there is a permutation $\pi : \{0,\ldots,n-1\} \to \{0,\ldots,n-1\}$ such that $T = \{\{\pi(i),\pi(j),\pi(k)\}|\ \{i,j,k\} \in S\}$.

STS(7) is pictured in Figure 1, each monochromatic triangle is a block. For a survey of basic results on STS see [13].

Theorem 3.1: For any $n \geq 3$, there exists a STS(n) if and only if $n \bmod 6 \in \{1,3\}$.

Definition 3.1 (Switching graphs): For a STS(n) S and two elements $a, b \in \{0,\ldots,n-1\}$, a graph $S_{a,b}$ is defined, called the (a-b)-*switching* of S:

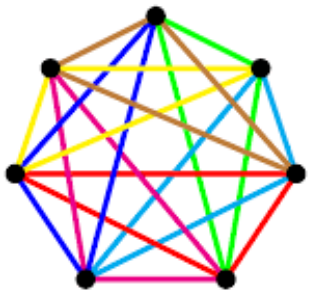


Fig. 1. STS(7): The blocks are formed by the vertices of monochromatic triangles.

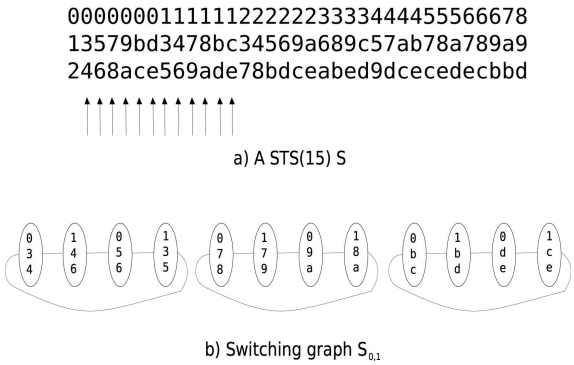


Fig. 2. Switching graph example.

Vertices: $V(S_{a,b})$ is the set of triples in S having either a or b as members but not both:

$$v \in V(S_{a,b}) \iff |v \cap \{a,b\}| = 1.$$

Edges: Two vertices $v_1, v_2 \in V(S_{a,b})$ form an edge, $v_1v_2 \in E(S_{a,b})$, if and only if $v_1 = \{a,x,y\}$ and $v_2 = \{b,x,z\}$, with $x,y,z \in \{0,\ldots,n-1\}$ pairwise different.

In other words, two triples form an edge e in $S_{a,b}$ if and only if one contains a, the other contains b, and both contain a common element x. The elements a and b are called the *axis* and x is the *pivot* of the edge $e = v_1v_2$.

In Figure 2 this definition is illustrated. The blocks are displayed as the columns of a $3 \times r$-array, where r is the number of triples. In a) it is given a STS(15), and the triples in $V(S_{0,1})$ are pointed in by arrows. In b) the graph $S_{0,1}$ is shown. This graph has three connected components and each one is a cycle. This structure appears in every decomposition graph as we prove in the following result.

Lemma 3.2: Let S, a and b be as in definition 3.1. Then,

1) $S_{a,b}$ is bipartite.
2) $S_{a,b}$ is a disjoint union of cycles, henceforth called *switching cycles*.

Proof: The graph $S_{a,b}$ is bipartite because each edge joins one triple containing a and another containing b.

In order to prove the second assertion observe that a vertex in $S_{a,b}$ is a triple, say $\{c_1, a, c_2\}$, adjacent in $S_{a,b}$ to $\{c_3, b, c_1\}$ and $\{c_2, b, c_4\}$ for some indices c_3 and c_4. No other possibility exists since every pair is contained in just one triple. Thus $S_{a,b}$ is two-regular, hence it is a cycle decomposition. □

Proposition 3.3: Let S be a STS(n) and let a, b be two different elements in $\{0, \ldots, n-1\}$. Then:

1) Each pivot i in $\{0, \ldots, n-1\}$ belongs to triples in at most one connected component of $S_{a,b}$.
2) The graph $S_{a,b}$ consists of $n-3$ vertices (namely, triples).

Proof: For the first assertion, suppose, on the contrary, that some pivot c belongs to two different connected components C_1 and C_2 of $S_{a,b}$. Then there exists an edge in C_1, and other in C_2, both with pivot c, say $(\{p_1, b, c\}, \{c, a, p_1'\}) \in E(C_1)$, and $(\{p_2, b, c\}, \{c, a, p_2'\}) \in E(C_2)$. So, $(\{p_1, b, c\}, \{c, a, p_2'\})$ is an edge in $S_{a,b}$ and $\{c, a, p_2'\} \in C_1$. Thus, $C_1 = C_2$ contradicting the choice of C_1 and C_2.

For the second assertion, since S is an STS the number of triples containing a particular index is $\frac{n-1}{2} - 1$. So, $n-3$ triples contain either a or b, but not a and b together. This is the number of vertices in $S_{a,b}$. □

Let S be a STS(n), $a, b \in \{0, \ldots, n-1\}$ different indices, and C a cycle in $S_{a,b}$. The $(a$-$b)$-*switching* transformation on C, denoted $sw_{a,b}(S, C)$, changes all the triples in S contained in C replacing a by b and conversely. In other words,

$$sw_{a,b}(S, C) = \{T \triangle \{a, b\} | T \in C\} \cup (S - C)$$

where $\triangle$ represents the symmetric difference operator.

Theorem 3.4: Let S be a STS(n), $a, b \in \{0, \ldots, n-1\}$, and C a cycle in $S_{a,b}$. Then $sw_{a,b}(S, C)$ is an STS(n) which in general is not isomorphic to S.

Proof: The cycle C can be expressed as $C = (t_0, \ldots, t_{2k-1})$ where $t_{2i} = \{a, p_{2i}, p_{2i+1}\}$, $t_{2i+1} = \{b, p_{2i+1}, p_{2i+2}\}$, $i = 0, \ldots, k-1$, (subscripts taken $\bmod(2k)$). After the switching transformation C is transformed into $C' = (t_0', \ldots, t_{2k-1}')$ with $t_{2i}' = \{b, p_{2i}, p_{2i+1}\}$, $t_{2i+1}' = \{a, p_{2i+1}, p_{2i+2}\}$, and the other triples in S remain unchanged.

We check now that any index pair contained in one triple t_i is contained in just one triple t_j'.

- Pairs of the form $\{p_{2i+\varepsilon}, p_{2i+1+\varepsilon}\}$ are such that $\{p_{2i+\varepsilon}, p_{2i+1+\varepsilon}\} = t_{2i+\varepsilon} \cap t_{2i+\varepsilon}'$, for $\varepsilon \in \{0, 1\}$.
- Pairs of the form $\{a, p_{2i+\varepsilon}\}$ are such that $\{a, p_{2i+\varepsilon}\} = t_{2i} \cap t_{2(i+\varepsilon)-1}'$, for $\varepsilon \in \{0, 1\}$.
- Pairs of the form $\{b, p_{2i+\varepsilon}\}$ are such that $\{b, p_{2i+\varepsilon}\} = t_{2(i+\varepsilon)-1} \cap t_{2i}'$, for $\varepsilon \in \{0, 1\}$.

We prove now, just by providing an example, that it is not the case that, in general, S and $sw_{a,b}(S, C)$ are isomorphic. Consider S as the STS(13):

```
0000001111122222333444556 6
13579B3478A3456A6895797878
2468AC569BC789CBACBBACCAB9
```

For $a = 9$ and $b = A$, $S_{9,A}(S, C)$ contains the cycle

$$\begin{aligned} C &= (\{2,5,9\}, \{2,A,B\}, \{3,9,B\}, \\ &\quad \{3,6,A\}, \{6,8,9\}, \{5,8,A\}) \end{aligned}$$

and

$$\begin{aligned} S' &= sw_{9,A}(S, C) \\ &= (S - C) \cup \{\{2,5,A\}, \{2,9,B\}, \{3,A,B\}, \\ &\quad \{3,6,9\}, \{6,8,A\}, \{5,8,9\}\} \end{aligned}$$

obtaining

```
0000001111122222333444556 6
13579B3478A3456968A5797878
2468AC569BC78ACB9CBBACC9BA
```

In fact, S is isomorphic to the STS(13) number 2, reported in Table 1.27 [5] and S' to the STS(13) number 1, reported in the same table. So, S and S' are not isomorphic. □

IV. Zero-knowledge identification protocol

Identification protocols, in addition to zero-knowledge requirements, must satisfy two conditions: completeness and soundness [14]. Completeness means that if the prover is genuine, then he/she must be able to satisfy the challenges posed by the verifier. Soundness means that if the prover is not genuine, then the probability that he/she provides right consecutive answers to challenge rounds should rapidly tend to zero.

Now we introduce our identification protocol.

a) Key generation: The prover generates the following:

1) An arbitrary STS S of order n and a list of triples $L = [(a_i, b_i, c_i)]_{i=0}^{m-1}$ such that $\forall i \leq m-1$: $a_i, b_i \in \{0, \ldots, n-1\}$, $a_i \neq b_i$, and $c_i \in \mathbb{N}$.
2) A sequence $[S_i]_{i=0}^{m-1}$ of STS(n) such that $S_0 = S$, $\forall i \geq 1$, $S_i = sw_{a_i,b_i}(S_{i-1}, C_i)$ where C_i is the c_i-th switching cycle of $S_{i-1}{}_{a_i,b_i}$ modulus the number of cycles in S_{i-1}. The ordering of the switching cycles in switching graphs is inherited by the lexicographical ordering of the triples.

The public key is (S_0, S_{m-1}) and the private key is L.

b) Challenge-response rounds: The verifier shall be convinced that the prover possesses the private key of the known public key.

- *Commitment.* The prover selects randomly a permutation π in the symmetric group $\mathbb{S}_n$ and builds the STS(n) T_0, T_1 isomorphic, respectively, to S_0 and S_{m-1}. He sends (T_0, T_1) to the verifier as a commitment.
- *Challenge.* The verifier chooses randomly a bit $b \in \{0, 1\}$ and sends it to the prover as challenge. The intention is the following:

  $b = 0$: "Which is the isomorphism between T_0 and S_0?"
  $b = 1$: "Which is the switching transformation on STS that carries T_0 into T_1?"
- *Response.* The prover gets b and responses accordingly.

Certainly, the protocol is complete because the genuine prover knows both the permutation π and the list L. If $b = 1$,

the correct response is $L(\pi) = [(\pi(a_i), \pi(b_i), c_i)]_{i=0}^{m-1}$. The protocol is sound because with k rounds, the probability of k correct replies is 2^{-k}, and it tends to zero as k increases.

The protocol is of zero-knowledge kind. When someone knows the switching transformation from T_0 into T_1, in order to recover the secret key it would be sufficient to solve the Isomorphism Problem for the pair (T_0, S_0). However, an attacker may consider a slight variation of this problem:

Open problem 4.1: Given four STS(n) S_0, S_1, T_0, T_1 such that S_0 and T_0 are isomorphic through a permutation in $\mathbb{S}_n$ and S_1 and T_1 are isomorphic through the same permutation, find that permutation.

Problem 4.1 is slightly simpler than the Isomorphism Problem. However, there is no an immediate modification of the algorithm proposed in [4] to solve the isomorphism problem in $n^{\log n}$ steps. As we said before, in the introduction, the proposed algorithm uses the isomorphism technique and the fact that STS can be expressed in terms of structures related to quasigroups, where the number m of generators is never longer than $\log\, n$. First there are constructed the generators associated to one of the STS and the permutations between them and all the subsets of size m in the other STS are considered. The presence of a second pair of isomorphic STS in Problem 4.1 may help to choose the STS with the smallest number of generators, but it is not hard to construct examples where the bound $\log\, n$ on the number of generators is reached by all the STS appearing in Problem 4.1. So, today, the best method to solve Problem 4.1 require $n^{\log\, n}$ steps. Considering that the paper [4] was published 40 years ago it seems difficult to hope further improvements. The problem of finding the switching transformations from S to S_n could be solved by trying all the STS reachable from S by applying switching transformations. But since there are of the order of 10^{28} labeled STS for orders as small as 19 that can be reached from any STS of order 19 we can see that the problem is extremely hard.

V. Implementation issues

The theoretical time complexity in our identification protocol is super-polynomial, however S_0 must be choosen carefully. If there is a pair a, b such that $(S_0)_{a,b}$ has exactly one connected component then the isomorphism problem between S_0 and T_0 is solvable in polynomial time. In fact, since T_0 is isomorphic to S_0 it is possible to find elements $a_1, \ldots, a_k$ and $b_1, \ldots b_k$, with $1 \leq k \leq \frac{n(n-1)}{2}$, such that the switching graphs in T_0 containing just one cycle are $(T_0)_{a_1,b_1}, \ldots, (T_0)_{a_k,b_k}$. So, in order to find an isomorphism between S_0 and T_0 it is enough to test at most $2k(n-3) \leq n(n-1)(n-3) = O(n^3)$ permutations. In order to guaranty super-polynomial time hardness, a STS S_0 should be selected without switching graphs with few connected components. Suppose for instance, that there exists a pair a, b such that $(S_0)_{a,b}$ contains cycles of lengths $n_1, n_2, \ldots n_l$. Certainly, $n_1 + n_2 + \ldots + n_l = n-3$. Again, T_0 contains elements $a_1, \ldots, a_k$ and $b_1, \ldots b_k$, $1 \leq k \leq \frac{n(n-1)}{2}$, such that all switching graphs containing cycles of lengths $n_1, n_2, \ldots, n_l$ in T_0 are $(T_0)_{a_1,b_1}, \ldots, (T_0)_{a_k,b_k}$. The number of permutation that should be tested to find an isomorphism $S_0 \to T_0$ is upper bounded by

$$U(n, S_0, a, b) = 2^{l-1} n(n-1) n_1 n_2 \ldots n_l. \qquad (1)$$

$U^*(n, S_0) := \min_{a,b}(U(n, S_0, a, b))$ is an upper bound on the number of permutations to be analyzed for finding an isomorphism $S_0 \to T_0$. So, the best option for S_0 is to find the STS(n) S_{max} satisfying

$$U_{max}(n) \triangleq U^*(n, S_{max}) = \max_S(U^*(n, S)).$$

The maximum number of permutations to test for the isomorphism problem within STS(n) is $U_{max}(n)$. We conjecture that this number is super-polynomial wrt n, otherwise the isomorphism problem would be polynomial time solvable, which would be a remarkable result in Design Theory.

Now, we are trying heuristic approximations to find S_{max}. We hope to publish soon the results.

Acknowledgment

The authors acknowledge the partial support of Mexican *Consejo Nacional de Ciencia y Tecnología* (Conacyt).

References

[1] J.-J. Quisquater, L. Guillou, M. Annick, and T. Berson, "How to explain zero-knowledge protocols to your children," in *Proceedings on Advances in Cryptology*, ser. CRYPTO '89. New York, NY, USA: Springer-Verlag New York, Inc., 1989, pp. 628–631. [Online]. Available: http://dl.acm.org/citation.cfm?id=118209.118269

[2] D. Stinson and J. Wu, "A zero-knowledge identification and key agreement protocol," 2007, j32wu@cs.uwaterloo.ca 13712 received 30 Mar 2007, last revised 18 Jul 2007. [Online]. Available: http://eprint.iacr.org/2007/116

[3] M. J. Colbourn and C. J. Colbourn, "Concerning the complexity of deciding isomorphism of block designs," *Discrete Applied Mathematics*, vol. 3, no. 3, pp. 155 – 162, 1981. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0166218X81900123

[4] G. L. Miller, "On the $n^{\log n}$ isomorphism technique: A preliminary report," in *Proceedings of Tenth Annual ACM Symposium on Theory of Computing*. ACM, May 1978, pp. 51–58.

[5] R. Mathon and A. Rosa, "2-(v, k, λ) designs of small order," in *Handbook of Combinatorial Designs*, 2007, pp. 29–45.

[6] R. Diestel, *Graph Theory*, ser. Graduate Texts in Mathematics. Springer, 1997, no. 173.

[7] C. C. Lindner and C. A. Rodger, *Design Theory*, 2nd ed. Chapman & Hall/CRC, 2008.

[8] R. M. Wilson, "Nonisomorphic Steiner triple systems," *Mathematische Zeitschrift*, vol. 135, no. 4, pp. 303–313, Dec 1974. [Online]. Available: https://doi.org/10.1007/BF01215371

[9] N. Linial and Z. Luria, "An upper bound on the number of high-dimensional permutations," *Combinatorica*, vol. 34, no. 4, pp. 471–486, Aug 2014. [Online]. Available: https://doi.org/10.1007/s00493-011-2842-8

[10] ——, "An upper bound on the number of Steiner triple systems," *Random Struct. Algorithms*, vol. 43, no. 4, pp. 399–406, 2013. [Online]. Available: https://doi.org/10.1002/rsa.20487

[11] P. Kaski, V. Mäkinen, and P. R. J. Östergård, "The cycle switching graph of the Steiner triple systems of order 19 is connected," *Graphs and Combinatorics*, vol. 27, no. 4, pp. 539–546, Jul 2011. [Online]. Available: https://doi.org/10.1007/s00373-010-0982-1

[12] M. Grannell, T. Griggs, and J. Murphy, "Switching cycles in Steiner triple systems," *Utilitas Math*, 1999.

[13] C. Colbourn and A. Rosa, *Triple Systems*, ser. Oxford mathematical monographs. Clarendon Press, 1999. [Online]. Available: https://books.google.com.mx/books?id=brpLTq_NMRUC

[14] M. Fürer, O. Goldreich, Y. Mansour, M. Sipser, and S. Zachos, "On completeness and soundness in interactive proof systems," *Advances in Computing Research*, vol. 5, pp. 429–442, 1989.

Generador de ruido Zener para incrementar la entropía de un TRNG

Alfonso Blanco Blanco, Víctor Gayoso Martínez, Luis Hernández Encinas,
Agustín Martín Muñoz, Fausto Montoya Vitini, Amalia B. Orúe López

Instituto de Tecnologías Físicas y de la Información (ITEFI)
Consejo Superior de Investigaciones Científicas (CSIC)
C/ Serrano 144. 28006-Madrid. España
Emails: [alfonso, victor.gayoso, luis, agustin, fausto, amalia.orue]@iec.csic.es

Abstract- **Los generadores físicos de números aleatorios (TRNG) se basan en alguna fuente física de ruido con suficiente entropía y en un mecanismo para extraerla y obtener una secuencia de bits "en bruto", a los que se suele aplicar un posprocesado para obtener la secuencia final de bits aleatorios. La calidad de las secuencias de bits así generadas se evalúa mediante test estadísticos y, además, mediante unos test continuos a la salida de la fuente de ruido blanco para garantizar que no se ha producido una pérdida de entropía. En este trabajo se propone un generador de ruido Zener para mejorar las características de entropía de los osciladores típicamente usados en los TRNG, y se presenta una simulación LabVIEW del comportamiento de un TRNG que incluye test continuos basados en el juego del caos y en la aplicación de retorno, y permite identificar qué parámetros garantizan un nivel de entropía adecuado.**

***Index Terms*- TRNG, ruido Zener, entropía, test continuos**

I. INTRODUCCIÓN

Como es bien sabido, los generadores de números aleatorios son componentes esenciales de los sistemas criptográficos (generación de claves de un solo uso y de números primos, protocolos de desafío-respuesta, etc.). La calidad de las secuencias de bits que generan se evalúa mediante unos test estadísticos. Los más conocidos y utilizados son los test del FIPS 140-2 [1] y los del paquete *SP800-22* del Instituto Nacional de Estándares y Tecnología de los Estados Unidos (NIST, *National Institute of Standards and Technology*) [2], los del *AIS 31* de la Oficina Federal Alemana para la Seguridad de la Información (BSI, *Bundesamt für Sicherheit in der Informationstechnik*) [3], y los de los paquetes *Diehard* [4] y *Dieharder* [5].

Los generadores físicos (TRNG, *True Random Number Generator*) se basan en alguna fuente física de ruido capaz de proporcionar suficiente entropía, y en un determinado mecanismo de extracción de dicha entropía, muestreando adecuadamente la señal obtenida del circuito físico para obtener una secuencia de bits "en bruto" (*raw*), a los que se suele aplicar un posprocesado, que puede ser aritmético o incluso criptográfico, para obtener la secuencia final de bits aleatorios. Hay que subrayar que, en el AIS 31, el posprocesado es opcional salvo para la clase PTG.3, la de requisitos de seguridad más elevados. Para las clases inferiores (PTG.1 y PTG.2), el posprocesado no es un requisito si los números *raw* ya son suficientemente buenos.

En los últimos años han sido varias las iniciativas internacionales para avanzar en el desarrollo de TRNG y de métodos para evaluar su calidad. Destacan la de la Unión Europea a través del proyecto HECTOR [6], y la del NIST [7]. En ambos casos se recomienda elaborar un modelo estocástico de la fuente de ruido (tras un análisis de sus principios de funcionamiento), para poder evaluar adecuadamente el TRNG. El modelo estocástico permite conocer qué factores pueden afectar a la entropía y es preciso poder garantizar que el TRNG proporciona un nivel de entropía mínimo, incluso si los test estadísticos detectan que la calidad de los bits generados no es adecuada [8].

Actualmente, en el caso de los TRNG, los estándares en uso indican que, además de los test estadísticos, es preciso realizar unos test continuos a la salida de la fuente de ruido para garantizar que no se ha producido una pérdida de entropía. La Fig. 1 muestra un esquema de un TRNG y los tipos de test que se deben realizar para analizar la calidad de las secuencias de bit que suministra.

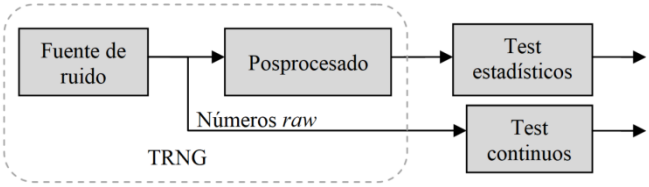


Fig. 1. Esquema de un TRNG y test necesarios [1].

Los test estadísticos mencionados anteriormente se aplican a las secuencias generadas una vez aplicado el posprocesado. Todos los números aleatorios obtenidos por muestreo de un proceso físico tienden a sufrir desviaciones de una estadística ideal. El posprocesado se aplica para intentar reducir esas desviaciones. Tal vez el mecanismo de posprocesado más utilizado es la compresión von Neumann [9].

Los test continuos se aplican a los bits obtenidos en bruto (*raw*). Estos son los bits que se obtienen a partir de la fuente física aplicando un proceso de digitalización (o muestreo, mediante un contador). Existen numerosos tipos de circuitos que pueden ser utilizados como fuentes de ruido físico, la mayoría de ellos basados en osciladores [10], [11], y diversas propuestas de test continuos para TRNG (por ejemplo, [12], [13]).

En este trabajo se propone un generador de ruido blanco Zener para mejorar las características de entropía de los osciladores típicamente usados en los TRNG, y se presenta una simulación LabVIEW de su comportamiento, incluyendo en la herramienta de simulación unos test continuos, basados en el juego del caos y en la aplicación de retorno [14], que permiten

detectar una disminución de entropía de la fuente de ruido blanco.

En la sección II se describe brevemente el generador Zener propuesto, en la sección III se muestran los resultados de la simulación LabVIEW del comportamiento de un TRNG resultante de la combinación del generador de ruido Zener con un oscilador astable típico y, finalmente, la sección IV incluye las principales conclusiones de este trabajo.

II. Circuito Zener generador de ruido blanco

La fuente de aleatoriedad más frecuentemente utilizada en dispositivos digitales es el *jitter* de reloj, entendiendo por *jitter* una variación rápida de un determinado evento, como puede ser una modificación en un flanco de subida o bajada de la señal. Por este motivo, el funcionamiento de la mayoría de los osciladores astables utilizados como fuente de ruido en los TRNG se basa en muestrear (contar) los bits comparando dos señales de distinta frecuencia de reloj, aprovechando el llamado *jitter* temporal, o la inestabilidad de la señal de reloj. Hay diversas formas de generar este tipo de inestabilidad. Como se indica en el bloque denominado oscilador astable en la Fig. 2, se suele utilizar un contador con un reloj estable con señal de cuarzo o similar (con una constante de tiempo CLK_H elevada) que compara su señal con otra proveniente de un reloj menos estable, basada en componentes R-C (con una constante de tiempo CLK_L, baja).

La idea de la fuente de ruido blanco propuesta consiste en inyectar ruido blanco a la entrada del oscilador, con lo que, de un periodo a otro, se consigue aumentar la incertidumbre en la decisión de si se genera un bit 0 o un bit 1. Mediante el generador de ruido Zener se consigue aumentar el *jitter* de manera aleatoria, obteniendo así desviaciones respecto del valor medio, que contribuyen a mejorar la entropía de la fuente.

En el bloque izquierdo de la Fig. 2 se muestra el esquema electrónico del circuito de ruido Zener con unos valores fijados para los distintos parámetros. Dichos valores son totalmente configurables pudiendo ser modificados para obtener diferentes niveles de ruido Zener a inyectar en el oscilador astable.

III. Simulación LabVIEW

El TRNG de la Fig. 2 se ha simulado utilizando LabVIEW. Variando los parámetros del circuito se obtienen diferentes niveles de ruido Zener. En las Figs. 3 y 4 se presentan capturas de pantalla de la herramienta de simulación elaborada para diferente valor de ruido inyectado (alto y bajo, respectivamente).

Como puede observarse en ambas figuras, la herramienta muestra gráficamente en la parte inferior izquierda de la pantalla la relación entre las frecuencias de ambos relojes. La señal ruidosa analógica se muestrea (digitaliza, contando así ceros y unos) a cada flanco de subida (o bajada) de la señal de reloj de referencia. Dependiendo del nivel de ruido blanco, el *jitter* temporal hará que la secuencia digital así obtenida sea más o menos aleatoria.

La parte derecha de las figuras incluye diferentes valores y diagramas que sirven para evaluar la calidad de las secuencias aleatorias obtenidas con los parámetros de ruido establecidos. Como puede observarse, se muestran dos valores (entropía y exponente de Lyapunov), y cuatro diagramas (histograma, transformada rápida de Fourier) el juego del caos y la aplicación de retorno [14].

A. Estimación de la entropía

En teoría de la información, la entropía es una medida de la incertidumbre de una variable aleatoria. La definición más general de entropía es la de *Rényi* pero, por razones prácticas, las medidas de entropía más utilizadas son la entropía mínima y la entropía de Shannon [7].

La entropía mínima $H_\infty(X)$ de una distribución de una variable X que puede tomar valores $x_1,\ldots,x_n$ con unas probabilidades $p_i = \Pr[X = x_i]$ se define como [11]:

$$H_\infty(X) = -log_2(\max_{i=1..n}(p_i)) \ , \quad p_i > 0 \ . \qquad (1)$$

Así, una distribución de una variable discreta X tiene una entropía mínima k si para cualquier elemento x_i de su rango, su probabilidad es $\Pr[X = x_i] \leq 2^{-k}$. La entropía mínima es la medida de incertidumbre más conservadora, porque se garantiza que hay al menos $H_\infty(X)$ bits de información en cada observación de X. Según el estándar del NIST [7] este es el valor que debe usarse para evaluar la calidad de la fuente de ruido blanco.

La entropía de Shannon se define por

$$H(X) = -\sum_{i=1}^{n} p_i \, log_2 \, p_i \ . \qquad (2)$$

Si el tamaño del espacio de salida es 2^k, esto es, el resultado puede representarse usando k bits, el valor máximo de la entropía (con cualquiera de sus medidas) es $H = k$.

Se ha hecho una prueba generando 2^{20} muestras de 8 bits. En la Fig. 3 puede observarse que, cuando se inyecta un alto nivel de ruido blanco en el oscilador, el valor obtenido de la entropía de Shannon es $H(X) = 7.9998$, muy próximo al máximo, siendo el valor de la entropía mínima $H_\infty(X) = 7{,}9501$. Estos son unos excelentes valores, pues están a un nivel similar a los obtenidos con varios PRNG (*Pseudo Random Number Generator*) de comprobada calidad [15, 16, 17] que presentan un mismo valor de la entropía de Shannon, $H\ (X)$= 7.9998, y valores respectivos $H_\infty(X) = 7{,}9329$, $H_\infty(X) = 7{,}9447$ y $H_\infty(X) = 7{,}9498$. Todos satisfacen el valor mínimo de la entropía de Shannon exigido por la norma AIS 31 del BSI: $H\ (X) > 0.997$ por bit, que para 8 bits supone $H\ (X) > 7.9760$. En el ejemplo mostrado en la Fig. 4 el ruido inyectado es bajo y la entropía de Shannon se reduce a $H(X) = 5{,}45089$, por ello se ajusta el nivel de ruido inyectado para obtener rutinariamente la entropía máxima.

B. Test continuos

Además de los test de arranque, diseñados para comprobar el sistema antes del primer uso de la fuente de entropía, los estándares definen diversos test a realizar de manera continua para analizar la calidad de los números aleatorios *raw* generados por un TRNG —también denominados test en línea (*online tests*)— deben poder ser realizados con un número de bits razonablemente bajo (sin llegar a los valores requeridos por los test estadísticos) y han de poder detectar con fiabilidad disminuciones importantes del nivel de entropía esperado en el TRNG, incluyendo una ruptura total de la fuente de entropía (*total failure test*). Además, deben ser rápidos, puesto que no debe salir ningún número aleatorio del generador cuando se detecta algún defecto en la fuente de entropía. Se admite, obviamente, que estos test continuos sean menos precisos que los test estadísticos (las pequeñas desviaciones de la aleatoriedad serán posteriormente corregidas en el posprocesado).

En todos los test continuos resulta muy importante definir el valor umbral a utilizar para que el test haga saltar una alarma

(indicando que el valor de la entropía no es tolerable). Un valor excesivamente restrictivo puede dar lugar a muchas falsas alarmas. Esto no es malo desde el punto de vista de la seguridad, pero puede hacer que el TRNG sea excesivamente lento para determinadas aplicaciones. Es preciso pues encontrar el valor umbral adecuado en un compromiso entre seguridad y velocidad.

En particular, el recientemente publicado estándar del NIST, SP800-90B [7] incluye un test que comprueba que un mismo valor (0 o 1) se repita más de un determinado número de veces seguidas. El NIST da una fórmula para calcular ese número máximo de ceros o unos repetidos en función de la probabilidad aceptable de falsos positivos y de la entropía mínima H_∞ de la fuente de ruido.

En la herramienta de simulación se han incluido dos test gráficos que pueden utilizarse además de los test continuos determinados en los estándares. Se trata del juego del caos y de la aplicación de retorno, herramientas de la criptografía caótica muy útiles para el análisis de la calidad de secuencias aleatorias [14]. Como puede apreciarse en la Fig. 3, cuando se inyecta un alto nivel de ruido blanco en el oscilador no es posible identificar ningún patrón ni usando el juego del caos ni usando la aplicación de retorno. Sin embargo, cuando el nivel de ruido es bajo aparecen claros patrones en ambos diagramas, como puede observarse en la Fig. 4.

El sistema puede ser atacado físicamente, por ejemplo alterando la tensión de alimentación, la temperatura, o sometiéndolo a un campo electromagnético intenso. En estos casos se podría disminuir drásticamente su entropía. Para evitar este tipo de ataques se requiere el uso de contramedidas físicas típicas de cualquier implementación hardware, como sensores térmicos o de luz que desconecten la alimentación y un apantallamiento electromagnético total.

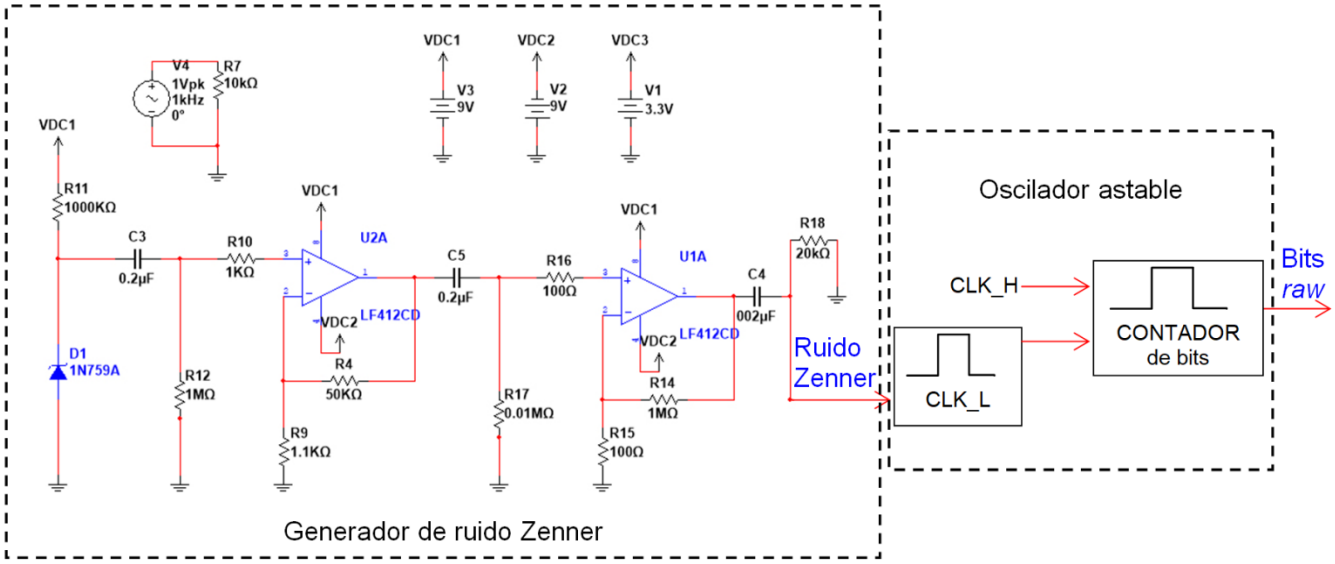


Fig. 2. Esquema de la fuente de ruido blanco propuesta: combinación de un generador de ruido Zener con un oscilador astable.

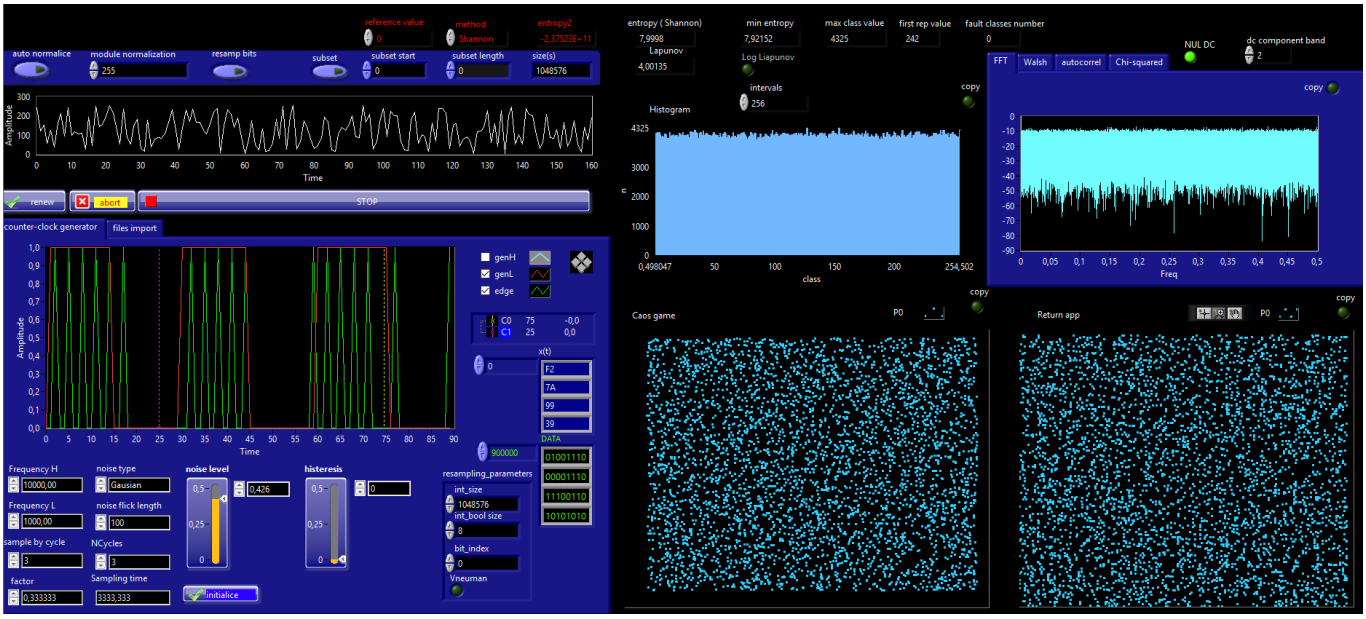


Fig. 3. Captura de pantalla de la herramienta de simulación LabVIEW. Se inyecta un ruido Zener alto.

Fig. 4 Captura de pantalla de la herramienta de simulación LabVIEW. Se inyecta un ruido Zener bajo.

IV. Conclusiones

En este trabajo se propone un generador de ruido Zener para mejorar las características de entropía de los osciladores típicamente usados en los TRNG. Se presenta, además, una herramienta de simulación LabVIEW del comportamiento del TRNG que, además de medir el valor de la entropía, incluye test continuos basados en el juego del caos y en la aplicación de retorno. La herramienta permite identificar de forma casi inmediata los parámetros que garantizan un nivel de entropía adecuado y detecta la ruptura de la fuente de entropía. Los ejemplos mostrados para casos con alto y bajo nivel de ruido Zener inyectado en el oscilador demuestran la utilidad del generador propuesto y de la herramienta de simulación desarrollada.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el Ministerio de Economía, Industria y Competitividad (MINECO), la Agencia Estatal de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER, UE), a través del proyecto COPCIS, referencia TIN2017-84844-C2-1-R, y por la Comunidad de Madrid, a través del proyecto CIBERDINE, referencia S2013/ICE-3095-CIBERDINE-CM, también cofinanciado con fondos FEDER de la Unión Europea.

Referencias

[1] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M.Vangel, D. Banks, A.Heckert, J. Dray, S. Vo, "A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications", NIST Special Publication SP800-22, revision 1a, April 2010.

[2] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M.Vangel, D. Banks, A.Heckert, J. Dray, S. Vo, "A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications", NIST Special Publication SP800-22, revision 1a, April 2010.

[3] W. Killmann and W. Schindler, "A proposal for: Functionality classes for random number generators, version 2.0", *Bundesamt für Sicherheit in der Informationstechnik (BSI)*, 2011.

[4] G. Marsaglia, "The Marsaglia Random Number CDROM including the Diehard Battery of Tests of Randomness", https://web.archive.org/web/20160125103112/http://stat.fsu.edu/pub/diehard/, último acceso abril 2018.

[5] R. G. Brown, D. Eddelbuettel, D. Bauer, "Dieharder: A Random Number Test Suite", *Duke University Physics Department* https://webhome.phy.duke.edu/~rgb/General/dieharder.php, ultimo acceso abril 2018.

[6] HECTOR project, "Hardware Enabled Crypto And Randomness", *H2020-ICT-2014-1 project*, https://hector-project.eu/, ultimo acceso abril 2018.

[7] M. S. Turan, E. Barker, J. Kelsey, K. McKay, M. L. Baish, M. Boyle, "Recommendation for the Entropy Sources Used for Random Bit Generation", NIST Special Publication SP800-90B, January 2018.

[8] V. Fischer and P. Haddad, "Random Number Generators for Cryptography", ch. 7 of Circuits and Systems for Security and Privacy, serie Devices, Circuits, Systems, L. S. Farhana Sheikh, Eds., CRC Press, pp. 245–286, 2016.

[9] M. Dichtl "Bad and Good Ways of Post-processing Biased Physical Random Numbers", *Proceedings of Fast Software Encryption, LNCS vol. 4593*, pp 137-152, 2007.

[10] B. Sunar, "True Random Number Generators for Cryptography", ch. 4 of *Cryptographic Engineering*, C.K. Koç, Ed., Springer Publishing Company, Incorporated, pp. 55–73, 2009.

[11] M. Stipčević and C. K. Koç, "True Random Number Generators", ch. 11 of *Open Problems in Mathematics and Computational Science*, Springer, pp. 275–315, 2014.

[12] V. Rožić, "Circuit-Level Optimizations for Cryptography. True random number generators: design, evaluation and testing", Ph.D. dissertation, Faculty of Engineering Science. KU Leuven (Belgium), September 2016, https://www.esat.kuleuven.be/cosic/publications/thesis-286.pdf.

[13] B. Yang, V. Rožić, N. Mentens, W. Dehaene, and I. Verbauwhede, "Embedded HW/SW Platform for On-the-fly Testing of True Random Number Generators", Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 345–350, 2015.

[14] A. B. Orúe López, A. Fúster Sabater, V. Fernández Mármol, F. Montoya Vitini, L. Hernández Encinas, A. Martín Muñoz, "Herramientas visuales usadas en criptografía caótica útiles para el análisis de secuencias pseudoaleatorias", *Actas de la XIV Reunión Española sobre Criptología y Seguridad de la Información*, pp. 180-185, 2016, http://recsi16.uib.es/wp-content/uploads/2016/10/ACTAS-RECSI-2016.pdf.

[15]Orúe A.B, Hernández, L. and Montoya, F. (2010). Trifork, a new Pseudorandom Number Generator Based on Lagged Fibonacci Maps. Journal of Computer Science and Engineering, 2(2):46-51. ISSN: 2043-9091.

[16] Orúe A.B., Álvarez G., Guerra A., Pastor G., Romera M. and Montoya F. , "Trident, a new Pseudo Random Number Generator Based on Coupled Chaotic Maps". Editors: Herrero A., Corchado E., Redondo C. and Alonso A. *Computational Intelligence In Security For Information Systems Advances in Intelligent and Soft Computing*, 85, pp. 183-190. Springer, 2010.

[17] A. B. Orúe López, L. Hernández Encinas, A. Martín Muñoz, F. Montoya Vitini. "A Lightweight Pseudorandom Number Generator for Securing the Internet of Things". *IEEE Access*, vol. 5, 2017.

Linear cellular automata, bivariate polynomials and 2^r-purely periodic binary sequences

(An extended abstract)

Domingo Gomez-Perez
Universidad de Cantabria
Department of Mathematics, Statistics and Computer Science
39005, Spain
domingo.gomez@unican.es

Ana I. Gomez
Universidad de Cantabria
Department of Mathematics, Statistics and Computer Science
39005, Spain
gomezperezai@unican.es

Jaime Gutierrez
Universidad de Cantabria
Department of Applied Mathematics and Computer Science
39005, Spain
jaime.gutierrez@unican.es

***Abstract*—In this paper, we study properties of 2^r-purely periodic binary sequences generated by cellular automata with specific rules. We obtain bounds on the lengths of these cellular automata and we propose a new and more efficient method to attacking such non-linear generators.**

***Index Terms*—Two dimensional binary array, cellular automata, shrinking generator, polynomials.**

I. Bivariate polynomials and two-dimensional cellular automata

We denote by $\mathbb{F}_2[X,Y]$ the polynomial ring in the variables X and Y, with coefficients over the prime finite field $\mathbb{F}_2$ with two elements $\{0,1\}$.

On the other hand, we consider infinite and periodic two-dimensional binary array, represented by a binary matrix. Each element in row $i \geq 0$ and column $j \geq 0$ is represented by the monomial X^iY^j. It is called the **cell** X^iY^j and the content is an element of $\mathbb{F}_2$. A general way to construct two-dimensional binary array is starting with a **initial state**, namely the content of a list of cells and **rules**, namely a list polynomials satisfying the content of the cells in the field $\mathbb{F}_2$ and generating (as polynomial ideal generated by the list of the polynomials) the content of rest of the cells, see [1] for a mathematical formalization. Here, we illustrate these concepts by an example.

Example 1: In Fig. 1 we show a Linear Feedback Shift Register(LFSR) implementation of a recurrence in an array, cells are joined by wires and signal transmit after one clock step. The time needed to populate the array is 11 clock-steps. The rules used to generate the two dimensional array

$$\begin{array}{ccccccc} 0 & 1 & 1 & 1 & 1 & . \\ 0 & 0 & 1 & 1 & 0 & . \\ 0 & 1 & 0 & 0 & 1 & . \\ . & . & . & . & . & . \end{array}$$

are

$$Y^5+1, X^3+1, 1+XY+XY^4$$

and the seed or initial state is:

$$X^0Y^0, X^0Y^3, XY, X^2Y^2 = 0100$$

Starting with the initial state:

$$\begin{array}{ccccccc} \underline{\mathbf{0}} & 1 & 1 & \underline{\mathbf{1}} & 1 & . \\ 0 & \underline{\mathbf{0}} & 1 & 1 & 0 & . \\ 0 & 1 & \underline{\mathbf{0}} & 0 & 1 & . \\ . & . & . & . & . & . \end{array}$$

1) The first clock-step takes the content of cell X^0Y^0, namely 0 and, the content of the cell XY, namely 0. Since $1+XY+XY^4$ is a rule, then $0+0+XY^4$ is 0. In other words, the content of the cell XY^4 is 0. Now, we have the content of 5 cells: the initial state and the new generated XY^4:

$$\begin{array}{ccccccc} \underline{\mathbf{0}} & 1 & 1 & \underline{\mathbf{1}} & 1 & . \\ 0 & \underline{\mathbf{0}} & 1 & 1 & \underline{\mathbf{0}} & . \\ 0 & 1 & \underline{\mathbf{0}} & 0 & 1 & . \\ . & . & . & . & . & . \end{array}$$

2) The second clock-step takes the content 0 of the cell XY and the content 0 of the cell X^2Y^2. We use the rule $1+XY+XY^4$ multiplied by XY, that is, $XY+X^2Y^2+X^2Y^5$. Applying the rule the Y^5+1, we get $XY+X^2Y^2+X^2$, then the content of the cell X^2 is 0. Now, we have the content of 6 cells:

$$\begin{array}{ccccccc} \underline{\mathbf{0}} & 1 & 1 & \underline{\mathbf{1}} & 1 & . \\ 0 & \underline{\mathbf{0}} & 1 & 1 & \underline{\mathbf{0}} & . \\ \underline{\mathbf{0}} & 1 & \underline{\mathbf{0}} & 0 & 1 & . \\ . & . & . & . & . & . \end{array}$$

3) The third clock-step takes the content 0 of the cell X^2Y^2 and the content 1 of the cell X^0Y^3. And the rule $1+XY+XY^4$ multiplied by X^2Y^2, that is, $X^2Y^2+X^3Y^3+X^3Y^6$. Using in this last polynomial the rules Y^5+1 and X^3+1, we get $X^2Y^2+Y^3+Y^6 = X^2Y^2+Y^3+Y^5Y = X^2Y^2+Y^3+Y$, then the content of the cell Y is 1. Now, we have the content of 7 cells:

$$\begin{array}{ccccccc} \underline{\mathbf{0}} & \underline{\mathbf{1}} & 1 & \underline{\mathbf{1}} & 1 & . \\ 0 & \underline{\mathbf{0}} & 1 & 1 & \underline{\mathbf{0}} & . \\ \underline{\mathbf{0}} & 1 & \underline{\mathbf{0}} & 0 & 1 & . \\ . & . & . & . & . & . \end{array}$$

And so on till complete the two-dimensional binary array on 11th clock-step.

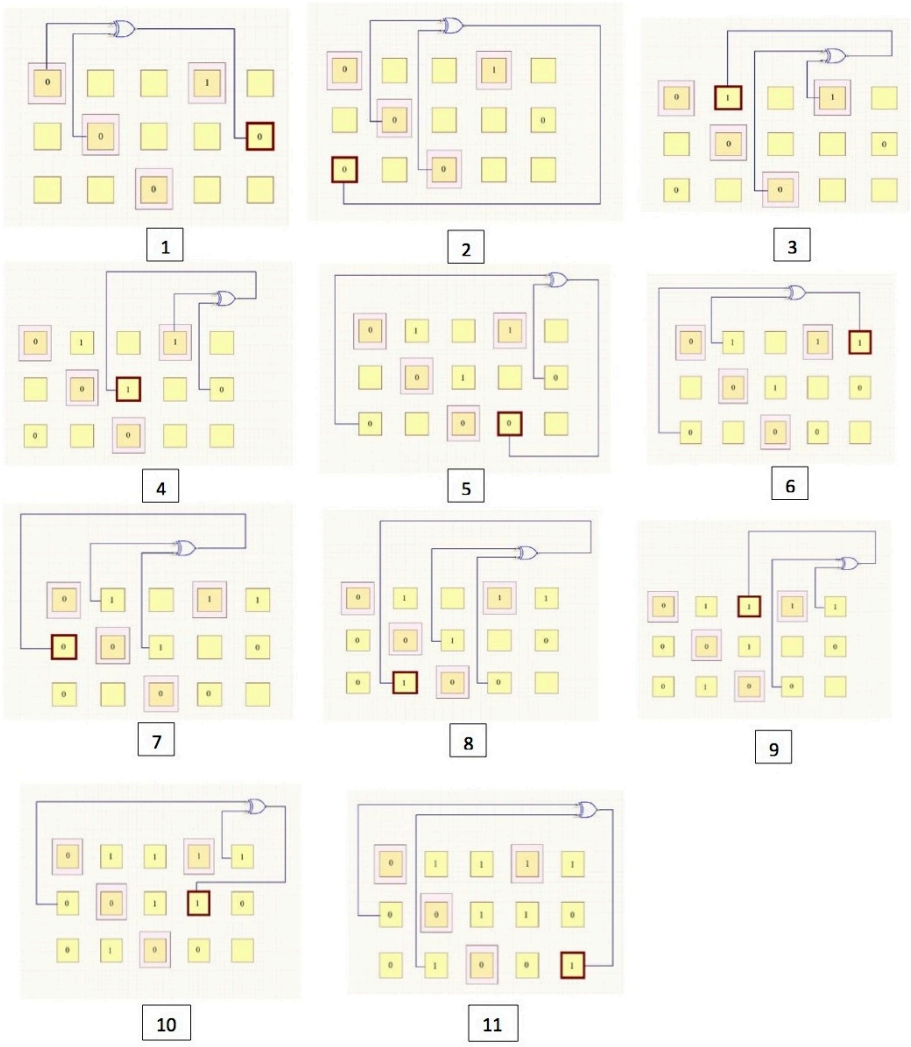


Figure 1. An array generated by a LFSR implementation.

A particular class of two dimensional arrays is the two-dimensional cellular automata (CA) of length l, see [3] and [4]. The CA is represented as a binary matrix with l columns and the cells notated x_i^t, where t represents the number of the row and i the column. The index t and i represents the i-th **cell** at time t. The content or state of a cell is an element of $\mathbb{F}_2$. A **rule** or local transition function determines the evolution of CA through time; that is, it governs the changes of the contents of each cell at a given instant $t+1$ taking into account its neighborhood at time t. To each rule, we associate a polynomial of $\mathbb{F}_2[X,Y]$. The identification is the following, for each term of the rule of the form $x_{i+m}^{t+k}, m, k \geq 0$ we associate the monomial $X^m Y^k$. The rule is associated to the polynomial which is sum of the monomials.

We are considering only *uniform or regular* automata (every cell follows the same rule) and *null* (cells with null content are added to the extreme left and right cells).

Example 2: Taking the rule 102, defined as $x_i^{t+1} + x_i^t + x_{i+1}^t = 0$, it is associated to polynomial $Y + 1 + X$. Notice that the same association can be done for rules in the columns. In Table I, taken from [5, table 1], the automata satisfies a recursion in the columns $x_i^{t+3} + x_i^{t+2} + x_i^{t+1} + x_i^t = 0$, which corresponds to the polynomial $Y^3 + Y^2 + Y + 1$.

A way to construct a uniform cellular automata from a sequence is to take a rule, put the sequence as the first column and try to fill up the rest of the matrix that satisfies the CA.

The previous example is quite illustrative for several reasons. First of all, all the columns satisfy $x_i^{t+3} + x_i^{t+2} + x_i^{t+1} + x_i^t = 0$, not only the first one. In this case, we say that the polynomial $Y^3 + Y^2 + Y + 1$ is **valid for the cellular automata** o the polynomial satisfied the CA. Also, there are other polynomials that are satisfied by this cellular automata, for example $x_{i+1}^{t+1} + x_{i+1}^t + x_{i+2}^t = 0$, which is associated to the polynomial $X(Y + 1 + X)$. Indeed there is an infinite number of valid polynomials for this cellular automata.

Table I
CELLULAR AUTOMATA CORRESPONDING TO THE SELF-SHRUNKEN

102	102	102
0	1	1
1	0	1
1	1	1
0	0	1
0	1	1
1	0	1
1	1	1
0	0	1

The following Theorem can be proven using the results and mathematical formalization in [1] and [2].

Theorem 1: Given a matrix defining an uniform cellular automata, such that all columns can be generated by the same LFSR, then the set I of all valid polynomials for that CA is the polynomial ideal generated by the characteristic polynomial $C(Y)$ generating all the columns and $R(X,Y)$ the polynomial associated to the rule, that is:

$$I = \{C(Y)h_1 + R(X,Y)h_2 \mid \forall\, h_1, h_2 \in \mathbb{F}_2[X,Y]\}.$$

Proof: The proof is based on Lemma 1 of [2]. ■

In Example 2, the polynomial $C(Y)$ is Y^3+Y^2+Y+1 and the polynomial $R(X,Y)$ is $1+X+Y$. We have that

$$X^4 = (1+Y+Y^2+Y^3)(1+Y)+(1+X+Y)(1+X+Y)^3$$

That is, $x^t_{i+4} = 0$ and, consequently the CA has length 3. On the other hand, in Table I we have that the content of the cells in last column is 1, then we have the rule $x^t_{i+2} + x^{t+1}_{i+2} = 0$ and, in terms of polynomial means $(Y+1)X^2 \in I$.

II. The 2^r-purely periodic binary sequences and bounds on the length of CA

A binary sequence $\{a_i\}$ is 2^r-purely periodic if $a_i = a_{i+2^r}$ for $i \geq 0$. The following result is probably known and, in any case the proof is straightforward.

Lemma 1: A binary sequence $\{a_i\}$ is 2^r-purely periodic if and only if the minimal polynomial is of the form $(1+X)^n$ for some integer n.

A very well known 2^r-purely periodic sequence is the denominated **self-shrunken** sequence. The self-shrinking generator [6], is a particular case of the celebrated **shrinking generator** [7], where an $\{a_i\}$ sequence generated by a maximum length LFSR is self-decimated following this: given two consecutive bits $\{a_{2i}, a_{2i+1}\}$ the output sequence $\{s_j\}$, called the self-shrunken sequence, is:

- If $a_{2i} = 1$, then $s_j = a_{2i+1}$
- If $a_{2i} = 0$, then a_{2i+1} is discarded.

The following result appears in [5]:

Theorem 2 (Theorem 2 in [5]): Given the self-shrunken sequence $\{s_i\}$ with characteristic polynomial $(Y+1)^n$, the uniform, null CA that only uses rule 102 and generates the sequence $\{s_i\}$ has length n.

We can generalize this theorem even further. We are going to denote **rule** R^k to the rule associated to the polynomial $(Y+1)+X^k$.

Theorem 3: Given a 2^r-purely periodic binary sequence $\{s_i\}$ with characteristic polynomial $(Y+1)^n$, any uniform, null CA that only uses rule R^k and generates the sequence $\{s_i\}$ has length nk.

Proof: By the above Lemma 1, the sequence has minimal polynomial $(Y+1)^n$ and by Theorem 1, we have that the set of valid polynomials is exactly the polynomial ideal I, where

$$I = \{(Y+1)^n h_1 + (Y+X^k+1)h_2; \forall\, h_1, h_2 \in \mathbb{F}_2[X,Y]\},$$

Now, it is easy to check that X^{nk} belongs to this ideal, because, we have $X^k = Y+1 \mod I$, then $X^{nk} = (Y+1)^{nk} \mod I$. This means that $x^t_{i+nk} = 0$. On the other hand, it is straightforward to check that $\{X^{nk}, Y+X^k+1\}$ is Groebner basis ([8]) of the ideal I with respect to the lexicographical monomial ordering with variable Y bigger than variable X, and the normal form of the polynomial X^{nk-1} is obviously X^{nk-1}, namely the polynomial X^{nk-1} does not belong to I, which implies the result. ■

III. Attacking 2^r-purely periodic binary sequences through CA

The **linear complexity** of sequence is the length of the shortest LFSR that can be generate such sequence.

Let LC be the linear complexity of 2^r-purely periodic binary sequence. Obviously, given LC consecutive bits of the sequence, then we can recover the whole sequence. Since we can generate the sequence trough CA using the rule 102, it is enough to intercept $LC-1$ consecutive bits. The complexity of the method in [9] to attack the so called **Generalized self-shrinking generator (GSSG)**, which is also a 2^r-purely periodic binary sequence, is

$$\frac{(LC-1)(LC-2)}{2}$$

that is, the number of XOR operations needed to obtain the sequence is $O(LC^2)$. The idea of [9] is computing the half intermediate cells. In our approach, we can compute directly the unknown bit using properties of the ideal I generated by the polynomials $(Y+1)^{LC}, X+Y+1$. We have that the monomials

$$1, Y, Y^2, \ldots, Y^{LC-2}, X^{LC-1}$$

modulo the ideal I is a basis of the $\mathbb{F}_2$ vector space $\mathbb{F}_2[X,Y]/I$. So, the class Y^{LC-1} modulo the ideal I is a linear combination of the above basis, in other words, we have to compute $(Y+1)^{LC-1}$ which gives a total complexity of $O(LC \log LC)$.

On the other hand, given $LC-2$ consecutive bits, the probability to recover the sequence is $1/2$.

IV. Conclusions

In this paper we have modeled the 2^r-purely periodic binary sequence by cellular automata and bivariate polynomials. We have obtained in a easy way some results about the sequences and also the associated CA. We think that some other results on this area could be improved by this approach.

Acknowledgements

This work was partially supported by the Spain Ministery Economía y Competitividad under grant number MTM2014-55421-P.

References

[1] Sakata, Shojiro: " Finding a Minimal Set of Linear Recurring Relations Capable of Generating a Given Finite Two-Dimensional Array", *J. Symb. Computation* col. 5(3), pp.321-337, 1988.

[2] Sakata, Shojiro: "Extension of the Berlekamp-Massey algorithm to N dimensions", *Information and Computation*, vol. 84-2, pp.2017-239, 1990.

[3] Wolfram S. "Cellular Automat" *Los Alamos Science* vol. 9, pp. 2 - 21, 1983.

[4] Das, A.K, A. Ganguly, A, . Dasgupta, A., S. Bhawmik, S., Chaudhuri,P., : "Efficient characterisation of cellular automata", *IEEE Proc Comp. Dig. Techn.*, vol. 137, pp. 81-87, 1990.

[5] Cardell, Sara and Fuster-Sabater, Amparo: ”Linear models for the self-shrinking generator on CA”, *Journal of Cellular Automata*, vol.11, pp. 195-2011, 2016.

[6] Willi Meier, Othmar Staffelbach: ”The self-shrinking generator”, *Advances in Cryptology – EUROCRYPT, LNCS*, vol. 950, pp. 205–214, 1994.

[7] Don Coppersmith, Hugo Krawczyk, Yishay Mansour. ”The shrinking generator”, *In Advances in Cryptology – CRYPTO ’93, LNCS*, vol. 773, pp.23–39, 1993.

[8] Cox, David A, Little, John, O’Shea, Donal,:” Ideals, Varieties, and Algorithms, An Introduction to Computational Algebraic Geometry and Commutative Algebra”, *Undergraduate Texts in Mathematics, Springer Verlag*, Second Edition 1997.

[9] Cardell, Sara and Fuster-Sabater, Amparo: ”Discrete linear models for the generalized self-shrunken sequences”, *Finite Fields and Their Applications*, vol. 4, pp. 222–241, 2017.

Sobre secuencias pseudoaleatorias a partir de autómatas celulares y funciones bent

Joan-Josep Climent
Departament de Matemàtiques
Universitat d'Alacant
jcliment@ua.es

Verónica Requena
Departament de Matemàtiques
Universitat d'Alacant
vrequena@ua.es

***Resumen*—En este artículo construimos un generador de secuencias de bits pseudoaleatorias usando autómatas celulares cuyas funciones de transición local son funciones equilibradas obtenidas a través de funciones bent.**

***Palabras Clave*—Generador de secuencias pseudoaleatorias, autómata celular, función bent, función equilibrada, criptografía**

I. Introducción

En muchas aplicaciones importantes es necesario generar secuencias de números aleatorios, por ejemplo, en simulaciones como las técnicas de Monte Carlo, en los cifrados en flujo, en los métodos de optimización estocástica, en la trasmisión de datos de un satélite, en la generación de claves, etc. Un generador de números aleatorios es un algoritmo que proporciona como salida una secuencia de números que son estadísticamente independientes. Sin embargo, la imposibilidad práctica de que una secuencia obtenida de forma aleatoria se pueda volver a obtener exactamente de la misma forma, hace que estos generadores no sean prácticos en algunas aplicaciones criptográficas, como por ejemplo los cifrados en flujo. Un generador de secuencias pseudoaleatorias es un algoritmo determinista en el que al darle como entrada una secuencia aleatoria de números, nos devuelve como salida una secuencia de longitud mayor y que parece ser aleatoria. Sin embargo, estas secuencias generadas no son aleatorias, pues cada vez que se ejecute el algoritmo con los mismos valores se obtendrán los mismos resultados; pero estas secuencias han de superar unos análisis de aleatoriedad a través de diferentes tests para ser aceptadas como secuencias pseudoaleatorias.

Un Autómata Celular (AC) es un sistema dinámico discreto el cual involucra reglas simples deterministas. Es adecuado para modelar sistemas naturales o artificiales que puedan ser descritos como una colección de objetos simples que interactúan localmente unos con otros, como por ejemplo, en el estudio de la evolución de ecosistemas, en la simulación de fenómenos físicos y biológicos complejos [29], [43], en el estudio del caos [14], entre muchos ejemplos. Los autómatas celulares ofrecen un gran número de ventajas sobre los generadores de números aleatorios tales como la simplicidad algorítmica y la fácil implementación en hardware.

Los generadores de secuencias pseudoaleatorias basados en autómatas celulares han sido ampliamente estudiado en las últimas décadas [13], [16], [22]. En 1986, Wolfram [42] aplicó por primera vez un autómata celular en la generación de números pseudoaleatorios. Otros autores [32], [36] también usaron los autómatas celulares para generar secuencias pseudoaleatorias para su uso en criptografía.

Las funciones booleanas juegan un papel muy importante en la criptografía moderna y son la pieza fundamental de numerosos criptosistemas gracias a su habilidad para proporcionar seguridad en las comunicaciones. El estudio de las funciones booleanas, tanto desde una perspectiva teórica como práctica, es crucial en la provisión de seguridad en las aplicaciones criptográficas como los cifradores en bloque, los cifradores en flujo y las funciones hash. Las propiedades de no linealidad y equilibrio son dos criterios esenciales criptográficamente para las funciones booleanas.

La no linealidad es la propiedad más importante en cualquier criptosistema de clave simétrica para alcanzar confusión. La familia de funciones bent son las funciones booleanas con la máxima no linealidad posible; estas funciones no son equilibradas, pero a partir de ellas se pueden construir funciones equilibradas. El uso de las funciones bent puede generar más seguridad debido a sus buenas propiedades criptográficas, además de proporcionar no linealidad al autómata.

En este artículo, construimos un generador de secuencias pseudoaleatorias usando un autómata celular finito y unidimensional donde las funciones de transición local se basan en funciones equilibradas que se obtienen a partir de funciones bent.

II. Preliminares

En esta sección introducimos los conceptos principales sobre los autómatas celulares, los generadores de secuencias pseudoaleatorias y las funciones bent que utilizaremos en el resto del artículo.

II-A. Autómata Celular

Los autómatas celulares fueron originalmente concebidos por Ulam [37] y von Neumann [38] en la década de los cuarenta del siglo pasado para proporcionar un marco en el que investigar el comportamiento de sistemas complejos y extendidos. El valor de su importancia y su aplicabilidad no fue reconocida hasta finales de la década de los ochenta, gracias a los trabajos de Wolfram [41], [42].

Los autómatas celulares son sistemas dinámicos discretos formados por un número finito o infinito de objetos idénticos llamados **celdas** (o células) [39]. El espacio, el tiempo y los estados del sistema son discretos. Cada celda es un punto en una cuadrícula en la cual la evolución o cambio de estado de cada celda depende de su estado actual y de los estados de sus vecinos inmediatos (véase [33], [40]). Los estados de las celdas se actualizan de acuerdo con una regla local, denominada **regla de transición**. Es decir, el estado de una

celda en un instante dado depende sólo de su propio estado en el instante anterior y de los estados de sus vecinos cercanos en el instante anterior.

Los autómatas celulares más utilizados suelen ser unidimensionales o lineales, es decir, las celdas se disponen según una línea recta; y también bidimensionales, donde las celdas se distribuyen según un plano. Nosotros nos centraremos en los **autómatas celulares finitos y unidimensionales**, que pueden ser definidos como una 4-tupla $\mathcal{A} = (C, S, V, f)$ donde:

1. C es el espacio del AC; es un vector de m celdas, donde cada una de ellas la denotamos por $\langle i \rangle$, para $i = 0, 1, \ldots, m-1$. Así, representamos un AC lineal con 6 celdas como:

$\langle 0 \rangle$	$\langle 1 \rangle$	$\langle 2 \rangle$	$\langle 3 \rangle$	$\langle 4 \rangle$	$\langle 5 \rangle$

2. S es el conjunto de todos los valores posibles que puede tomar cada celda. Denotamos por $a_i^{(t)} \in S$, el estado de la celda $\langle i \rangle$ en el instante t.
3. V es el conjunto de celdas cuyos estados en el instante t influyen en el estado de una determinada celda en el instante $t+1$. Así, llamamos **vecindad** de la celda $\langle i \rangle \in C$ al conjunto V_i formado por las celdas de C cuyo estado en el instante t influirá en el estado de $\langle i \rangle$ en el instante $t+1$ de acuerdo con f.
4. f es la **función de transición local** o **regla de transición** que determina la evolución del AC a través del tiempo; es decir, rige los cambios de los estados de cada celda teniendo en cuenta su vecindad. Así, $f : S^k \longrightarrow S$ con k el tamaño de las vecindades. El conjunto de todas las posibles reglas a tener en cuenta para un autómata celular es numerable. Se puede considerar una única función de transición (es decir, f no depende de t) o varias funciones de transición (es decir, f depende de t).

Aunque existen muchas clases de vecindades, en este artículo consideramos solamente vecindades simétricas; es decir, vecindades para las que la celda $\langle i \rangle$ es la celda central. Las representamos como

$$V_i^{(r)} = \{\langle i-r \rangle, \ldots, \langle i-1 \rangle, \langle i \rangle, \langle i+1 \rangle, \ldots, \langle i+r \rangle\},$$

y su tamaño es $2r+1$ con r el **radio de vecindad**. Para este tipo de vecindades, el estado de la celda $\langle i \rangle$ en el instante $t+1$ viene dado por

$$a_i^{(t+1)} = f\left(a_{i-r}^{(t)}, \ldots, a_{i-1}^{(t)}, a_i^{(t)}, a_{i+1}^{(t)}, \ldots, a_{i+r}^{(t)}\right) \quad (1)$$

con $f : S^{2r+1} \longrightarrow S$ la función de transición local.

Cuando $i < r$ o $i > m-r$, la determinación de la vecindad $V_i^{(r)}$ depende de determinadas condiciones de contorno del AC. En este artículo consideramos condiciones de contorno periódicas; es decir, suponemos que las m celdas del AC están dispuestas uniformemente según una circunferencia donde las celdas del autómata celular se comportan de forma cíclica; así, a la izquierda de la celda $\langle 0 \rangle$ estarán las celdas vecinas $\langle m-1 \rangle$, $\langle m-2 \rangle$, etc. y a la derecha de la celda $\langle m-1 \rangle$ estarán las celdas vecinas $\langle 0 \rangle$, $\langle 1 \rangle$, etc. Formalmente, esto implica que los índices de las celdas se calculan módulo m.

El conjunto de estados de todas las celdas en el instante t se denomina **configuración del AC en el instante** t y se representa por el vector

$$C^{(t)} = \left(a_0^{(t)}, a_1^{(t)}, \ldots, a_{m-1}^{(t)}\right) \in S^m.$$

En particular, $C^{(0)}$ es la **configuración inicial**, también llamada **punto semilla**. Por lo tanto, la **evolución** del AC está determinada por la secuencia de configuraciones $\left(C^{(0)}, C^{(1)}, C^{(2)}, \ldots\right)$ que podemos representar mediante una matriz (o tabla) conocida como **diagrama de evolución**, donde cada una de sus filas corresponde a cada una de las configuraciones $C^{(t)}$, para $t = 0, 1, 2, \ldots$ Por ejemplo, si consideramos un AC con $m = 6$ celdas, tendríamos:

$C^{(0)}$:	$a_0^{(0)}$	$a_1^{(0)}$	$a_2^{(0)}$	$a_3^{(0)}$	$a_4^{(0)}$	$a_5^{(0)}$
$C^{(1)}$:	$a_0^{(1)}$	$a_1^{(1)}$	$a_2^{(1)}$	$a_3^{(1)}$	$a_4^{(1)}$	$a_5^{(1)}$
$C^{(2)}$:	$a_0^{(2)}$	$a_1^{(2)}$	$a_2^{(2)}$	$a_3^{(2)}$	$a_4^{(2)}$	$a_5^{(2)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

II-B. Generadores de secuencias pseudoaleatorias

En la introducción hemos comentado la importancia de los generadores de secuencias aleatorias o pseudoaleatorias, debido a su aplicación en diversas áreas de la ciencia. En esta sección introducimos estos conceptos y comentamos los tests estadísticos que hemos utilizado para estudiar la aleatoriedad de nuestro generador.

Definición 1: Un **generador de bits aleatorios** es un dispositivo o algoritmo que proporciona como salida una secuencia de bits estadísticamente independientes.

Definición 2: Un **generador de bits pseudoaleatorios** es un algoritmo determinista que genera un secuencia binaria de longitud p que "parece" aleatoria, a partir de una secuencia binaria verdaderamente aleatoria de longitud m, con $p \gg m$. La entrada del generador se llama **semilla** y la secuencia de salida se conoce como **secuencia de bits pseudoaleatorios** o **secuencia pseudoaleatoria**.

La salida de un generador de bits pseudoaleatorios no es aleatoria, ya que éste produce siempre la misma secuencia de salida para una semilla fija. Sin embargo, podemos conseguir que estos generadores nos proporcionen salidas que parezcan aleatorias. Para ello, necesitamos que posean buenas propiedades de aleatoriedad para prevenir ataques estadísticos y que el generador sea criptográficamente seguro.

No se dispone de ninguna demostración matemática que asegure la aleatoriedad de una secuencia de bits; no obstante, existen diversas pruebas estadísticas que si son superadas por el generador, éste podría ser aceptado como generador de secuencias paseudoaleatorias. En nuestro estudio, hemos considerado las seis pruebas estadísticas más utilizada, que nos permiten verificar si las secuencias obtenidas se pueden considerar aleatorias (véase, por ejemplo, [23]):

1. **Test de Frecuencia (Monobit Test)**: El propósito de esta prueba es determinar si el número de unos y ceros en una secuencia es aproximadamente el mismo, que es lo que se esperaría de una secuencia verdaderamente aleatoria. Todas las pruebas posteriores dependen de la

superación de esta prueba; no hay evidencia que indique que la secuencia probada no sea aleatoria.

2. **Test de Series (Two-bit Test)**: El propósito de esta prueba es determinar si el número de ocurrencias de 00, 01, 10 y 11 como subsecuencias de la cadena a comprobar, son aproximadamente las mismas, como se esperaría de una cadena de bits aleatorios.
3. **Test de Póker**: Sea q un entero positivo tal que $\left\lfloor \frac{p}{q} \right\rfloor \geq 5 \cdot (2^q)$, y $k = \left\lfloor \frac{p}{q} \right\rfloor$. Dividamos la secuencia s en k partes no solapadas, cada una de ellas de longitud q; y sea n_j el número de ocurrencias del tipo j-ésimo de la secuencia de longitud q, para $j = 1, 2, \ldots, 2^q$. Este test determina si cada una de las secuencias de longitud q aparece aproximadamente el mismo número de veces en s.
4. **Test de Rachas (Runs Test)**: El test toma una secuencia s y calcula el número de subsecuencias de s formadas por una determinada cantidad consecutiva de ceros (huecos) o de unos (bloques), llamadas rachas. El fallo de este test indica que hay una mala distribución de las longitudes de las rachas o que no hay rachas registradas por encima de una cierta longitud que es la esperada para las transmisiones del tamaño de la muestra.
5. **Test de Autocorrelación**: El objetivo de este test es comprobar las correlaciones entre la secuencia s y las secuencias obtenidas al desplazar (sin rotación) la secuencia s.
6. **Test de Complejidad Lineal** Este test comprueba el mínimo número de bits necesario para reconstruir toda la secuencia. Cada secuencia finita, s, puede ser producida por un Registro de Desplazamiento Realimentado Linealmente o LFSR (Linear Feedback Shift Register). La complejidad lineal de una secuencia es la longitud del LFSR más corto que producirá dicha secuencia. Por tanto, determina qué porción de subsecuencia se necesita para reproducir la totalidad de la secuencia. Si el valor de la complejidad lineal es L, se pueden usar $2L$ términos consecutivos para reconstruir toda la secuencia usando el algoritmo de Berlekamp-Massey (véase [5], [19], [23]). Por lo tanto, para evitar la reconstrucción de la secuencia, el valor de la complejidad lineal debe ser grande.

Las descripciones anteriores sólo nos dan una idea básica de los tests empleados. Además, hay que señalar, que pasar los tests citados aquí, no implica que el generador sea bueno; pero el no pasarlos, sí que significa que el generador deba rechazarse.

II-C. Funciones Bent

En *criptografía* se hace un uso intensivo de las funciones booleanas en la construcción y análisis de criptosistemas. Las funciones booleanas tiene un papel importante en los cifradores en bloque, cifradores en flujo y las funciones hash [6], [7], [8], [21]. Por ejemplo, en los modelos más comunes de cifradores en flujo, la clave se obtiene utilizando una función booleana.

Las funciones **bent** son un tipo especial de funciones booleanas que poseen la máxima no linealidad posible. Debido a sus buenas propiedades criptográficas, tales como la más alta no linealidad y la más baja autocorrelación [20], [28], las funciones bent pueden ser empleadas en cifradores simétricos para resistir ataques tales como el criptoanálisis lineal y el diferencial [2], [15], [24].

A continuación daremos algunas definiciones y resultados necesarios sobre funciones bent que utilizaremos en el diseño de nuestro generador.

En este artículo denotamos por $\mathbb{F}_2$ el cuerpo de Galois de dos elementos y consideramos, para cualquier entero positivo n, el espacio vectorial $\mathbb{F}_2^n$. Decimos que $\boldsymbol{u} = (u_1, u_2, \ldots, u_{n-1}, u_n) \in \mathbb{F}_2^n$ es la **expansión binaria** de n dígitos del entero positivo

$$u = u_1 2^{n-1} + u_2 2^{n-2} + \cdots + u_{n-1} 2^1 + u_n 2^0$$

y, por tanto, $\mathbb{F}_2^n = \{\boldsymbol{u} \mid 0 \leq u \leq 2^n - 1\}$.

Una **función booleana** de n variables es una aplicación $f : \mathbb{F}_2^n \longrightarrow \mathbb{F}_2$. El conjunto de todas las funciones booleanas de n variables se denota por $\mathcal{B}_n$ y $|\mathcal{B}_n| = 2^{2^n}$. La **función complementaria** de $f \in \mathcal{B}_n$ es la función booleana $1 + f$ dada por $(1 + f)(\boldsymbol{x}) = 1 + f(\boldsymbol{x})$ para todo $\boldsymbol{x} \in \mathbb{F}_2^n$.

Si $f \in \mathcal{B}_n$, llamamos **tabla de verdad** de f (véase, por ejemplo, [25], [26]) la secuencia binaria de longitud 2^n dada por

$$\boldsymbol{f} = (f(\mathbf{0}), f(\mathbf{1}), \ldots, f(\mathbf{2^n - 1})).$$

Llamamos **peso** de f, denotado por $w(f)$, al número de 1 de la tabla de verdad de f. No es difícil comprobar que

$$w(1 + f) = 2^n - w(f).$$

Decimos que f es **equilibrada** si $w(f) = 2^{n-1}$. Claramente f es equilibrada si y sólo si $1 + f$ es equilibrada.

Decimos que $f \in \mathcal{B}_n$ es una **función afín** si

$$f(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x} \rangle + b$$

donde $\boldsymbol{a} \in \mathbb{F}_2^n$, $b \in \mathbb{F}_2$ y $\langle \boldsymbol{a}, \boldsymbol{x} \rangle$ es el producto interno de los vectores $\boldsymbol{a}$ y $\boldsymbol{x}$. Si $b = 0$, decimos que f es una **función lineal**. Las funciones afines son equilibradas, pero no todas las funciones equilibradas son afines.

Definimos la **nolinealidad** de una función $f \in \mathcal{B}_n$ como

$$\mathrm{NL}(f) = \min\{d(f, \varphi) \mid \varphi \in \mathcal{A}_n\}$$

donde $\mathcal{A}_n$ es el conjunto de todas las funciones afines y la distancia entre $f, g \in \mathcal{B}_n$ se define como $d(f, g) = w(f + g)$. La nolinealidad de f está acotada superiormente por (véase [21], [31])

$$\mathrm{NL}(f) \leq 2^{n-1} - 2^{\frac{n}{2}-1}.$$

Llamamos **funciones bent** a aquellas funciones booleanas que alcanzan la máxima nolinealidad (véase [21], [31]). Consecuentemente, las funciones bent sólo existen para n par.

El teorema siguiente (véase, por ejemplo [30], [31]) nos proporciona una caracterización de las funciones bent.

Teorema 1: Supongamos que $f \in \mathcal{B}_n$. Las siguientes propiedades son equivalentes:

1. f es una función bent.
2. Si $\boldsymbol{a} \in \mathbb{F}_2^n \setminus \{\mathbf{0}\}$, la función $f_{\boldsymbol{a}} \in \mathcal{B}_n$ dada por $f_{\boldsymbol{a}}(\boldsymbol{x}) = f(\boldsymbol{x}) + f(\boldsymbol{a} + \boldsymbol{x})$, para todo $\boldsymbol{x} \in \mathbb{F}_2^n$, es equilibrada.
3. $1 + f$ es una función bent.

El número β_n de funciones bent de n variables es, en general, desconocido y sólo se conoce para $n \leq 8$, así $\beta_2 = 8$, $\beta_4 = 896$, $\beta_6 = 5\,425\,430\,528$ y $\beta_8 = 99\,270\,589\,265\,934\,370\,305\,785\,861\,242\,880$; para $n \geq 10$ solamente se conocen cotas inferiores de β_n (véase [12], [17], [27], [34], [35]). En [9], [10], [11] se pueden encontrar distintas construcciones de funciones bent, así como el número de funciones bent que dichas construcciones proporcionan; es decir, una cota inferior de β_n.

III. Diseño del generador

En esta sección presentamos el diseño del generador de bits pseudoleatorios construido a partir de un autómata celular cuya función de transición, que varía en cada instante, es una función booleana equilibrada obtenida a partir de una función bent, o bien es una función bent.

El generador de bits pseudoaleatorios que presentamos se basa en los siguientes pasos:

- Elegimos $S = \mathbb{F}_2$.
- Fijamos un entero positivo m y elegimos aleatoriamente una configuración inicial
$$C^{(0)} = \left(a_0^{(0)}, a_1^{(0)}, \ldots, a_{m-1}^{(0)}\right) \in \mathbb{F}_2^m.$$
- Elegimos un radio de vecindad r.
- Tomamos $n = 2r$ y clasificamos las β_n funciones bent[1] de n variables en dos conjuntos $\mathcal{F}_n$ y $\mathcal{G}_n$ de forma que $|\mathcal{F}_n| = |\mathcal{G}_n| = \beta_n/2$ y $f \in \mathcal{F}_n$ si y sólo si $1 + f \in \mathcal{G}_n$.
- Para cada instante $t = 0, 1, 2, \ldots$, consideramos la función de transición local $f_t : \mathbb{F}_2^{2r+1} \longrightarrow \mathbb{F}_2$ definida de la siguiente forma:
  - $\boldsymbol{a}_t \in \mathbb{F}_2^n$ es la representación binaria del entero $a_t = t \bmod 2^n$,
  - c_t es el entero cuya representación binaria de m dígitos es $C^{(t)}$ y $j_t = c_t \bmod (\beta_n/2)$,
  - g_t es la j_t-ésima función de $\mathcal{F}_n$,
  - $f_t = \widehat{f}_t \circ \pi_{2r}$ con $\pi_{2r} : \mathbb{F}_2^{2r+1} \longrightarrow \mathbb{F}_2^{2r}$ la aplicación que a cada vector de $2r+1$ componentes le asigna el vector que se obtiene al suprimir la componente central y $\widehat{f}_t : \mathbb{F}_2^{2r} \longrightarrow \mathbb{F}_2$ la aplicación dada por
$$\widehat{f}_t(\boldsymbol{x}) = \begin{cases} g_t(\boldsymbol{x}) + g_t(\boldsymbol{a} + \boldsymbol{x}), & \text{si } \boldsymbol{a}_t \neq \boldsymbol{0}, \\ 1 + g_t(\boldsymbol{x}), & \text{si } \boldsymbol{a}_t = \boldsymbol{0}. \end{cases}$$
  Notemos que si $\boldsymbol{a}_t \neq \boldsymbol{0}$, por el teorema 1, $\widehat{f}_t(\boldsymbol{x})$ es una función equilibrada, mientras que si $\boldsymbol{a}_t = \boldsymbol{0}$, entonces $\widehat{f}_t(\boldsymbol{x})$ es la función bent complementaria de $g_t(\boldsymbol{x})$.

Hemos considerado solamente $\beta_n/2$ funciones bent ya que si $\boldsymbol{a}_t \neq \boldsymbol{0}$, entonces $g_t(\boldsymbol{x})$ y $1 + g_t(\boldsymbol{x})$ proporcionan la misma función equilibrada $\widehat{f}_t(\boldsymbol{x})$.

IV. Resultados

En esta sección, presentamos los resultados de los tests descritos en la subsección II-B aplicados al generador descrito anteriormente.

Todos los puntos semilla considerados en nuestro modelo han sido creados de forma aleatoria. Como requisito de seguridad mínimo, se ha de tener en cuenta que la longitud m de cualquier punto semilla escogido en un generador de bits pseudoaleatorios, debe ser lo suficientemente grande como para hacer que una búsqueda por fuerza bruta, sobre todas las semillas posibles, no sea factible para ningún atacante. Hemos considerado semillas de longitud $m = 256$ y las secuencias pseudoaleatorias obtenidas han sido de longitud $p = 1000$. El radio de vecindad ha sido de $r = 2$ y, en consecuencia, hemos consideramos funciones bent de 4 variables cuyo número es $\beta_4 = 896$.

Así, el diagrama de evolución de nuestro autómata celular es

$C^{(0)}$:	$a_0^{(0)}$	$a_1^{(0)}$	$a_2^{(0)}$	$\cdots$	$a_{m-1}^{(0)}$
$C^{(1)}$:	$a_0^{(1)}$	$a_1^{(1)}$	$a_2^{(1)}$	$\cdots$	$a_{m-1}^{(1)}$
$C^{(2)}$:	$a_0^{(2)}$	$a_1^{(2)}$	$a_2^{(2)}$	$\cdots$	$a_{m-1}^{(2)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$C^{(p)}$:	$a_0^{(p)}$	$a_1^{(p)}$	$a_2^{(p)}$	$\cdots$	$a_{m-1}^{(p)}$

lo cual nos ha proporcionado, para cada semilla de longitud m, las secuencias de bits de longitud p

$$\left(a_i^{(0)}, a_i^{(1)}, a_i^{(2)}, \ldots, a_i^{(p)}\right), \quad \text{para } i = 0, 1, \ldots, m-1.$$

Para analizar el generador propuesto, hemos generado 10 000 semillas distintas. A las secuencias obtenidas les hemos pasado los tests estadísticos introducidos en la subsección II-B. Solamente el 5 % de dichas secuencias no superaron el test de Póker, aunque los valores obtenidos fueron muy cercanos a los intervalos de confianza. El resto de secuencias superaron todos los tests estadísticos.

En la tabla I se muestran los tests estadísticos que hemos ejecutado sobre tres muestras obtenidas por nuestro generador de secuencias, y los intervalos de confianza en el que nuestros valores se pueden mover para considerarse correctos. Si el intervalo es superado por arriba, se considera que la secuencia estudiada falla en dicho test (*FAILED*); si el valor está dentro del intervalo de confianza o es menor, se considera que lo pasa (*OK*). Los tests utilizados en nuestro estudio fueron programados siguiendo las directrices de [23, capítulo 5] y los programas están disponible en [3].

V. Conclusiones

En este artículo hemos presentado un generador de secuencias binarias pseudoaleatorias basado en un autómata celular y con función de transición local construida a través de funciones bent de n variables. Los resultados obtenidos en una primera fase, a través de los principales tests estadísticos, nos indica que el generador construido, aparentemente, tiene buenas propiedades de aleatoriedad. Ahora falta hacer una segunda revisión de dicho generador, con tests más potentes como pueden ser los tests del NIST [4], de Diehard [1] y de Tuftest [18]. Además, en un trabajo futuro incluiremos el estudio de otras modificaciones de nuestro generador.

Agradecimientos

Este trabajo ha sido parcialmente financiado por la Generalitat Valenciana mediante el proyecto AICO/2017/128 y la Universitat d'Alacant mediante el proyecto VIGROB-287.

[1] Si $n \geq 10$, entonces β_n denota una cota inferior del número de funciones bent de n variables.

Tests (Intervalo de confianza)	Resultados		
FREQUENCY (2.7060, 10.8300)	OK (0.6760, 0=513 1=487)	OK (1.2960, 0=518 1=482)	OK (0.4, 0=510 1=490)
SERIAL (4.6050, 13.8200)	OK (0.9046)	OK (3.5118)	OK (0.7642)
POKER 8 (284.3000, 330.5000)	OK (257.9760)	OK (286.6480)	FAILED (352.1840)
POKER 16 (65999.3906, 66659.4688)	OK (65474)	OK (65474)	OK (65474)
RUNS (13.3616, 26.1245)	OK (8.0920, $k = 5$)	OK (4.8557, $k = 5$)	OK (6.7874, $k = 5$)
AC (1.2820, 3.0900)	OK (0.8112)	OK (0.7848)	OK (0.8466)
CL (500)	OK (500)	OK (501)	OK (501)

Tabla I
RESULTADOS TESTS ESTADÍSTICOS

REFERENCIAS

[1] "The Marsaglia random number CDROM including the Diehard battery of tests of randomness," https://web.archive.org/web/20160125103112/http://stat.fsu.edu/pub/diehard/, 1995.

[2] C. M. Adams, "Constructing symmetric ciphers using the CAST design procedure," *Designs, Codes and Cryptography*, vol. 12, pp. 283–316, 1997.

[3] R. Álvarez, "RANDTEST: A simple suit of randomness tests for bit sequences". https://github.com/rias/randtest, 2004.

[4] L. E. Bassham, A. L. Rukhin, J. Soto, J. R. Nechvatal, M. E. Smid, S. D. Leigh, M. Levenson, M. Vangel, N. A. Heckert, and D. L. Banks, "A statistical test suite for random and pseudorandom number generators for cryptographic applications," NIST, NIST Pubs 800-22 Rev 1a, 2010.

[5] E. Berlekamp, *Algebraic coding Theory*, revised ed. Singapore: World Scientific, 2015.

[6] A. Braeken, V. Nikov, S. Nikova, and B. Preneel, "On Boolean functions with generalized cryptographic properties," in *Progress in Cryptology – INDOCRYPT 2004*, ser. Lecture Notes in Computer Science, A. Canteaut and K. Viswanathan, Eds. Berlin: Springer-Verlag, 2004, vol. 3348, pp. 120–135.

[7] C. Carlet and Y. Tarannikov, "Covering sequences of Boolean functions and their cryptographic significance," *Designs, Codes and Cryptography*, vol. 25, pp. 263–279, 2002.

[8] C. Charnes, M. Rötteler, and T. Beth, "Homogeneous bent functions, invariants, and designs," *Designs, Codes and Cryptography*, vol. 26, pp. 139–154, 2002.

[9] J.-J. Climent, F. J. García, and V. Requena, "On the construction of bent functions of $n + 2$ variables from bent functions of n variables," *Advances in Mathematics of Communications*, vol. 2, no. 4, pp. 421–431, 2008.

[10] ——, "Construction of bent functions of $2k$ variables from a basis of $\mathbb{F}_2^{2k}$," *International Journal of Computer Mathematics*, vol. 89, no. 7, pp. 863–880, 2012.

[11] ——, "On the construction of new bent functions from the max-weight and min-weight functions of old bent functions," *SeMA Journal*, vol. 72, no. 1, pp. 13–36, 2015.

[12] T. W. Cusick and P. Stănică, *Cryptographic Boolean Functions and Applications*, 2nd ed. San Diego, CA: Academic Press, 2017.

[13] A. Fúster-Sabater, M. E. Pazo-Robles, and P. Caballero-Gil, "Modelización del generador auto-shrinking mediante autómatas celulares," in *Actas X Reunión Española sobre Criptología y Seguridad de la Información (RECSI 2008)*, L. Hernández Encinas and A. Martín del Rey, Eds., Salamanca, España, 2008, pp. 187–193.

[14] J. Gleick, *Chaos: Making a New Science*. New York, NY: Penguin Books, 1988.

[15] K. C. Gupta and P. Sarkar, "Improved construction of nonlinear resilient S-boxes," *IEEE Transactions on Information Theory*, vol. 51, no. 1, pp. 339–348, 2005.

[16] L. Hernández Encinas, A. Martín del Rey, and G. Rodríguez Sánchez, "Aplicaciones de los autómatas celulares a la generación de bits," *Boletín de la Sociedad Española de Matemática Aplicada*, vol. 21, pp. 65–87, 2002.

[17] P. Langevin and G. Leander, "Counting all bent functions in dimension eight 99270589265934370305785861242880," *Designs, Codes and Cryptography*, vol. 59, pp. 193–201, 2011.

[18] G. Marsaglia and W. W. Tsang, "Some difficult-to-pass tests of randomness," *Journal of Statistical Software*, vol. 7, no. 3, pp. 1–9, 2002.

[19] J. L. Massey, "Shift-register synthesis and BCH decoding," *IEEE Transactions on Information Theory*, vol. 15, no. 1, pp. 122–127, 1969.

[20] W. Meier and O. Staffelbach, "Fast correlation attacks on certain stream ciphers," *Journal of Cryptology*, vol. 1, no. 3, pp. 159–176, 1989.

[21] ——, "Nonlinearity criteria for cryptographic functions," in *Advances in Cryptology – EUROCRYPT'89*, ser. Lecture Notes in Computer Science, J. J. Quisquater and J. Vandewalle, Eds. Berlin: Springer-Verlag, 1990, vol. 434, pp. 549–562.

[22] ——, "Analysis of pseudo random sequences generated by cellular automata," in *Advances in Cryptology – EUROCRYPT'91*, ser. Lecture Notes in Computer Science, D. W. Davies, Ed. Berlin: Springer-Verlag, 1991, vol. 547, pp. 186–199.

[23] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography*. Boca Raton, FL: CRC Press, 1996.

[24] K. Nyberg, "Perfect nonlinear S-boxes," in *Advances in Cryptology – EUROCRYPT'91*, ser. Lecture Notes in Computer Science, D. W. Davies, Ed. Berlin: Springer-Verlag, 1991, vol. 547, pp. 378–386.

[25] D. Olejár and M. Stanek, "On cryptographic properties of random Boolean functions," *Journal of Universal Computer Science*, vol. 4, no. 8, pp. 705–717, 1998.

[26] E. Pasalic and T. Johansson, "Further results on the relation between nonlinearity and resiliency for Boolean functions," in *Crytography and Coding*, ser. Lecture Notes in Computer Science, M. Walker, Ed. Berlin: Springer-Verlag, 1999, vol. 1746, pp. 35–44.

[27] B. Preneel, "Analysis and design of cryptographic hash functions," Ph.D. dissertation, Katholieke University Leuven, Jan. 1993.

[28] B. Preneel, W. Van Leekwijck, L. Van Linden, R. Govaerts, and J. Vandewalle, "Propagation characteristics of Boolean functions," in *Advances in Cryptology – EUROCRYPT'90*, ser. Lecture Notes in Computer Science, I. B. Damgard, Ed. Berlin: Springer-Verlag, 1991, vol. 473, pp. 161–173.

[29] P. Sarkar, "A brief history of cellular automata," *ACM Computing Surveys*, vol. 32, no. 1, pp. 80–107, 2000.

[30] J. Seberry and X.-M. Zhang, "Constructions of bent functions from two known bent functions," *Australasian Journal of Combinatorics*, vol. 9, pp. 21–35, 1994.

[31] J. Seberry, X.-M. Zhang, and Y. Zheng, "Nonlinearity and propagation characteristics of balanced Boolean functions," *Information and Computation*, vol. 119, pp. 1–13, 1995.

[32] M. Sipper and M. Tomassini, "Computation in artificially evolved, non-uniform cellular automata," *Theoretical Computer Science*, vol. 217, no. 1, pp. 81–98, 1999.

[33] T. Toffoli and N. Margolus, *Cellular Automata Machines: A New Environment for Modeling*. Cambridge, MA: MIT Press, 1987.

[34] N. Tokareva, "On the number of bent functions from iterative constructions: lower bounds and hypotheses," *Advances in Mathematics of Communications*, vol. 5, no. 4, pp. 609–621, 2011.

[35] ——, *Bent Functions. Results and Applications to Cryptography*. Amsterdam: Academic Press, 2015.

[36] M. Tomassini and M. Perrenoud, "Cryptography with cellular automata," *Applied Soft Computing*, vol. 1, pp. 151–160, 2001.
[37] S. M. Ulam, "On some mathematical problems connected with patterns of growth of figures," in *Essays on Cellular Automata*, A. W. Burks, Ed. Urbana, IL: University of Illinois Press, 1970, pp. 219–231.
[38] J. von Neumann, *Theory of Self-Reproducing Automata*. Champaign, IL: University of Illinois Press, 1966.
[39] S. Wolfram, "Cellular automata," *Los Alamos Science*, vol. 9, pp. 2–21, 1983.
[40] ——, "Universality and complexity in cellular automata," *Physica D*, vol. 10, pp. 1–35, 1984.
[41] ——, "Cryptography with cellular automata," in *Advances in Cryptology – CRYPTO '85*, ser. Lecture Notes in Computer Science, H. C. Williams, Ed. Berlin: Springer-Verlag, 1986, vol. 218, pp. 429–432.
[42] ——, "Random sequence generation by cellular automata," *Advances in Applied Mathematics*, vol. 7, pp. 123–169, 1986.
[43] ——, *A New Kind of Science*. Champaign, IL: Wolfram Media, 2002.

Fingerprinting in weighted noisy adder channels

Marcel Fernandez
Department of Network Engineering.
Universitat Politècnica de Catalunya
Barcelona, Spain

Elena Egorova
Skolkovo Institute of Science
and Technology (Skoltech).
Moscow region, Russia

Grigory Kabatianski
Skolkovo Institute of Science
and Technology (Skoltech).
Moscow region, Russia

***Abstract*—We propose a new coding scheme for multimedia fingerprinting resistant to noise. Our scheme is based on signature codes for the weighted noisy adder channel. The scheme (codes) can trace the entire coalition of pirates and provides significantly better rate than previously known fingerprinting schemes. We also establish a relationship between these two problems and the compressed sensing problem.**

I. Introduction

There are the following well established concepts of digital fingerprinting codes known as *tracing traitors* [1] or *codes with the identifiable parent property* [2], and *digital fingerprinting codes* [3],[4], dealing with the problem of protection against unauthorized copying of discrete data.

The fingerprinting technique consists in making the copies of a digital object unique by embedding a different set of marks in each copy. Having unique copies of an object clearly rules out plain redistribution, but still a coalition of dishonest users can collude, compare their copies and by changing the marks where their copies differ, they are able to create a pirate copy that tries to disguise their identities. Thus, the fingerprinting problem consists in finding, for each copy of the object, the right set of marks that help to prevent collusion attacks. If the marks are the codewords of a collusion secure code, then it is guaranteed that at least one of the members of the traitor coalition can be traced back with high probability.

A theoretical model for data protection of continuous data was proposed in [5] and [6]. The corresponding codes, called *multimedia fingerprinting codes*, were investigated in [5] - [12].

This paper can be viewed as a sequel of our work in [13] and [14]. The story runs as follows. First multimedia fingerprinting codes with non-vanishing rate, namely with rate of order t^{-2}, where t is the coalition's size, were constructed in our paper [13]. Generalizations of results from [13] to the case of presence of noise was done in [14]. All previous considerations of multimedia fingerprinting codes were based on some discretisation, introduced in [5], i.e., only *hard decoding* of such codes was investigated. In this paper we consider the output of the corresponding channel without discretisation, i.e. we consider *soft decoding* of multimedia fingerprinting codes. We show that the corresponding model of multimedia fingerprinting channel without discretisation is equivalent to some modification of some model of the adder channel, considered in [15], which we call *weighted adder channel*. Moreover, signature codes for the weighted real adder channel can be considered as multimedia fingerprinting codes with soft decoding. We show that the rate of multimedia fingerprinting codes with soft decoding is at least $1/t$. Therefore, it is much larger than for the hard decoding case where the rate cannot exceed $O(t^{-2}\log t)$ even for the noiseless case, see [13].

II. Channel model for multimedia fingerprinting

Consider, as discussed in [5], [6], multimedia content being represented as a real-valued vector $\mathbf{x} \in R^m$, called the *host signal*. To prevent unauthorized redistribution of $\mathbf{x}$, the dealer constructs a set of digital fingerprints using a linear modulation scheme that employs n *noise-like* orthonormal signals (vectors)

$$\{\mathbf{f}_i \in R^m \mid i = 1, \ldots, n,\ n \leq m\}.$$

The fingerprint $\mathbf{w}_j$ of the $j-$th user ($j \in \{1, \ldots, M\}$) is represented as follows:

$$\mathbf{w}_j = \sum_{i=1}^{n} h_{ij}\mathbf{f}_i, \tag{1}$$

where:

- $h_{ij} \in \{+1, -1\}$ if antipodal signals are used.
- $h_{ij} \in \{0, 1\}$ if on-off keying type of modulation is used.

In this paper we consider the case on-off keying type of modulation. The sets of fingerprints $\{\mathbf{w}_j\}$ will be called a *fingerprinting code*.

Equivalently and more convenient is to consider the set of n-dimensional binary vectors

$$\mathbf{h}_j = (h_{1j}, \ldots, h_{nj})$$

as *binary multimedia fingerprinting code* of length n, cardinality M and rate $R = n^{-1}\log_2 M$.

The dealer distributes to the j-th user the vector

$$\mathbf{y}_j = \mathbf{x} + \mathbf{w}_j,$$

where one assumes that $\|\mathbf{x}\|_2 \gg \|\mathbf{w}_j\|_2$ in order to be sure that the fingerprinting scheme does not introduce significant changes in the host signal.

A group of users, called *colluders* or *pirates*, aims to create an unauthorized copy of the contents such that the dealer cannot trace its origins to any of them. An important concept in digital fingerprinting, the *Marking Assumption* of [3], in the context of multimedia fingerprinting can be expressed in the following way: we *assume* that the members of the pirate coalition $J \subset \{1, \ldots, M\}$ cannot manipulate individual signals $\mathbf{f}_j$, and are limited to *linear attacks*. By a linear attack we mean that pirates can generate a forged copy $\hat{\mathbf{y}}$ of the host

content only as a linear combination of their copies $\mathbf{y}_j$ with some coefficients λ_j

$$\hat{\mathbf{y}} = \sum_{j\in J} \lambda_j \mathbf{y}_j \tag{2}$$

where $\lambda_j > 0$ for all j and $\sum_{j\in J} \lambda_j = 1$. Hence

$$\hat{\mathbf{y}} = \mathbf{x} + \sum_{j\in J} \lambda_j \mathbf{w}_j. \tag{3}$$

The dealer calculates the vector $\mathbf{S} = \mathbf{S}(J, \{\lambda_j\}) = (s_1, \ldots, s_n)$, where

$$s_k = (\widehat{\mathbf{y}} - \mathbf{x}, \mathbf{f}_k) = \langle \sum_{i=1}^{n} \sum_{j\in J} \lambda_j h_{ij} \mathbf{f}_i, \mathbf{f}_k \rangle = \sum_{j\in J} \lambda_j h_{kj} \tag{4}$$

and $\langle \cdot, \cdot \rangle$ denotes the inner product. Equivalently,

$$\mathbf{S} = \sum_{j\in J} \lambda_j \mathbf{h}_j, \tag{5}$$

where $\mathbf{h}_j = (h_{1j}, \ldots, h_{nj})$. It was suggested in [5], [6] to consider the following discretisation, when the dealer knows only whether $s_k = 0$, or $s_k = 1$, or that $0 < s_k < 1$, but does not know the value of s_k in the last case. Note, that $s_k = 0$ means that $h_{kj} = 0$ for all $j \in J$, $s_k = 1$ means that $h_{kj} = 1$ for all $j \in J$ and finally $0 < s_k < 1$ means that

$$\{\cup_{j\in J} h_{kj}\} = \{0, 1\} \tag{6}$$

This means that the dealer observesa ternary output, i.e., the output of the 2-frequency noiseless multiple-access channel without intensity information [16] (the A-channel, for short), see a more detailed explanation in the next section. This approach to constructing multimedia fingerprinting codes as signature codes for the A-channel was first developed in [13].

In almost all previous works on multimedia fingerprinting codes it was assumed that the dealer observes the forged copy $\hat{\mathbf{y}}$ *without noise*. This simplified assumption is not very realistic. Indeed, there are at least two sources of noise: a measurement noise, and a malicious noise, produced by the coalition. Therefore we assume that the dealer observes the following vector as a result of coalition forgery

$$\mathbf{z} = \mathbf{x} + \sum_{j\in J} \lambda_j \mathbf{w}_j + \mathbf{E}, \tag{7}$$

where $\mathbf{E} = (E_1, \ldots, E_m)$ and let for simplicity E_i be independent identically distributed random variables. Similarly to (4) and (5) the dealer evaluates

$$s_k(\mathbf{E}) = (\widehat{\mathbf{z}} - \mathbf{x}, \mathbf{f}_k)$$

and the following vector

$$\mathbf{S}(\mathbf{E}) = (s_1(\mathbf{E}), \ldots, s_n(\mathbf{E})) = \sum_{j\in J} \lambda_j \mathbf{h}_j + \mathbf{e}, \tag{8}$$

where

$$\mathbf{e} = (e_1, \ldots, e_n)$$

and

$$e_k = \langle \mathbf{E}, \mathbf{f}_k \rangle.$$

Based on the known value of the vector $\mathbf{S}(\mathbf{E})$ the dealer tries to reveal (to trace) the whole coalition that produced this forgery, or at least one of its members. We shall consider $\mathbf{S}(\mathbf{E})$ as the output of *continuous* multimedia fingerprinting channel which is in fact a modification of well known adder channel with noise. We call this channel the *weighted adder channel* (WAC, for short).

Recall, that a code C is called a uniquely decodable (u.d.) $t - signature$ code for the multiple-access channel (MAC) if any set $U \subset C$ of cardinality at most t can be uniquely recovered from the corresponding output of the channel. This means that for $U, U' \subset C$ such that $U \neq U'$ and $|U| \leq t$, $|U'| \leq t$, the outputs of the MAC corresponding to the input sets U and U', are distinct. In the case of the binary A-channel considered above this property is exactly the property that any coalition of codewords U with $|U| \leq t$ can be uniquely recovered from its signature S_U.

We see that indeed, equation (8) with $\lambda_j \equiv 1$ and $\mathbf{e} = \mathbf{0}$ describes the output of the adder channel with binary input in the case of t active users among a total number of M, and λ_j plays the role of weights. In this case, multimedia fingerprinting codes capable to trace the whole coalition of size t are in fact the same as t-signature code for the weighted adder channel, and the number of users M equals the cardinality of the corresponding signature code. Note that an intermediate case of the weighted adder channel $\lambda_j > 0$ was firstly considered in [15] (where weights λ_j were called gains) but consideration of noise was restricted in [15] to the very particular case $\mathbf{e} = b\mathbf{1}$, where b is a real number. Below, we will also establish a relationship between the weighted adder channel and the compressed sensing problem [17],[18].

III. Codes for noiseless weighted adder channel

In the previous section we established that the dealer observes the output a multiple-access channel, which we called the weighted adder channel, see (8), with unknown weights λ_j and an unknown set J of active users of the WAC, i.e., the set of traitors. The dealer's goal is to find the set J, that corresponds to zero-error coding for multiple-access channel (MAC) with partial activity, i.e. not all possible users of the MAC are active, but only at most t of them. We shall consider the set of n-dimensional binary vectors $\mathbf{h}_j = (h_{1j}, \ldots, h_{nj})$ as a binary multimedia fingerprinting code of cardinality M capable to find any set of t or less traitors in the case of a noiseless WAC if and only if for any subsets $J, J' \subset \{1, \ldots, n\}$ such that $|J| \leq t$, $|J'| \leq t$ the following equation

$$\sum_{j\in J} \lambda_j \mathbf{h}_j = \sum_{j\in J'} \lambda_j \mathbf{h}_j \tag{9}$$

implies that $J = J'$.

Since weights λ_j are (almost) arbitrary real numbers the above condition means that any $2t$ vectors $\mathbf{h}_j$ are linearly independent. Such an effective set of vectors is well known in coding theory as columns of a parity-check matrix of a binary BCH code of length M and redundancy n, correcting t errors. Indeed, any $2t$ columns are linearly independent over the field of residues module 2, and hence they are independent over the rational numbers and also over the real numbers since

in all cases dependency means that the determinant of the corresponding $t \times t$ minor equals to 0. It is not clear if this construction is even weakly optimal, i.e., has the minimal possible order of n as a function of M - for binary BCH-codes it gives

$$n = t \log_2 M(1 + o(1)).$$

Anyway we see that in the case of absence of quantization, i.e. soft decision, the rate of the best multimedia fingerprinting code is at least $1/t$ whereas for hard decoding the rate of any multimedia fingerprinting code is $O(t^{-2} \log t)$ [13].

IV. Codes for noisy weighted adder channel and the compressed sensing problem

Now we want to construct codes which are capable to recover a set J by its "syndrome" $\mathbf{S}(\mathbf{E}) = \sum_{j \in J} \lambda_j \mathbf{h}_j + \mathbf{e}$, see equation (8).
It is very similar to the compressed sensing problem, see [17],[18], with the only three differences:

- for "weakly" multimedia fingerprinting code it is sufficient to find at least one $j \in J$;
- for multimedia fingerprinting we need to find only the "error" set J, whereas the solution of the compressed sending problem means finding J and values λ_j;
- vectors $\mathbf{h}_j$ are *binary*.

Consider an slightly different statement of the problem, more in the spirit of coding theory, namely, that the number of errors is limited, i.e., $wt(\mathbf{e}) \leq T$. A solution to this problem was given in [19], see also [20]. To omit details (which can be found in [19]), in order to correct T syndrome's errors one should "add"

$$T \log(t \log M)$$

redundant symbols. This gives a total redundancy

$$n = t \log_2 M + T \log_2 \log_2 M + T \log_2 t.$$

It is easy to check that for $T - const$ it keeps the same order of rate, namely

$$R = t^{-1}(1 + o(1)).$$

Note that a similar technique was used in [21].

V. Conclusion

Let us note in conclusion that the problem of noise in multiple-access channels was almost ignored probably because of the following remark in [16] "A more sophisticated model taking such errors into account could easily be developed. One method would be to use a noisy channel in cascade with our noiseless channel. It is our contention, however, that although the details are different, the basic ideas are the same in the noisy and noiseless cases." Our construction is much better than such simple cascade.

Acknowledgments

The work of M. Fernández has been supported by the Spanish Government through projects TEC2015-68734-R (MINECO/FEDER) "ANFORA" and 2017 SGR 782.

References

[1] B.Chor, A.Fiat, and M.Naor. "Tracing traitors". *Advances in Cryptology-Crypto'94, LNCS*, 839, pp. 480–491, 1994.
[2] H. D. L. Hollmann, J. H. van Lint, J.-P. Linnartz, and L. M. G. M. Tolhuizen, "On codes with the Identifiable Parent Property," *J. Combinatorial Theory, Ser. A*, vol. 82, no. 2, pp. 121–133, May 1998.
[3] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1897–1905, 1998.
[4] G. Tardos, "Optimal probabilistic fingerprint codes." in *Proceedings of the 35th Annual ACM Symposium on Theory of Computing, June 9-11, 2003, San Diego, CA, USA*, ACM, 2003, pp. 116–125.
[5] W. Trappe , M. Wu ,Z.J. Wang ,K.J.R. Liu , "Anti-collusion fingerprinting for multimedia", *IEEE Trans. Signal Process.* vol. 51, pp. 1069-1087, 2003.
[6] K.J.R. Liu, W. Trappe, Z.J.Wang, M. Wu and H.Zhao. Multimedia fingerprinting forensics for traitor tracing. Vol. 4. Hindawi Publishing Corporation, 2005.
[7] M.Cheng and Y.Miao, "On Anti-Collusion Codes and Detection Algrithms for Multimedia Fingerprinting". *IEEE Trans. Info. Theory*, vol. 57, no. 7, pp. 4843-4851, 2011.
[8] M.Cheng, L. Ji and Y.Miao, "Separable Codes", *IEEE Trans. Info. Theory*, vol. 58, no. 3, pp. 1791-1803, 2012.
[9] Fei Gao and Gennian Ge, "New bounds on separable codes for multimedia fingerprinting", *IEEE Trans. Info. Theory*, vol. 60, pp. 5257-5262, 2014.
[10] M.Cheng, H-L Fu, J. Jiang, Y-H Lo and Y.Miao, "New bounds on 2-separable codes of length 2", *Designs, Codes, and Cryptography*, vol. 74, no. 3, pp. 31-40, 2015.
[11] M.Cheng, L. Ji and Y.Miao, "Strongly separable codes", *Designs, Codes and Cryptography*, 2015
[12] J. Jian, Mi. Cheng, and Y. Miao, "Strongly Separable Codes", '*Designs, Codes, and Cryptography*, vol. 72, no. 2, pp. 303-318, 2016.
[13] E. Egorova, M. Fernandez, G. Kabatiansky, and Moon Ho Lee, "Signature codes for A-channel and collusion-secure multimedia fingerprinting codes", Proceedings 2016 IEEE International Symposium on Information Theory, Barcelona, 2016.
[14] E. Egorova, M. Fernandez and G. Kabatiansky, "Multimedia Fingerprinting Codes Resistant Against Colluders and Noise", Proceedings of 8th IEEE International Workshop on Information Forensic and Security, Abu Dhabi, 2016.
[15] P.Mathys, "A Class of Codes for a T Active Users Out of N Multiple-Access Communvation System", *IEEE Trans. Info. Theory*, vol. 36, No 6, pp. 1206-1219, 1990.
[16] Chang S. C., Wolf J. K., "On the T-user M-frequency noiseless multiple-access channel with and without intensity information", *IEEE Trans. Inform. Theory* vol. 27, no. 1, pp. 41-48, 1981.
[17] D. L. Donoho, "Compressed sensing", *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289-1306, 2006
[18] E. J. Candes, T. Tao, "Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies? ", *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 5406 - 5425, 2006
[19] Kabatiansky, G., Lomakov, V., Vladuts, S. "On codes correcting errors in channel and syndrom", Problems of Information Transmission, v.51, N.2, pp.50-57, 2015.
[20] Kabatiansky G., Vladuts S., Tavernier C., "On the Doubly Sparse Compressed Sensing Problem", IMAAC 2015, *LNCS*, vol. 9496, pp. 1-6, 2015.
[21] Gritsenko V., Kabatiansky G., Lebedev V. and Maevskiy A., "Signature codes for noisy multiple access adder channel", Designs, Codes and Cryptography, 82(1), pp.293-299, 2017.

Diagnóstico de CSM en estegoanálisis

Daniel Lerch-Hostalot y David Megías,
Internet Interdisciplinary Institute (IN3),
Universitat Oberta de Catalunya (UOC),
CYBERCAT-Center for Cybersecurity Research of Catalonia
Avgda. Carl Friedrich Gauss, 5, 08860, Castelldefels
E-mail: {dlerch,dmegias}@uoc.edu

***Resumen*—En este artículo se presenta una metodología para detectar el problema de *Cover Source Mismatch* (CSM) en estegoanálisis en imágenes. El método propuesto determina si un clasificador ha sido entrenado con un conjunto de datos incompleto y si no es apropiado para clasificar una imagen concreta. En este caso, la técnica desarrollada detecta que estamos intentando clasificar una muestra no alineada y elige no clasificarla. En el artículo se muestra que esta metodología permite incrementar considerablemente la precisión del clasificador a cambio de no clasificar ciertas muestras. Este método permite aplicar estegoanálisis en escenarios reales donde aparece el problema del CSM. Además, se presenta un procedimiento simple para completar el conjunto de entrenamiento proporcionando nuevas imágenes para completar las regiones del espacio no cubiertas por los datos de entrenamiento iniciales.**

***Index Terms*—Estegoanálisis, *Cover-Source Mismatch*, *Machine Learning*.**

I. Introducción

La esteganografía es una colección de técnicas usadas para ocultar información dentro de objetos aparentemente inocentes. Hoy en día, estos objetos suelen ser medios digitales debido a su amplio uso. Por otra parte, se denomina estegoanálisis a las diferentes técnicas usadas para detectar la existencia de mensajes ocultos mediante esteganografía.

La mayoría de los métodos de estegoanálisis en el estado del arte usan *Machine Learning* [20], [6], [21], [23], [22]. En los sistemas de estegoanálisis basados en *Machine Learning* se utiliza un conjunto de imágenes esteganográficas –también llamadas imágenes *stego*– y de imágenes sin alterar –conocidas como imágenes *cover*– para entrenar un clasificador. Posteriormente, ese clasificador ya entrenado se utiliza para identificar qué imágenes, de un conjunto de test, son *cover* y cuáles son *stego*. Este enfoque funciona muy bien en condiciones de laboratorio, es decir, cuando el conjunto de entrenamiento es similar al conjunto de test. Sin embargo, en el mundo real, el conjunto de imágenes usado por el esteganógrafo para ocultar información puede ser muy diferente al usado por el estegoanalista para entrenar el clasificador [11], incluso aunque el estegoanalista intente disponer de un conjunto de imágenes lo más completo posible. Esto ocurre cuando las imágenes de test no están bien representadas en el conjunto de entrenamiento. Las imágenes de test pueden haber sido obtenidas usando otros modelos de cámaras o pueden haberse tomado en condiciones muy diferentes. En estegoanálisis, este problema se conoce como *cover source mismatch* (CSM), descrito por primera vez en [4].

Se han propuesto diferentes enfoques para abordar el problema del CSM. Durante la competición BOSS [2], algunos participantes intentaron una solución conocida como "entrenamiento con bases de datos contaminadas" que consiste en eliminar el ruido de imágenes del conjunto de test e incluirlas en el conjunto de entrenamiento [7]. Otro enfoque diferente consiste en crear un conjunto de entrenamiento lo más completo posible. En [17], los autores entrenan un clasificador usando grandes cantidades de imágenes. Debido a los altos tiempos de procesamiento y a la gran cantidad de memoria requeridas, los autores optan por usar clasificación *online* y clasificadores ligeros. En [14], se presentan tres estrategias diferentes: (1) entrenamiento con una mezcla de bases de datos de imágenes; (2) uso de diferentes clasificadores para cada una de las bases de datos y test mediante la base de datos más cercana; y (3) utilizar el segundo enfoque pero añadiendo las imágenes de una en una. Otro método similar, conocido como *islet* [19], introduce una paso de preprocesamiento que consiste en organizar las imágenes en grupos, asignando un clasificador a cada grupo. En [24], se presenta un esquema para construir una base de datos completa. Finalmente, en [16] se presenta un método no supervisado que no requiere de ningún conjunto de entrenamiento. En esta técnica, el problema del CSM se evita generando un conjunto de datos de entrenamiento basado en los datos del conjunto de test, transformando el problema en supervisado.

Tal y como se propone en [12], puede ser interesante que los métodos de estegoanálisis produzcan como resultado una muestra *no clasificada* cuando la muestra que se pretende clasificar no está bien representada en el conjunto de entrenamiento. En este artículo, presentamos una metodología a partir de este principio. La idea propuesta está basada en [16], extendiéndola para detectar este tipo de muestras. Al no clasificar las muestras que no están bien representadas en el conjunto de entrenamiento, lo que produciría una cantidad significativa de errores, la precisión final de la clasificación se ve incrementada significativamente. Adicionalmente, se propone una metodología para completar el conjunto de entrenamiento teniendo en cuenta el número de muestras no clasificadas.

Este trabajo se ha financiado parcialmente con los fondos del proyecto TIN2014-57364-C2-2-R "SMARTGLACIS".

El resto del artículo está organizado de la manera siguiente. La Sección II introduce algunos conceptos relevantes que se usan en el método presentado. La Sección III presenta el método propuesto. Los resultados experimentales obtenidos usando cinco bases de datos de imágenes diferentes en los que aparece el problema del CSM se presentan en la Sección IV. Finalmente, en la Sección V se presentan las conclusiones y algunas sugerencias para realizar investigación futura.

II. Preliminares

Consideremos un escenario en el que el algoritmo y la ratio de inserción –pero no la clave– son conocidos (al menos aproximadamente) por el estegoanalista. Dado un conjunto de imágenes *cover* podemos construir un conjunto de entrenamiento usando una mitad de imágenes sin alterar y ocultando un mensaje en la otra mitad. Para ello, se usará el algoritmo y la ratio de inserción conocidos y un valor aleatorio de la clave secreta. A continuación, si usamos una técnica de *Machine Learning* basada en la extracción de características (no todas las técnicas de *Machine Learning* requieren extracción de características [24]), necesitaremos extraerlas. Llamaremos a este conjunto de entrenamiento A^{train}. En el estegoanálisis basado en *Machine Learning* la metodología habitual consiste en entrenar un clasificador usando A^{train}. A continuación, dicho clasificador se utiliza para clasificar A^{test}, es decir, un conjunto de imágenes para el cual, *a priori*, no tenemos información acerca de qué imágenes son *cover* y cuáles son *stego*.

La metodología que proponemos usa algunos conjuntos de datos adicionales. Primero, definimos los conjuntos B^{train} y C^{train} tal y como se describe en [16]: el conjunto B^{train} es el resultado de ocultar información aleatoria en todas las imágenes de A^{train} usando el mismo algoritmo, una ratio de inserción similar y una clave aleatoria; y el conjunto C^{train} es el resultado de ocultar información aleatoria en todas las imágenes de B^{train}, siguiendo el mismo procedimiento. Por lo tanto dispondremos de tres conjuntos: el conjunto A^{train} que contiene imágenes *cover* y *stego*, el conjunto B^{train} que contendrá imágenes *stego* y "doble *stego*", y el conjunto C^{train} que contendrá imágenes"doble *stego*" y "triple *stego*".

Introducimos la siguiente notación: α_i es una muestra del conjunto A^{train}, β_i es la muestra correspondiente del conjunto B^{train}, donde $\beta_i = \mathrm{Embed}(\alpha_i, Bitrate)$, y γ_i es la muestra correspondiente en C^{train}: $\gamma_i = \mathrm{Embed}(\beta_i, Bitrate)$; donde "Embed" significa que incrustamos un mensaje usando información aleatoria, una clave aleatoria, una ratio de inserción ($Bitrate$) similar y el mismo algoritmo.

Siguiendo el mismo procedimiento, usando las imágenes de A^{test} preparamos los conjuntos B^{test} y C^{test}. En este caso a_i, b_i y c_i hacen referencia a las muestras de los conjuntos A^{test}, B^{test} y C^{test}, respectivamente, donde $b_i = \mathrm{Embed}(a_i, Bitrate)$ y $c_i = \mathrm{Embed}(b_i, Bitrate)$.

La única diferencia con los conjuntos de entrenamiento es que en los conjuntos de test no se conocen las clases (etiquetas), es decir, no sabemos si las imágenes son *cover* o *stego* en A^{test}, si son *stego* o "doble *stego*" en B^{test} o si son "doble *stego*" o "triple *stego*" en C^{test}.

Ahora, supongamos que disponemos de un clasificador C_A (con su correspondiente extractor de características, si es necesario) capaz de clasificar imágenes en *cover* ($\mathcal{C}_A$) y *stego* ($\mathcal{S}_A$) con una tasa de error aceptable. De la misma manera, supongamos que disponemos de los clasificadores C_B y C_C, capaces de clasificar imágenes en *stego* ($\mathcal{S}_B$) y "doble *stego*" ($\mathcal{D}_B$), y en "doble *stego*" ($\mathcal{D}_C$) y "triple *stego*" ($\mathcal{T}_C$) respectivamente.

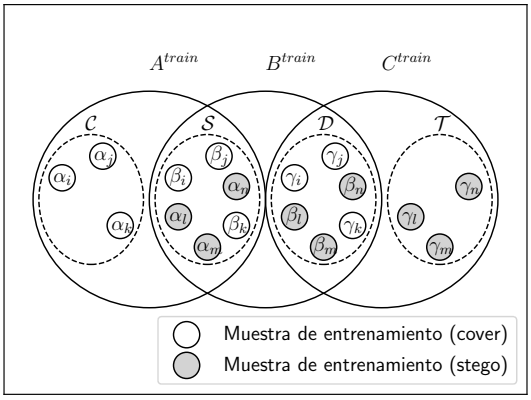


Figura 1. Conjuntos A^{train}, B^{train} y C^{train}

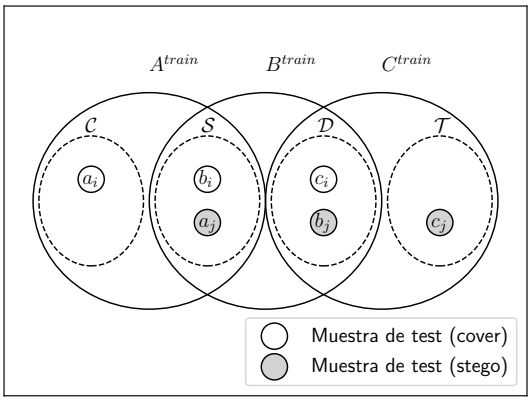


Figura 2. Muestras alineadas

III. Método propuesto

Si disponemos de un conjunto de test que queremos clasificar y de un clasificador entrenado con un conjunto de entrenamiento sin CSM, podremos clasificar el conjunto de test con una probabilidad de error determinada [6]. Si extendemos esta idea a los conjuntos de datos introducidos en la sección anterior, podemos obtener una herramienta para detectar muestras no alineadas entre el conjunto de entrenamiento y el conjunto de test. Dado que existe una biyección entre los elementos de A^{test}, B^{test} y C^{test}, si una imagen de A^{test} se clasifica como *cover* por C_A, esperaremos que la imagen correspondiente de B^{test} se clasifique como *stego* (y no "doble

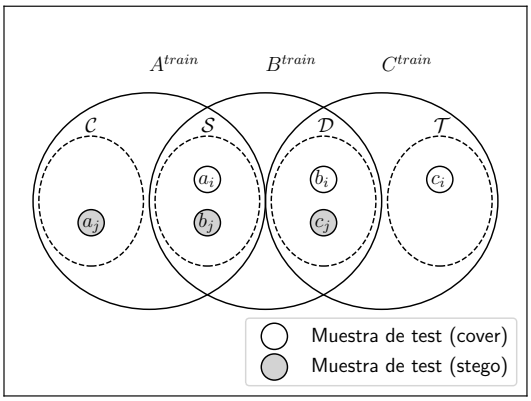


Figura 3. Muestras no alineadas indetectables

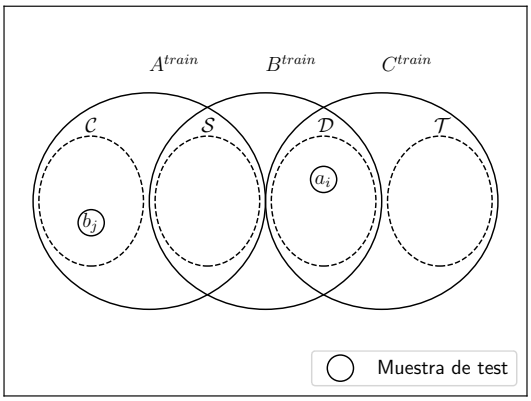


Figura 5. Muestras detectables por F_2

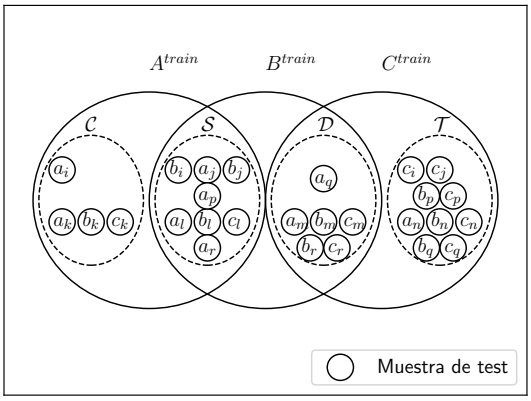


Figura 4. Muestras no alineadas detectables por F_1

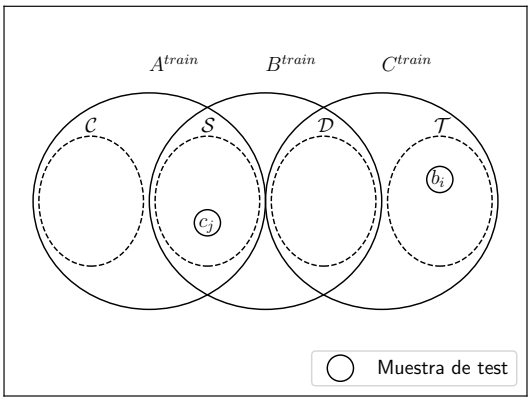


Figura 6. Muestras detectables por F_3

stego") por C_B. Adicionalmente, esperamos que la imagen correspondiente en C^{test} se clasigique como "doble *stego*" y no como "triple *stego*". De forma similar, si una imagen en A^{test} se clasifica como como *stego* por C_A, esperaremos que la imagen correspondiente en B^{test} se clasifique como "doble *stego*" por C_B, y que la imagen correspondiente en C^{test} se clasifique como "triple *stego*" por C_C.

Llamaremos muestra no alineada a la imagen que no cumpla las restricciones esperadas. Si una imagen no es clasificada de forma consistente usando los tres clasificadores, será mejor no clasificarla, pues nuestros clasificadores podrían no ser apropiados para trabajar con dicha imagen. Por lo tanto, el método propuesto puede retornar tres posibles valores para cada imagen clasificada: *cover*, *stego* o no clasificada.

Hay más restricciones que podemos introducir para detectar muestras no alineadas. Llamaremos a estas restricciones *filtros*, y al filtro descrito en el párrafo anterior lo llamaremos F_1. Es decir, F_1 es el resultado de clasificar a_i con C_A, b_i con C_B y c_i con C_C, y verificar que los resultados son consistentes:

$$F_1(i) \equiv \begin{cases} \text{Si } C_A(a_i) = \mathcal{S}_A, & \text{Si } (C_B(b_i) \neq \mathcal{D}_B) \vee (C_C(c_i) \neq \mathcal{T}_C), \\ & \text{etiquetar como "no clasificada",} \\ \text{Si no,} & \text{Si } (C_B(b_i) \neq \mathcal{S}_B) \vee (C_C(c_i) \neq \mathcal{D}_C), \\ & \text{etiquetar como "no clasificada".} \end{cases}$$

en la Figura 1 podemos ver una representación gráfica de las clases. La alineación perfecta de las muestras de entrenamiento y de test se muestra en la Figura 2. en la Figura 4, podemos ver muestras no alineadas que pueden ser detectadas con el filtro F_1.

Ahora podemos definir otros filtros. Si clasificamos $a_i \in A^{test}$ usando C_B, esperamos que se clasifique como *stego* y no "doble *stego*". Si a_i se clasifica "doble *stego*" por C_B es que hay algún tipo de desalineación. La misma idea se puede externder a imágenes *cover*: si usamos C_A para clasificar una muestra $b_i \in B^{test}$ esperamos que b_i se clasifique como *stego* y no como *cover*. Llamaremos a este tipo de filtro F_2. en la Figura 5 podemos ver el tipo muestras que detecta.

$$F_2(i) \equiv \begin{cases} \text{Si } C_B(a_i) \neq \mathcal{S}_B, & \text{etiquetar como "no clasificada",} \\ \text{Si } C_A(b_i) \neq \mathcal{S}_A, & \text{etiquetar como "no clasificada".} \end{cases}$$

Una idea similar se puede aplicar usando los conjuntos B^{test} y C^{test} en lugar de A^{test} y B^{test}, respectivamente. Si clasificamos una muestra $b_i \in B^{test}$ con C_C esperamos obtener "doble *stego*" y no "triple *stego*". De la misma manera, si usamos C_B para clasificar una muestra $c_i \in C^{test}$ esperaremos obtener "doble *stego*" y no *stego*. Llamaremos a este filtro F_3

y podemos ver, en la Figura 6, qué tipo de muestras detecta:

$$F_3(i) \equiv \begin{cases} \text{Si } C_C(b_i) \neq \mathcal{D}_C, & \text{etiquetar como “no clasificada”,} \\ \\ \text{Si } C_B(c_i) \neq \mathcal{D}_B, & \text{etiquetar como “no clasificada”.} \end{cases}$$

Nótese que existe un caso que no puede ser detectado por estos filtros: cuando una muestra se clasifica de forma incorrecta por todos los clasificadores. en la Figura 3 podemos ver una representación de este tipo de casos.

Existen varias formas de aplicar estos filtros. Hemos elegido las siguientes combinaciones (1) filtrado *soft*: únicamente aplicamos F_1; (2) filtrado *medium*: aplicamos F_1 y F_2; y (3) filtrado *hard*: aplicamos los filtros F_1, F_2 y F_3. Los diferentes tipos de filtrado se han seleccionado de esta manera debido a que es más difícil cumplir los requisitos en C^{test} comparado con B^{test} o A^{test}, pues cuantas más operaciones de inserción se realizan, más difícil es para el clasificador separar las dos clases. Los experimentos realizados muestran que C_C es menos preciso al clasificar imágenes en $\mathcal{D}_C$ y $\mathcal{T}_C$ que C_B al clasificar las mismas imágenes en $\mathcal{S}_B$ y $\mathcal{D}_B$. Así, el filtrado *soft* no tiene que cumplir ninguna restricción en C^{test}, el filtrado *medium* introduce algunas restricciones en C^{test} y, finalmente, el filtrado *hard* usa todos los filtros, ampliando así la dificultad para cumplir todas las restricciones.

Tal y como se muestra en la Sección IV, el filtrado incrementa la precisión de el clasificador de forma considerable, lo que nos permite lidiar con el problema del CSM, a cambio de obtener muestras sin clasificar. Podemos aprovechar esta situación para mejorar el clasificador. Si el número de muestras sin clasificar es elevado, podemos intentar mejorarlo añadiéndole nuevas imágenes. Después de reentrenar el clasificador, podemos clasificar de nuevo el conjunto de test para comprobar si el número de muestras no clasificadas se reduce. Si es así, sabremos que las nuevas imágenes completan nuestro conjunto de entrenamiento.

IV. Resultados experimentales

Los experimentos se han realizado usando imágenes de las bases de datos siguientes:

- BOSS, que contiene imágenes de la competición *Break Our Steganographic System!* [2]. Está formada por 10.000 imágenes de 512×512 píxeles, obtenidas de siete cámaras diferentes.
- ESO del European Southern Observatory [5], con imágnes de tamaño variable alrededor de 1200×1200 píxeles.
- Interactions (INTE), que contiene imágenes de Interactions.org [10] de tamaño variable alrededor de 600×400 píxeles.
- NOAA, que contiene imágenes de la National Oceanic and Atmospheric Administration [18] con tamaños variables alrededor de 2000×1500 píxeles.
- Albion (ALBI), que contiene imágenes de la Plant Image Database del Albion College [1] de tanaño 1024×685 píxeles.
- Calphotos (CALP), que contiene imágenes del Regents of the University of California [3] de tamaño variable alrededor de 700×500 píxeles.

Para los siguientes experimentos, hemos preparado una base de datos de entrenamiento usando 5000 imágenes escogidas aleatoriamente de la base de datos BOSS. Cada una de las 5000 imágenes aparece dos veces en el conjunto de entrenamiento, como *stego* (con mensaje) y como *cover* (sin mensaje). Para los conjuntos de test A^{test} hemos obtenido 250 imágenes de las bases de datos de manera que sean diferentes a las imágenes de entrenamiento. De las 250 imágenes, 125 se usan como *cover* y a las 125 restantes se les incrusta un mensaje aleatorio con una clave aleatoria. Estos conjuntos de datos han sido seleccionados con CSM, por lo que siempre usaremos el mismo conjunto de entrenamiento (que proviene de la base de datos BOSS) y diferentes conjuntos de test. Uno de los conjuntos de entrenamiento proviene también de la base de datos BOSS para poder analizar el método propuesto sin CSM.

Para los experimentos se han usado tres algoritmos de esteganografía diferentes en el dominio espacial: *Highly Undetectable steGO* (HUGO) [7], *Wavelet Obtained Weights* (WOW) [8] y *UNIversal WAvelet Relative Distortion* (UNIWARD). [9]. Además, para cada imagen, se ha generado una clave aleatoria diferente. Cada uno de los experimentos se ha realizado usando el mismo algoritmo y ratio de inserción en el conjunto de entrenamiento y en el de test. Con los algoritmos de esteganografía seleccionados, la inserción requerida para construir los conjuntos A^{test}, B^{test} y C^{test} no siempre es posible. En algunos casos (unas 2 ó 3 imágenes por conjunto de test), el proceso de inserción falla. Hemos decidido conservar dichas imágenes y marcarlas como no clasificadas, dado que en una situación real no podríamos descartarlas.

Además, hemos usado el conocido *framework Rich Models* (RM) [6] con *Ensemble Classifiers* (EC) [13]. Este *framework* cumple los requisistos introducidos en la Sección III, es decir, permite clasificar los conjuntos A^{test} (en $\mathcal{C}_A$ y $\mathcal{S}_A$), B^{test} (en $\mathcal{S}_B$ y $\mathcal{D}_B$) y C^{test} (en $\mathcal{D}_C$ y $\mathcal{T}_C$) para HUGO, WOW, UNIWARD y algoritmos similares. No pretendemos que el método propuesto funcione para algoritmos diseñados para eludir el *framework* RM+EC, como el presentado en [15], dado que el método propuesto depende del clasificador subyacente.

IV-A. Detección de muestras no alineadas

En esta sección se presentan los resultados obtenidos con el método *Non-Aligned Sample Detection* (NASD). En la Tabla I, podemos ver diferentes bloques, cada uno con su correspondiente base de datos de entrenamiento, el algoritmo y la ratio de incrustación usados. En cada fila, la base de datos de test va variando, mostrando a continuación los resultados de clasificación. La primera fila de cada bloque corresponde al caso en el que no hay CSM. Las otras filas muestran situaciones en que la base de datos de entrenamiento y la de test son diferentes y, por lo tanto, se produce CSM. Los resultados se expresan en ratio total de errores “Err_T” (error sobre todas las imágenes), ratio de error “Err_C” (error sobre las imágenes clasificadas). También se proporcionan los positivos

Tabla I
RESULTADOS DE CLASIFICACIÓN PARA NASD

		Train	Test	RM+EC	H-NASD						M-NASD						S-NASD						
ALGO	BR	DB	DB	Err	Err_C	Err_T	TP	TN	FP	FN	NC	Err_C	Err_T	TP	TN	FP	FN	NC	Err_C	Err_T	TP	TN	FP	FN	NC
HUGO	0.4	BOSS	BOSS	0.11	0.04	0.03	74	87	3	4	82	0.04	0.03	78	95	3	5	69	0.05	0.04	93	105	5	6	41
HUGO	0.4	BOSS	ALBI	0.07	0.05	0.02	41	43	4	0	162	0.04	0.02	57	45	4	0	144	0.06	0.04	57	110	5	5	73
HUGO	0.4	BOSS	CALP	0.25	0.16	0.02	14	12	4	1	219	0.17	0.03	26	13	7	1	203	0.16	0.06	32	55	14	2	147
HUGO	0.4	BOSS	ESO	0.50	0.00	0.00	2	1	0	0	247	0.00	0.00	2	1	0	0	247	0.40	0.02	2	4	0	4	240
HUGO	0.4	BOSS	INTE	0.50	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250
HUGO	0.4	BOSS	NOAA	0.50	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250
HUGO	0.2	BOSS	BOSS	0.22	0.10	0.03	37	38	2	6	167	0.00	0.06	43	54	3	13	137	0.13	0.09	60	84	6	16	84
HUGO	0.2	BOSS	ALBI	0.35	0.43	0.12	7	34	11	20	178	0.45	0.16	8	40	16	24	162	0.41	0.21	31	46	21	32	120
HUGO	0.2	BOSS	CALP	0.39	0.50	0.01	0	2	2	0	246	0.75	0.02	0	2	3	3	242	0.34	0.08	11	29	10	11	189
HUGO	0.2	BOSS	ESO	0.47	0.00	0.00	0	2	0	0	248	0.00	0.00	1	2	0	0	247	0.44	0.15	1	47	0	38	164
HUGO	0.2	BOSS	INTE	0.47	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250	0.48	0.34	0	93	1	85	71
HUGO	0.2	BOSS	NOAA	0.49	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250	0.49	0.30	0	77	0	75	98
WOW	0.4	BOSS	BOSS	0.16	0.05	0.02	55	65	2	4	124	0.07	0.04	62	75	2	8	103	0.10	0.08	77	98	7	13	55
WOW	0.4	BOSS	ALBI	0.50	0.40	0.01	3	0	2	0	245	0.76	0.05	4	0	13	0	233	0.72	0.05	4	1	13	0	232
WOW	0.4	BOSS	CALP	0.50	1.00	0.00	0	0	1	0	249	1.00	0.00	0	0	1	0	249	0.50	0.00	0	1	1	0	248
WOW	0.4	BOSS	ESO	0.43	0.00	0.00	0	1	0	0	249	0.00	0.00	0	1	0	0	249	0.46	0.21	0	62	0	52	136
WOW	0.4	BOSS	INTE	0.50	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250	0.59	0.08	0	13	0	19	218
WOW	0.4	BOSS	NOAA	0.50	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250	0.58	0.30	0	54	0	74	122
WOW	0.2	BOSS	BOSS	0.33	0.12	0.01	6	9	1	1	233	0.31	0.06	14	19	8	7	202	0.28	0.14	30	57	16	18	129
WOW	0.2	BOSS	ALBI	0.48	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250	0.48	0.10	0	28	0	26	196
WOW	0.2	BOSS	CALP	0.51	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250	0.50	0.02	2	4	1	5	238
WOW	0.2	BOSS	ESO	0.50	0.00	0.00	0	0	0	0	250	0.77	0.04	3	0	10	0	237	0.55	0.09	8	11	16	7	208
WOW	0.2	BOSS	INTE	0.50	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250	0.63	0.04	5	1	7	3	234
WOW	0.2	BOSS	NOAA	0.48	0.00	0.00	0	0	0	0	250	0.68	0.06	7	0	15	0	228	0.54	0.22	36	12	41	15	146
UNIW	0.4	BOSS	BOSS	0.13	0.06	0.04	59	72	2	7	110	0.07	0.04	68	78	2	9	93	0.09	0.08	88	99	6	13	44
UNIW	0.4	BOSS	ALBI	0.28	0.42	0.04	7	8	7	4	224	0.63	0.11	8	8	23	4	207	0.36	0.16	8	63	23	17	139
UNIW	0.4	BOSS	CALP	0.46	0.56	0.02	3	1	4	1	241	0.61	0.09	10	4	21	1	214	0.54	0.12	16	9	28	1	196
UNIW	0.4	BOSS	ESO	0.44	0.50	0.00	0	1	1	0	248	0.67	0.02	0	2	4	0	244	0.48	0.12	1	32	4	27	186
UNIW	0.4	BOSS	INTE	0.47	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250	0.62	0.10	0	15	0	24	211
UNIW	0.4	BOSS	NOAA	0.38	0.00	0.00	0	0	0	0	250	1.00	0.00	0	0	1	0	249	0.44	0.14	0	46	1	35	168
UNIW	0.2	BOSS	BOSS	0.24	0.12	0.03	21	29	1	6	193	0.15	0.05	33	37	2	10	168	0.13	0.07	54	71	3	15	107
UNIW	0.2	BOSS	ALBI	0.44	0.36	0.02	5	2	2	2	239	0.48	0.10	10	17	7	18	198	0.47	0.15	25	18	19	19	169
UNIW	0.2	BOSS	CALP	0.48	0.00	0.00	0	1	0	0	249	0.60	0.01	1	1	3	0	245	0.44	0.03	6	4	8	0	232
UNIW	0.2	BOSS	ESO	0.47	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250	0.44	0.08	1	24	1	19	205
UNIW	0.2	BOSS	INTE	0.51	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250	0.78	0.03	0	2	1	6	241
UNIW	0.2	BOSS	NOAA	0.51	0.00	0.00	0	0	0	0	250	0.00	0.00	0	0	0	0	250	0.49	0.08	0	21	0	20	209
TOTAL				**0.40**	**0.12**		**334**	**408**	**49**	**56**		**0.21**		**435**	**494**	**148**	**103**		**0.32**		**648**	**1396**	**258**	**702**	

Tabla II
RESULTADOS PARA NASD-R

		Train	Test	RM+EC	H-NASD						
ALGO	BR	DB	DB	Err	Err_C	Err_T	TP	TN	FP	FN	NC
HUGO	0.40	BOSS	CALP	0.25	0.16	0.02	14	12	4	1	219
HUGO	0.40	BOSS+ALBI	CALP	0.23	0.12	0.03	29	29	3	5	184
HUGO	0.40	BOSS+ESO	CALP	0.23	0.06	0.01	29	31	4	0	186
HUGO	0.40	BOSS+NOAA	CALP	0.34	0.66	0.06	5	3	16	0	226
HUGO	0.40	BOSS+INTE	CALP	0.25	0.36	0.03	8	6	7	1	228
HUGO	0.40	BOSS+ALBI+ESO	CALP	0.21	0.07	0.02	40	39	3	3	165
HUGO	0.40	BOSS+ALBI+ESO+NOAA	CALP	0.18	0.07	0.02	40	41	2	5	160
HUGO	0.40	BOSS+ALBI+ESO+NOAA+INTE	CALP	0.20	0.09	0.03	36	39	2	6	167

verdaderos (TP), los falsos positivos (FP), los negativos verdaderos (TN) y los falsos negativos (FN), así como el número de muestras no clasificadas. Como referencia, también indicamos los resultados obtenidos usando el *framework* EC+RM. Sin embargo, una comparación directa no sería justa, puesto que EC+RM clasifica todas las muestras, mientras que el método propuesto puede elegir no clasificar. Por otra parte, el método propuesto está también basado en RM+EC, pero con una fase adicional de filtrado.

En la Tabla I, se observa que NASD con filtrado *hard* ("H-NASD") detecta el CSM de forma precisa. Cuando hay CSM los falsos positivos y falsos negativos son casi cero, dejando una gran cantidad de imágenes sin clasificar, tal y como se espera. A medida que se va usando un filtrado menos estricto, el número de no clasificados decrece, aunque incrementa el error de clasificación. Al usar H-NASD se produce un considerable número de muestras no clasificadas, incluso en casos donde no hay un CSM obvio. En contraste, obtenemos unos errores de clasificación muy reducidos. En la última fila de la Tabla I mostramos el número total de TP, TN, FP y FN. El método propuesto, con todas las opciones de filtrado, nos llevan a un error de clasificación mejorado (Err_C) respecto a RM+EC. Como se puede observar, el método propuesto permite identificar las imágenes *cover* y *stego* con una alta fiabilidad a cambio de dejar las muestras con CSM sin clasificar.

IV-B. NASD reforzando la base de datos de entrenamiento

En esta sección presentamos los resultados para NASD con refuerzo de la base de datos (NASD-R). La idea tras este experimento consiste en mostrar que el conjunto de entrenamiento puede ser completado sin añadir imágenes innecesarias. En este caso hemos usado algunas de las bases de datos para reforzar el entrenamiento.

La primera fila de la Tabla II muestra un ejemplo con CSM: se usan imágenes de BOSS para entrenar e imágenes

de Calphotos para el test. Bajo estas condiciones, y usando el filtrado H-NASD, se producen 219 no clasificados. En la segunda fila se muestra el resultado después de reforzar el conjunto de entrenamiento con imágenes de la base de datos Albion. El número de no clasificados disminye a 184. La tercera fila muestra el resultado después de reforzar el conjunto de entrenamiento con imágenes de ESO. En este caso, el número de no clasificados es de 186. En la cuarta y la quinta filas la base de datos se refuerza con imágenes de NOAA e Interactions, respectivamente. El número de no clasificados no mejora, de hecho, empeora. A la vista de los resultados, vemos que Albion y ESO pueden ser convenientes para mejorar el conjunto de entrenamiento. La sexta fila muestra el resultado obtenido después de reforzar el conjunto de entrenamiento con ambas bases de datos, Albion y ESO. En este caso, el número de no clasificados disminuye hasta 165. El error de clasificación es de $0{,}02$ ($\mathrm{Err_T}$), o $0{,}07$ si consideramos únicamente las imágenes clasificadas. En la séptima y octava filas se muestra el resultado después de reforzar el conjunto de entrenamiento también con NOAA e Interactions. El resultado no mejora significativamente, lo que tiene sentido si vemos los resultados anteriores. Este procedimiento puede ser repetido con otras imágenes hasta que la mayoría de las imágenes queden clasificadas, lo cual nos indicaría que el conjunto de entrenamiento es apto para clasificar el conjunto de test. De esta manera, el problema de CSM quedaría eliminado.

V. Conclusiones

En este artículo, se presenta un método de estegoanálisis supervisado capaz de diagnosticar CSM. Adicionalmente, se muestra una metodología que permite mejorar el conjunto de entrenamiento añadiendo nuevas imágenes hasta obtener uno más adecuado para clasificar un conjunto de test dado. El método propuesto se ha verificado usando tres algoritmos de esteganografía: HUGO, WOW y UNIWARD, y cinco bases de datos de imágenes diferentes que producen CSM. Los experimentos muestran que el método propuesto permite clasificar imágenes con una alta precisión a cambio de dejar las imágenes con CSM sin clasificar.

Como trabajo futuro, sería interesante analizar si otras condiciones también causan muestras no alineadas, como por ejemplo un algoritmo o una ratio de incrustación incorrectos.

Referencias

[1] "Plant Image DataBase from Albion College," Available online at: http://www4.albion.edu/plants/, n.d, accessed on 30 de junio de 2018. [Online].

[2] P. Bas, T. Filler, and T. Pevný, ""Break Our Steganographic System": The Ins and Outs of Organizing BOSS," in *Proceedings of the 13th International Conference on Information Hiding*, ser. IH'11. Springer-Verlag, 2011, pp. 59–70.

[3] "Regents of the University of California, Berkeley," Available: http://calphotos.berkeley.edu/, n.d, (Accessed on 30 de junio de 2018).

[4] G. Cancelli, G. Doërr, M. Barni, and I. J. Cox, "A Comparative Study of ± 1 Steganalyzers," in *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, 2008, pp. 791–796.

[5] "European Southern Observatory," Available: http://www.eso.org/public/images/, n.d, accessed on 30 de junio de 2018.

[6] J. Fridrich and J. Kodovský, "Rich Models for Steganalysis of Digital Images," *IEEE Transactions on Information Forensics and Security*, vol. 7, pp. 868–882, 2012.

[7] G. Gul and F. Kurugollu, "A New Methodology in Steganalysis: Breaking Highly Undetectable Steganography (HUGO)," in *Proceedings of the 13th International Conference on Information Hiding*, 2011, pp. 71–84.

[8] V. Holub and J. Fridrich, "Designing Steganographic Distortion Using Directional Filters." in *International Workshop on Information Forensics and Security (WIFS)*, 2012, pp. 234–239.

[9] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," *EURASIP Journal on Information Security*, vol. 2014, no. 1, pp. 1–13, 2014.

[10] "Interactions.org Particle Physics News and Resources," Available: http://www.interactions.org/cms/?pid=1900, n.d, accessed on 30 de junio de 2018. [Online].

[11] A. D. Ker, P. Bas, R. Böhme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, and T. Pevný, "Moving Steganography and Steganalysis from the Laboratory into the Real World," in *Proceedings of the First ACM Workshop on Information Hiding and Multimedia Security*, 2013, pp. 45–58.

[12] A. D. Ker and T. Pevný, "A Mishmash of Methods for Mitigating the Model Mismatch Mess," in *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 9028, 2014, pp. 90 280I–90 280I–15.

[13] J. Kodovský, J. J. Fridrich, and V. Holub, "Ensemble Classifiers for Steganalysis of Digital Media." *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 432–444, 2012.

[14] J. Kodovský, V. Sedighi, and J. Fridrich, "Study of Cover Source Mismatch in Steganalysis and Ways to Mitigate its Impact," in *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 9028, 2014.

[15] S. Kouider, M. Chaumont, and W. Puech, "Adaptive steganography by oracle (ASO)," in *International Conference on Multimedia and Expo (ICME)*. IEEE, 2013, pp. 1–6.

[16] D. Lerch-Hostalot and D. Megías, "Unsupervised steganalysis based on artificial training sets," *Engineering Applications of Artificial Intelligence*, vol. 50, pp. 45–59, 2016.

[17] I. Lubenko and A. D. Ker, "Going from Small to Large Data in Steganalysis," in *Media Watermarking, Security, and Forensics 2012*, ser. Proceedings of SPIE - The International Society for Optical Engineering, vol. 8303, 2012, pp. 0M01–0M10.

[18] "National Oceanic and Atmospheric Administration (NOAA)," Available: http://www.photolib.noaa.gov/, n.d, accessed on 30 de junio de 2018. [Online].

[19] J. Pasquet, S. Bringay, and M. Chaumont, "Steganalysis with Cover-Source Mismatch and a Small Learning Database," in *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 2425–2429.

[20] T. Pevný, P. Bas, and J. Fridrich, "Steganalysis by Subtractive Pixel Adjacency Matrix," *IEEE Transactions on Information Forensics and Security*, vol. 5, pp. 215–224, 2010.

[21] Y. Qian, J. Dong, W. Wang, and T. Tan, "Deep Learning for Steganalysis via Convolutional Neural Networks," in *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 9409, 2015, pp. 94 090J–94 090J-10.

[22] S. Wu, S. Zhong, and Y. Liu, "Deep Residual Learning for Image Steganalysis," *Multimedia Tools and Applications*, pp. 1–17, 2017.

[23] G. Xu, H. Z. Wu, and Y. Q. Shi, "Structural Design of Convolutional Neural Networks for Steganalysis," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 708–712, 2016.

[24] X. Xu, J. Dong, W. Wang, and T. Tan, "Robust Steganalysis based on Training Set Construction and Ensemble Classifiers Weighting," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 1498–1502.

Técnica de Detección de Manipulaciones de Imágenes Digitales Basada en la Matriz del Filtro de Color

Edgar González Fernández, Esteban Alejandro Armas Vega, Ana Lucila Sandoval Orozco, Luis Javier García Villalba*

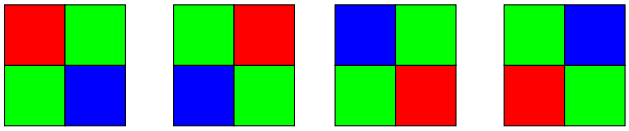
Grupo de Análisis, Seguridad y Sistemas (GASS),
Departamento de Ingeniería del Software e Inteligencia Artificial (DISIA)
Facultad de Informática, Despacho 431, Universidad Complutense de Madrid (UCM)
Calle Profesor José García Santesmases, 9, Ciudad Universitaria, 28040 Madrid, España
Emails: {edggonza, esarmas}@ucm.es, {asandoval, javiergv}@fdi.ucm.es

***Resumen*—La existencia de dispositivos móviles con cámaras de alto rendimiento y potentes aplicaciones de procesamiento de imágenes digitales facilita la modificación de las mismas con fines maliciosos. Este trabajo propone una técnica de detección de manipulación de imágenes digitales basada en las alteraciones de la matriz del filtro de color. En primer lugar, la técnica obtiene una estimación de la matriz de interpolación utilizada en el canal verde de la imagen. Posteriormente, analiza la distribución de los residuales y asigna un valor a cada píxel mediante la función error complementaria, generando un mapa de probabilidades. Finalmente, calcula los coeficientes de la transformada discreta del coseno de las submatrices de este mapa, considerando que la región de la imagen ha sido manipulada si el coeficiente del componente de mayor frecuencia toma valores inusuales.**

***Index Terms*—Manipulación de Imágenes, Matriz de Filtro de Color, Transformada de Coseno Discreta.**

I. Introducción

Los sensores basados en la matriz de filtros de color (CFA, del inglés Color Filter Array) son usados ampliamente en la industria de cámaras fotográficas. La luz es capturada por el sensor, filtrada en distintas longitudes de onda y finalmente, transformada en valores numéricos para generar una imagen visible para el humano. Idealmente se utiliza un sensor para cada longitud de onda, pero supone un alto coste de los dispositivos fotográficos. Por tanto, es común usar filtros que capturen distintos rangos de color. Uno de los más comunes es el conocido filtro de Bayer, que filtra la luz en tres distintos colores: azul, verde y rojo. Existen cuatro configuraciones distintas de dicho filtro, todas ellas consisten de un arreglo matricial (50 % verde, 25 % rojo y 25 % azul).



(a) RGGB (b) GRBG (c) BGGR (d) GBRG

Figura 1. Filtros de Bayer

La presencia de este tipo de filtro en las cámaras fotográficas se hace presente en los llamados artefactos de interpolación (inconsistencias en los píxeles estimados debido a la aplicación de métodos de interpolación). Estas inconsistencias pueden ser visibles, en el caso de algoritmos que no contemplan la detección de orillas, como son la interpolación bilineal o bicúbica, pero también pueden estar presentes sin ser tan evidentes al ojo humano. Se puede identificar en este último la inconsistencia en la distribución de los valores estimados a partir de la interpolación, que están correlacionados con los valores de sus vecinos, a diferencia de valores adquiridos a través del sensor. Se denota por $\mathcal{I}$ y $\mathcal{A}$ a los píxeles interpolados y adquiridos respectivamente.

La presencia de dos distribuciones distintas en imágenes obtenidas mediante filtros de Bayer es estudiada en [1]. A partir de una estimación de los elementos en $\mathcal{I}$ se calcula la varianza local teniendo en cuenta los errores de los N-vecinos más cercanos. Posteriormente, se extrae un valor característico en ventanas de tamaño B considerando los valores en $\mathcal{I}$ y $\mathcal{A}$ por separado. De acuerdo con [2], la varianza de los elementos en $\mathcal{I}$ es mucho menor que en $\mathcal{A}$, por lo que aunado a este hecho se puede exponer una modificación si existen inconsistencias.

Otro enfoque utilizado para detectar alteraciones en el patrón CFA puede encontrarse en [3] donde se utiliza el algoritmo esperanza-maximización (EM) para estimar los coeficientes de interpolación. La propuesta supone que los valores provienen de dos fuentes distintas. Del algoritmo de interpolación estimado y del originado desde una fuente externa (que se considera puramente aleatoria) se genera un mapa de probabilidad, el cuál, en el caso ideal adquiere valores cercanos a 1 en aquellas posiciones que han sido interpoladas y cercanos a 0 en cualquier otro caso, generando un patrón que resalta el filtro RGGB. Posteriormente, se aplica la transformada de Fourier discreta al mapa de probabilidades, haciendo visible una figura distintiva que se presenta únicamente si la rejilla no ha sido alterada.

Por último, mencionamos el aporte hecho en [4], donde los dos procesos principales para la detección de alteraciones son: la extracción de ruido usando la Transformada Wavelet Discreta y el cálculo de la característica definida por

$$F_2 = \text{máx}\left(\frac{var(A_1)}{var(A_2)}, \frac{var(A_2)}{var(A_1)}\right), \qquad (1)$$

donde A_1 y A_2 son las valores del ruido en los pixeles interpolados y no interpolados respectivamente. Valores cercanos a 1 exhiben posibles manipulaciones en el bloque analizado.

En este trabajo se propone una técnica que consiste de manera muy general de 4 pasos:

1. Asumiendo que se conoce la configuración del patrón CFA, se genera una sencilla estimación de los coeficientes de interpolación para el canal verde basada en los mínimos cuadrados ordinarios (OLS del inglés Ordinary Least Squares).
2. Se genera la matriz de errores residuales, es decir, la diferencia entre la estimación y los valores originales. A partir de estos valores se extrae la desviación en el conjunto de valores interpolados y adquiridos, σ_I y σ_A, respectivamente. De acuerdo a [2], se debe cumplir $\sigma_I \approx \frac{\sigma_A}{2}$.
3. De manera similar a [3], se genera un mapa de probabilidad. Sin embargo, se utiliza la función error complementaria, que evaluará la probabilidad de un píxel de pertenecer al conjunto $\mathcal{I}$ a partir del residual en la estimación.
4. Finalmente, se aplica la transformada de coseno discreta (DCT) en ventanas de tamaño $B \times B$ para comprobar la presencia del patrón CFA.

Valores inusuales en el coeficiente de mayor frecuencia al aplicar DCT podrán ser utilizados como evidencia de modificación de la imagen.

El resto del trabajo se divide como sigue: En la Sección II se aborda brevemente el método de mínimos cuadrados aplicado a imágenes para obtener la matriz de residuales. La Sección III explica el uso de los residuales para ser procesado con el mapa de probabilidad y posteriormente con la tranformada DCT. Los resultados obtenidos en los experimentos con imágenes del dataset en [5] se presentan en la Sección IV. Finalmente, la Sección V muestra las principales conclusiones extraídas del trabajo y el trabajo futuro.

II. Estimación de los Coeficientes de Interpolación

El método propuesto se basa en el método de regresión lineal por mínimos cuadrados ordinarios para estimar la interpolación. Considérese una matriz de Bayer C como la que se muestra en la Fig. 2 y sea G la matriz del canal verde construida a partir de C como se muestra en Ec. (2).

$$G(i,j) = \begin{cases} C(i,j) & \text{si } i+j \text{ es par} \\ 0 & \text{en otro caso} \end{cases} \tag{2}$$

Suponiendo que únicamente se emplean los valores de G para interpolar la matriz del canal verde, se tiene que los valores estimados $\hat{g}_{i,j}$ que no han sido obtenidos por el filtro se estiman aplicando Ec. (3).

$$\hat{g}_{i,j} = \sum_{m=-N}^{N} \sum_{n=-N}^{N} h_{m,n} g_{i+m,j+n}. \tag{3}$$

con $h_{m,n} \in \mathbb{R}$ para $m, n = -N, \ldots, N$. Se tiene que de los $(2 * N + 1)^2$ valores $h_{m,n}$, aquellos que coinciden con las posiciones (i, j) tales que $G(i, j) = 0$ son irrelevantes, por lo que les se les asigna el valor 0.

Aunque existen métodos de interpolación más sofisticados que consideran las demás bandas, como los propuestos en [6], por simplicidad, en el método propuesto se considera solo el canal verde debido a los buenos resultados obtenidos en los experimentos. Ya que en general los valores de $h_{m,n}$ son desconocidos, estos se estiman utilizando OLS. Un ejemplo de esto se presenta utilizando $N = 1$ y considerando la matriz de Bayer con los elementos de la Fig. 2.

Las estimación del valor $g_{2,2}$ se muestra en Ec. (4).

$$\hat{g}_{2,2} = h_{1,2}g_{1,2} + h_{2,1}g_{2,1} + h_{2,3}g_{2,3} + h_{3,2}g_{3,2} \tag{4}$$

Se requiere una aproximación $h'_{i,j}$ de tal forma que los residuales de los valores interpolados en Ec. (5) sean tan pequeños como sea posible.

$$\epsilon_{i,j} = \hat{g}_{i,j} - g'_{i,j} \tag{5}$$

Para esto, se aproximan estos coeficientes aplicando OLS donde la variable dependiente Y es el conjunto de valores interpolados y las variables independientes X_i son los N-vecinos correspondientes. Así, las matrices empleadas en la estimación a partir de la matriz de valores interpolados $\hat{G}$ serán de la forma presentada en Ec. (6).

$$Y = \begin{pmatrix} \hat{g}_{1,1} \\ \hat{g}_{1,2} \\ \hat{g}_{2,2} \\ \hat{g}_{3,1} \\ \hat{g}_{3,3} \end{pmatrix}, X = \begin{pmatrix} g_{0,1}, g_{1,0}, g_{1,2}, g_{2,1} \\ g_{0,3}, g_{1,2}, g_{1,4}, g_{2,3} \\ g_{1,2}, g_{2,1}, g_{2,3}, g_{3,2} \\ g_{2,3}, g_{3,2}, g_{3,4}, g_{4,3} \\ g_{2,3}, g_{3,2}, g_{3,4}, g_{4,3} \end{pmatrix}. \tag{6}$$

El valor del vector de coeficientes h' que minimiza la suma de cuadrados de los residuales $\sum_{i,j} \epsilon^2_{i,j}$ está dado por Ec. (7).

$$H' = (X^T X)^{-1} X^T Y. \tag{7}$$

Empleando notación matricial, el conjunto de coeficientes es organizado y escrito como se muestra en Ec. (8).

$$H' = \begin{pmatrix} 0 & h'_{1,2} & 0 \\ h'_{1,4} & 0 & h'_{2,1} \\ 0 & h'_{2,3} & 0 \end{pmatrix}. \tag{8}$$

La aproximación G' obtenida al aplicar la interpolación puede expresarse como la *convolución* $G' = H' \star \hat{G}$. Cabe resaltar que con este método se reduce al mínimo la suma de residuos cuadráticos de los valores interpolados, ya que no se tiene en cuenta los valores adquiridos.

Este método puede emplearse para cualquier valor de N, eligiendo entonces bloques de tamaño $(2N+1)^2$ y extrayendo los valores adquiridos para conformar las filas de la matriz

$r_{0,0}$	$g_{0,1}$	$r_{0,2}$	$g_{0,3}$	$r_{0,4}$
$g_{1,0}$	$b_{1,1}$	$g_{1,2}$	$b_{1,3}$	$g_{1,4}$
$r_{2,0}$	$g_{2,1}$	$r_{2,2}$	$g_{2,3}$	$r_{2,4}$
$g_{3,0}$	$b_{3,1}$	$g_{3,2}$	$b_{3,3}$	$g_{3,4}$
$r_{4,0}$	$g_{4,1}$	$r_{4,2}$	$g_{4,3}$	$r_{4,4}$

Figura 2. Valores del filtro de Bayer

X. En este caso, el número de columnas será $\frac{(2N+1)^2-1}{2}$, ya que sólo la mitad de los valores en este bloque pertenece al conjunto de píxeles adquiridos. El número de filas dependerá del tamaño de la imagen a analizar.

III. Método de Detección Propuesto

Para poder realizar un análisis estadísticamente significativo, es importante realizar una buena estimación de los coeficientes de interpolación, como se ha discutido en la sección anterior. La distribución de los errores da un buen indicador de la bondad de ajuste de la estimación lineal. Esto podrá observase en la sección IV donde se detallan los resultados obtenidos.

Por ahora se considera que la regresión lineal por mínimos cuadrados provee una estimación aceptable, por lo que se utiliza este resultado para la detección de alteraciones en el patrón del filtro.

III-A. Mapa de probabilidad

Como se ha referido anteriormente, ante la presencia de una imagen creada a partir de un filtro de color, aquellos valores calculados a partir de los píxeles adquiridos tendrán una mayor desviación. Esto se aplicará también al estimar los parámetros de la distribución de los residuales. Por tanto, se utiliza esta información para detectar inconsistencias. Para esto, el método propuesto se apoya en el uso una función de probabilidad para decidir si un píxel pertenece al conjunto $\mathcal{I}$ o proviene una muestra con una distribución distinta.

En una imagen sin alteraciones existirán únicamente dos muestras con distinta distribución, a saber $\mathcal{I}$ y $\mathcal{A}$. Idealmente, se requiere una función f que asigne el valor 1 a los elementos del conjunto $\mathcal{I}$ y 0 a aquellos que no.

I	A
A	I

$\xrightarrow{\quad f \quad}$

1	0
0	1

Figura 3. Clasificación de pixeles mediante la función de probabilidad

Se asume que ambos conjuntos se comportan como variables aleatorias normales con media 0 y desviaciones σ_A, σ_I respectivamente, por lo que la función de densidad de $\mathcal{I}$ estará dada por Ec. (9).

$$f(x) = \frac{1}{\sigma_I\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_I^2}}. \tag{9}$$

Adicionalmente, se introduce la *función de error de Gauss* definida como se muestra en Ec. (10) y la *función de error complementaria*, dada por $\operatorname{erfc}(x) = 1 - \operatorname{erf}(x)$, cuya gráfica se muestra en la Fig. 4.

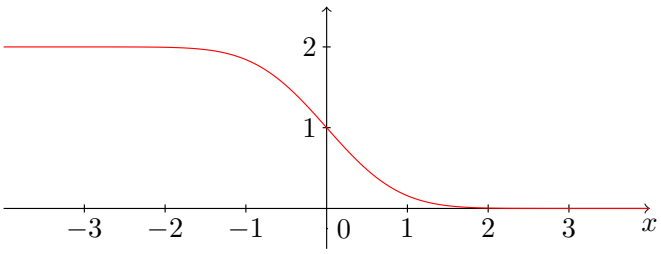


Figura 4. Gráfica de la función error complementaria

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \tag{10}$$

La función de error complementaria se relaciona con la función de distribución normal con media cero y desviación σ, denotada $\Phi_{0,\sigma}$, mediante Ec. (11).

$$\Phi_{0,\sigma}(x) = \frac{1}{2}\operatorname{erfc}\left(\frac{-x}{\sigma\sqrt{2}}\right). \tag{11}$$

A partir de esta relación, y observando que erfc adquiere el valor 1 cuando $x = 0$, y decrece rápidamente conforme x crece, se ha considerado que esta función es ideal para realizar la clasificación, ya que, por hipótesis, los errores siguen una distribución normal y aquellos elementos en $\mathcal{I}$ que son estimados con mayor precisión obtienen valores muy cercanos a 1. Finalmente, se define la función de probabilidad con la Ec. (12).

$$f(x) = \operatorname{erfc}\left(\frac{|x|}{\sigma_I\sqrt{2}}\right). \tag{12}$$

Al aplicar f a la matriz de errores, se desea que aquellos valores que han sido interpolados obtengan una probabilidad más alta a comparación de aquellos que no lo son, debido a que los errores son mucho menores. El patrón 0-1 mostrado en la Fig. 3 es prácticamente imposible de obtener. Sin embargo, si la aproximación de los $h'_{i,j}$ es suficientemente buena, y la hipótesis sobre la distribución de los errores es válida, los valores en la diagonal principal serán menores a aquellos de la antidiagonal. Para establecer una medida de similitud y obtener evidencia de la presencia y, en su caso, alteración del patrón CFA se utiliza la transformada de coseno discreta.

III-B. Transformada de Coseno Discreta

La transformada de coseno es ampliamente utilizada en procesamiento de imágenes y es la herramienta más popular utilizada para compresión de imágenes. Existen diversas definiciones del DCT, las cuales dependen de la forma en la que se extiende la muestra finita de la señal para crear una señal periódica en el dominio de los naturales. Para el método propuesto se ha decidido utilizar la primer variación de la DCT, denotada DCT-I. Para un vector X de tamaño B, la transformada DCT-I unidimensional está dada por

$$y_u = x_0 + (-1)^u x_{B-1} + 2\sum_{i=1}^{B-2} x_i \cos\left(\frac{\pi u i}{B-1}\right). \tag{13}$$

La transformada DCT-I bidimensional se consigue aplicando Ec. (13) en las filas y posteriormente en las columnas resultantes de la matriz a analizar.

Para explicar la utilidad del DCT-I, lo aplicaremos a una matriz $X = (x_{ij})$ de tamaño $B \times B$ definida por

$$x_{ij} = \begin{cases} 1 & \text{si } i+j \text{ es par} \\ 0 & \text{en otro caso} \end{cases} \tag{14}$$

Considerando la i-ésima fila de X para i par, se tiene un vector $X_i = (1, 0, \ldots, 1, 0)$. Aplicando Ec. (13) obtenemos:

$$y_u = \begin{cases} B-1 & \text{si } u = 0, B-1 \\ 0 & \text{en otro caso.} \end{cases} \tag{15}$$

Mientras tanto, para i impar se tiene $X_i = (0, 1, \ldots, 0, 1)$ y

$$y_u = \begin{cases} B-1 & \text{si } u = 0 \\ -(B-1) & \text{si } u = B-1 \\ 0 & \text{en otro caso.} \end{cases} \quad (16)$$

Creamos ahora entonces la matriz cuyas filas están conformadas por los vectores y_u y aplicamos 13, obteniendo la matriz $Y = (y_{uv})$ definida por

$$y_{uv} = \begin{cases} 2(B-1)^2 & \text{si } (u,v) = (0,0), (B-1, B-1) \\ 0 & \text{en otro caso.} \end{cases} \quad (17)$$

Finalmente, normalizamos los coeficientes $y_{u,v}$ para obtener valores que no dependen del tamaño del bloque seleccionado considerando

$$y'_u = \frac{y_u}{2(B-1)} \quad (18)$$

Cabe resaltar que en la práctica, no se ve afectado el método propuesto si se utiliza alguna otra definición de DCT, como lo es el tipo 2 (DCT-II), que es una de las más comunes. Pueden entonces utilizarse implementaciones estándar para procesamiento de imágenes, como la que provee el paquete opencv. Nos referiremos de aquí en adelante por DCT .

Si se consideran bloques de un tamaño pequeño en imágenes comunes, se espera que los valores de los pixeles tengan variación. Esto significa que las bajas frecuencias serán más comunes y la imagen podrá reconstruirse con pocas diferencias al utilizar los coeficientes de las frecuencias más bajas. Para el análisis en curso se requiere un enfoque similar. Sin embargo, en el método propuesto se considera el valor del componente de mayor frecuencia. En la Fig. 5 se observa un bloque de tamaño 2×2 resultado de aplicar la función f a la rejilla de píxeles. Al bloque se le aplica DCT sobre la matriz 0-1 obteniendo los valores mostrados en la segunda matriz.

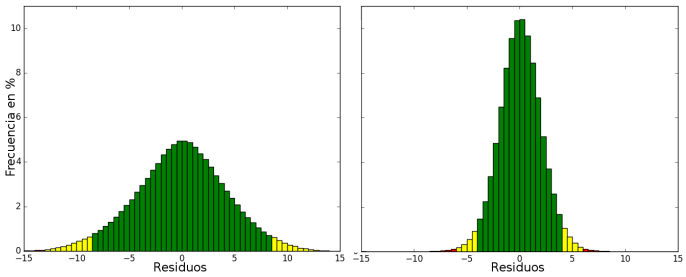
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \xrightarrow{\text{DCT}} \begin{bmatrix} .5 & 0 \\ 0 & .5 \end{bmatrix}$$

Figura 5. Aplicación de DCT a una matriz 2×2

De forma más general, si se aplica DCT en un bloque con valores alternados 0, 1 de tamaño $B \times B$ los coeficientes tomarán el valor 0 a excepción de aquellos en las posiciones $(0,0)$ y $(B-1, B-1)$, cuyos valores serán .5 en cada caso. Como se ha mencionado anteriormente, es prácticamente imposible obtener valores tan precisos mediante las estimaciones que se realizan, pero si estas son suficientemente buenas se pueden obtener valores relativamente altos que indiquen presencia del patrón CFA, aunque se espera que estos sean mucho menores.

IV. Resultados

Para evaluar la propuesta se han analizado 165 imágenes extraídas del dataset utilizado en [6]. Los resultados del análisis han detectado satisfactoriamente alteraciones en 158 de ellas. Sin embargo, en algunos casos la detección ha sido parcial. Además, en algunas de las imágenes originales se aprecian indicativos de manipulación, lo cual indica que el método es proclive a emitir falsos positivos.



(a) (b)

Figura 6. Distribución de residuos en píxeles adquiridos (a) e interpolados (b) en una imagen con filtro RGGB.

A continuación se muestra el análisis realizado a las imágenes de la Fig. 8 y los datos obtenidos del mismo. La imagen de la Fig. 8e ha sido modificada agregando nuevos elementos en ella y recortada a un tamaño de 1024×1024 para poder visualizarla más claramente. Sin embargo, los resultados aquí mostrados son obtenidos aplicando el análisis a la imagen sin recorte.

Primero se ha realizado la estimación de los coeficientes utilizando mínimos cuadrados. Para esta estimación se ha seleccionado $N = 2$, por lo que se tendrán 12 variables independientes. Puede apreciarse en Fig. 6 que los residuos siguen una distribución acorde a lo expuesto anteriormente. Por tanto, es posible concluir que la imagen presenta indicios de haber sido interpolada a partir del patrón RGGB (o posiblemente BGGR). En ambos casos se aprecia la forma de campana de Gauss distintiva de la distribución normal. Es visible además que los residuos de píxeles adquiridos siguen una distribución con una desviación mucho mayor. Posteriormente se calcula el mapa de probabilidad aplicando la función error complementaria y por último se aplica el DCT. Los resultados de estos pasos pueden observarse en la Fig. 8.

En el mapa de probabilidades mostrado en Fig. 8f, obtenidos al aplicar la función de error, se observa de manera tenue la existencia de una sección con una distribución de probabilidad distinta al resto de la imagen. En la Fig. 8g se observa el mapa de calor resultante de aplicar DCT en ventanas de tamaño 16, identificando con mayor claridad el área modificada.

Finalmente, en las Fig. 8d y 8h se muestran los histograma de los valores obtenidos al aplicar DCT. En dichos histogramas se observa la acumulación de un número mayor de valores cercanos al 0 correspondientes a la silueta negra visible en el mapa de calor.

Entre las alteraciones en imágenes que se han podido detectar con nuestra técnica encontramos:

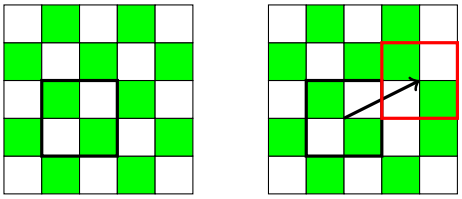


Figura 7. Alteración del patrón RGB por copy move

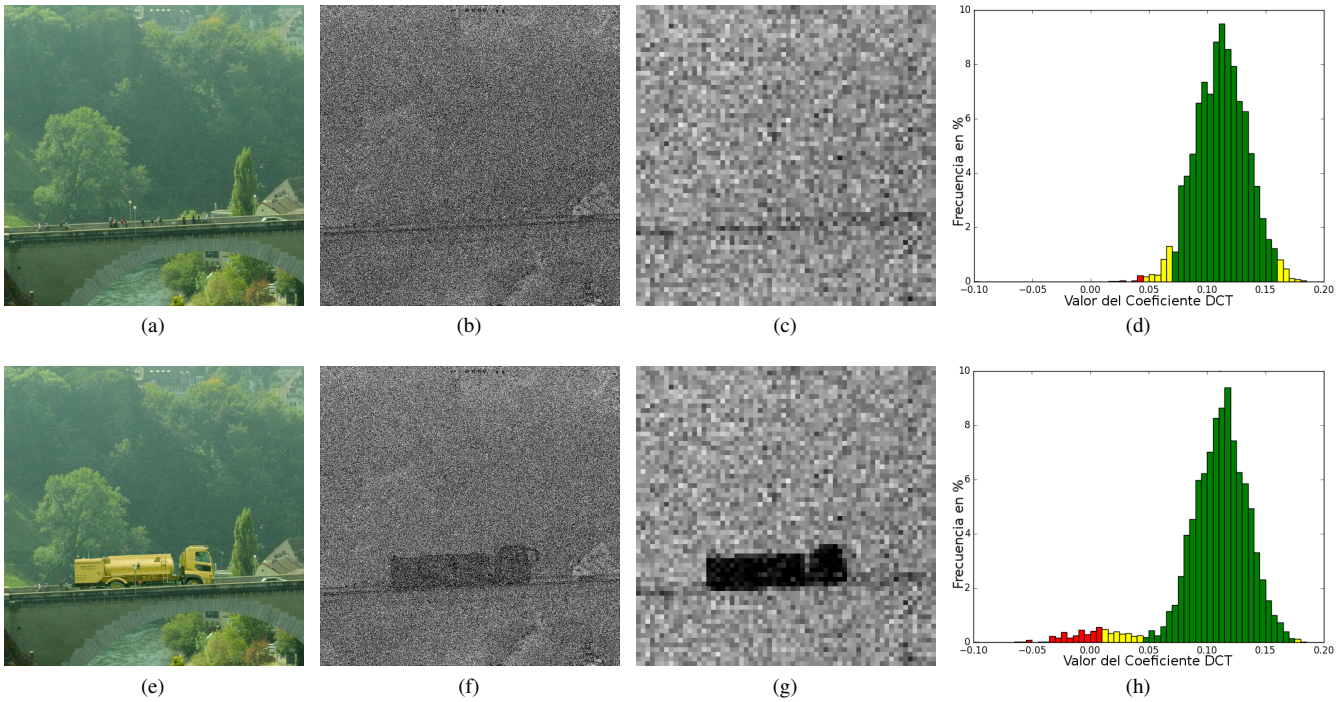


Figura 8. Mapas de probabilidad y de calor en blanco y negro de la imagen original vs la imagen modificada. La parte superior corresponde a la imagen original. La imagen inferior muestra los resultados en la imagen manipulada.

- **Copy-move**. Si el canal verde del filtro de color es alterado de manera similar a lo mostrado en la Fig. 7, entonces la matriz de probabilidades exhibirá esta modificación, lo que se verá reflejado al momento de calcular el DCT. Esto puede ocurrir al agregar elementos externos o bien al mover una sección de la misma imagen en un lugar distinto al original.
- **Filtros de difuminado**. Se ha podido localizar el uso de filtros de difuminado (blur), como lo son: el filtro gaussiano, el de media y el de mediana entre otros. En las secciones que son suavizadas las altas frecuencias son filtradas, y la diferencia entre el valor de las desviaciones σ_I y σ_A desaparece, lo que hace visible la modificación.
- **Inserción de imágenes**. La detección ocurre principalmente por dos motivos. Como se explica en el primer punto, puede perderse la rejilla característica del color verde al insertar una imagen desde una fuente externa. Sin embargo, también puede detectarse una modificiación al agregar elementos obtenidos mediante una interpolación distinta. En este caso es muy probable que la distribución de residuos sea distinta en la zona analizada, y esto será reflejado en los resultados.

Por otro lado, también se han tenido fallos que incluyen tanto falsos positivos como falsos negativos. Al agregar elementos en la imagen ya sea de fuentes externas o de la imagen misma no siempre se altera la rejilla del canal verde. Si además la fuente proviene del mismo dispositivo, es posible que esta no genere ninguna inconsistencia en la distribución de los residuos. Se ha notado también que las zonas que contienen registros de color extremos, como puede ser una zona demasiado iluminada o muy oscura, y por lo tanto con una suavidad poco común, sea detectada como una modificación mediante alguno de los filtros antes mencionados. Mostramos en Fig. 9 un histograma y el mapa de calor de una imagen sin alteraciones, pero en la cual se exhibe un aparente alteración. Esto se da por las zonas extremadamente claras y oscuras que se aprecian en la imagen original. Puede apreciarse que el histograma muestra también una distribución sospechosa, similar al análisis satisfactorio mencionado anteriormente.

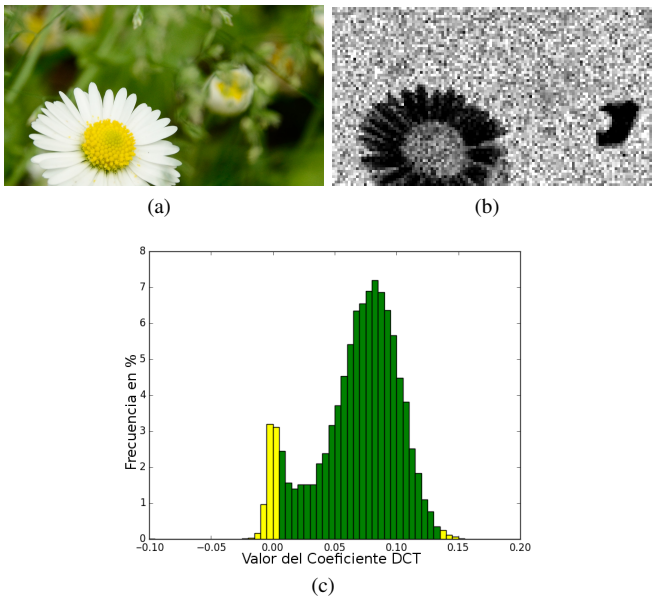


Figura 9. Análisis de una imagen sin alteraciones. Las zonas con colores extremadamente claros y oscuros son detectadas erróneamente como alteraciones en una imagen sin modificar.

Dentro del dataset utilizado en los experimentos se consideran 4 dispositivos distintos: Nikon D90, Nikon D7000, Sony A57 y Canon 60D con 220 imágenes, 55 por dispositivo. A

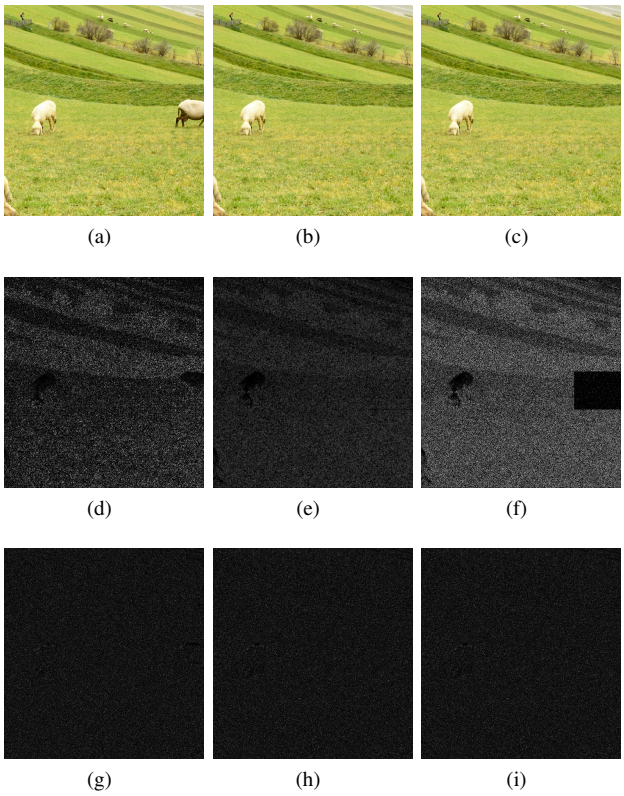


Figura 10. Comparativa entre imágenes analizadas con el método propuesto (d), (e), (f) y con la característica F_2 propuesta en [4] (g), (h), (i). Se Puede notar que ningún método ha reconocido la manipulación realizada en (b), que se trata de un copy-move simple. Sin embargo, al reflejar la sección copiada en (c), nuestro método muestra la zona manipulada. Esto no ocurre con el cálculo de F_2.

diferencia de los 3 primeros, el método descrito ha fracasado en el caso de la cámara Canon 60D. En este caso no se observa la diferencia entre distribuciones de residuos adquiridos e interpolados, no cumpliéndose la hipótesis inicial. Esto puede ser a causa de un proceso de suavizado posterior efectuado por el dispositivo.

IV-A. Comparación con otros métodos

Hemos mencionado 3 métodos propuestos en trabajos relacionados con la detección de alteraciones en el patrón CFA en la sección I. De estos, sólo hemos podido comparar nuestros resultados obtenidos con 2 de ellos, a saber, el propuesto por Popescu [3] y Dirik [4]. En el primero se utiliza el algoritmo EM para estimar el kernel de interpolación. Se genera un mapa de probabilidad el cual es después analizado mediante la Transformada de Fourier Discreta para detectar alteraciones. Este mapa de probabilidad es calculado suponiendo que el ruido puede provenir de una fuente completamente aleatoria o de la interpolación de los píxeles adquiridos. Se ha notado que la función de error complementaria, utilizada en el presente trabajo, clasifica de mejor manera a los píxeles adquiridos e interpolados. Además el cálculo del DCT es más veloz y nuestro análisis funciona bien incluso si utilizamos un kernel muy básico, como lo es el bilineal.

En el segundo trabajo, se tiene que la característica F_2 calculada mediante Ec. 1 detecta adecuadamente manipulaciones siempre que la distribución en el conjunto de pixeles interpolados y adquiridos sea indistinguible, para lo cual se calcula la varianza. Con esto, manipulaciones de tipo copy-move tienen una alta probabilidad de no ser detectadas, pues aún cuando se rompe el patrón CFA como en el caso mostrado en Fig. 7, F_2 mostrará un valor habitual. Esto puede verse en Fig. IV-A, donde además se muestra el caso de un falso negativo, no reconociéndose la alteración de tipo copy-move de Fig. 10b.

V. Conclusiones y Trabajo futuro

Se ha evaluado la técnica desarrollada en imágenes obtenidas cámaras digitales de gama alta. Se realizaron los experimentos utilizando 165 imágenes de un dataset público para probar la fiabilidad de la técnica propuesta. Se optó por esta opción ya que fabricar imágenes ad hoc mediante métodos de interpolación básicos, como son bilineal y bicúbico entre otros, se obtienen resultados muy exactos, pero que no reflejan con exactitud la utilidad de la técnica propuesta. En los resultados de los experimentos se obtuvo un 96 % de acierto en la detección de imágenes manipuladas, siendo un resultado muy satisfactorio. Sin embargo, para robustecer la técnica propuesta se deben considerar las siguientes mejoras: hacer uso de métodos más precisos para el cálculo de las matrices de interpolación, ampliar la técnica aquí utilizada a las bandas de color rojo y azul y obtener datos alternativos que permitan reducir la probabilidad de fallos, principalmente en cuanto a los falsos positivos se refiere. Para lograr esto, se planea hacer un estudio mas avanzado de métodos de estimación para imagenes de color, considerar también el coeficiente de menor frecuencia en el DCT, que representaría un promedio del mapa de probabilidad, y recopilación de datos en una cantidad mayor de imágenes para realizar un análisis numérico más preciso basado en un entrenamiento previo.

Agradecimientos

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 700326. Website: http://ramses2020.eu

Referencias

[1] P. Ferrara, T. Bianchi, A. De Rosa, and A. Piva, "Image forgery localization via fine-grained analysis of CFA artifacts," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1566–1577, October 2012.

[2] A. C. Gallagher and T. Chen, "Image authentication by detecting traces of demosaicing," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Anchorage, AK, USA, July 2008.

[3] A. Popescu and H. Farid, "Exposing digital forgeries in color filter array interpolated images," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3948–3959, October 2005.

[4] A. E. Dirik and N. Memon, "Image tamper detection based on demosaicing artifacts," in *Proceedings of the IEEE International Conference on Image Processing*, Cairo, Egypt, November 2009, pp. 1497–1500.

[5] P. Korus and J. Huang, "Multi-scale analysis strategies in PRNU-based tampering localization," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 809–824, April 2017.

[6] D. Menon and G. Calvagno, "Color image demosaicking: An overview," *Signal Processing: Image Communication*, vol. 26, pp. 518–533, October 2011.

Técnica de Agrupamiento de Vídeos de Dispositivos Móviles

Raquel Ramos López, Ana Lucila Sandoval Orozco, Luis Javier García Villalba*

Grupo de Análisis, Seguridad y Sistemas (GASS),
Departamento de Ingeniería del Software e Inteligencia Artificial (DISIA)
Facultad de Informática, Despacho 431, Universidad Complutense de Madrid (UCM)
Calle Profesor José García Santesmases, 9, Ciudad Universitaria, 28040 Madrid, España
Emails: raqram01@ucm.es, {asandoval, javiergv}@fdi.ucm.es

***Resumen*—La realización de vídeos con dispositivos móviles es una actividad habitual en nuestra sociedad dada su facilidad de grabación y distribución. Estos vídeos pueden utilizarse para deslindar responsabilidades o como parte de la evidencia en procesos judiciales. De ahí la relevancia del análisis forense de vídeos para determinar su origen y autenticidad. El propósito de esta rama del análisis forense multimedia es relacionar a un individuo con un dispositivo. Este artículo presenta una técnica de identificación de la fuente de vídeos en escenarios abiertos, es decir, el analista no conoce el conjunto de posibles videocámaras que pudo generar el vídeo investigado. Este es un planteamiento muy cercano a situaciones de la vida real puesto que en la mayoría de los casos el analista desconoce el conjunto de videocámaras sospechosas. Para evaluar la eficiencia de la técnica se realizaron experimentos con vídeos de diferentes resoluciones.**

***Index Terms*—Analysis, Clustering, PRNU, Sensor Pattern Noise, Video Forensics, Video Source.**

I. Introducción

Las cámaras digitales de los teléfonos inteligentes son tan útiles y simples de usar que han conseguido poner fin al rápido crecimiento de las cámaras digitales compactas. Precisamente, la posibilidad de usar el teléfono móvil como videocámara es una de las funcionalidades a las que los usuarios dan mayor importancia. Conscientes de esta realidad, los fabricantes de dispositivos móviles enfocan sus esfuerzos en mejorar las prestaciones de la videocámara.

La aparición de estos dispositivos ha supuesto un cambio importantísimo en los hábitos de conducta y de comunicación en las personas. En muy poco tiempo, estos dispositivos se han convertido en el epicentro del interés de una gran mayoría de personas, de modo que gran parte de sus actividades cotidianas se canalizan a través del teléfono inteligente.

Como consecuencia de este fenómeno y de la gran cantidad de tiempo que una persona pasa junto a un teléfono inteligente, éste se ha convertido en el primer dispositivo de grabación de vídeos para muchos usuarios en la sociedad actual. A su vez estos vídeos digitales pueden ser realizados en cualquier momento en cualquier lugar y con diferentes fines, distribuyéndose en internet en un corto periodo de tiempo y mostrando en ocasiones actos ilegales. Pero al mismo tiempo, los vídeos se exhiben con mayor frecuencia en procesos judiciales como pruebas o evidencias para la aplicación de la ley.

Si además se considera que en la actualidad existen herramientas de manipulación de vídeos cada vez más sofisticadas, que dificultan la identificación de las características de interés, resulta útil la aplicación de este tipo de técnicas. En el análisis forense de vídeos digitales existen dos principales ramas: la identificación de la fuente de adquisición y la detección de manipulaciones malintencionadas. Este trabajo se centra en la primera de ellas. El análisis forense de vídeos ha demostrado ser más difícil con respecto al análisis de imágenes puesto que los datos que contienen los vídeos tienen formatos de compresión más altos, que pueden comprometer las "huellas" existentes haciendo más complicado recuperar el procesamiento de un vídeo desde su origen.

Dentro de la identificación de la fuente de adquisición existen dos grandes enfoques: escenarios cerrados y abiertos. Un escenario se considera que es cerrado cuando la identificación de la fuente de vídeos se realiza sobre un conjunto de videocámaras concreto y conocido a priori. Para este enfoque se suele utilizar normalmente un conjunto de vídeos de cada videocámara para entrenar un clasificador y posteriormente predecir la fuente de adquisición de los vídeos objeto de la investigación. Una de las técnicas más utilizadas para la clasificación de vídeos digitales es la máquina de soporte vectorial (SVM del inglés *Support Vector Machine*), aunque hay otros enfoques como pueden por ejemplo el uso de redes neuronales.

En un escenario abierto el analista forense no conoce el conjunto de videocámaras a las que pertenecen los vídeos a identificar. Este tipo de clasificación, en la que no se tienen datos de vídeos conocidos a priori el objetivo no es identificar la marca y modelo de éstos, sino poder agrupar en conjuntos los fotogramas de los vídeos que pertenecen al mismo vídeo.

Este trabajo se centra en la identificación de la fuente de vídeos digitales en escenarios abiertos, puesto que este enfoque es muy cercano a situaciones de la vida real, debido a que es casi imposible poder tener un conjunto de vídeos de todas las videocámaras de dispositivos móviles.

Además de esta sección introductoria, el trabajo se estructura en 4 secciones más: En la Sección II se presentan brevemente los trabajos previos relacionados con las técnicas de agrupamiento. En la Sección III se presenta las contribuciones de este trabajo. La Sección IV describe en detalle la experimentación realizada para evaluar la efectividad de las técnicas propuestas, mostrando los resultados obtenidos. Por último, la Sección V resume las conclusiones de este trabajo.

II. Trabajos Relacionados

En los últimos años han aparecido numerosas investigaciones prometedoras para identificar el dispositivo de adquisición de la fuente de imágenes fotográficas estáticas siendo casi inexistente la literatura en el caso de vídeo. Hay que tener en cuenta que un vídeo está formado entre otros elementos por una secuencia de imágenes que aparecen en pantalla en sucesión rápida produciendo sensación de movimiento, que se denominan fotogramas. Los fotogramas que componen un vídeo contienen demasiada información redundante. Para averiguar el contenido esencial de un vídeo la cuestión que debe abordarse es la eliminación de información redundante que reducirá significativamente la cantidad de información a procesar.

Los fotogramas claves son las imágenes fijas que mejor representan el contenido de la secuencia de un vídeo de manera abstracta. Cabe destacar que el vídeo como ente unitario no es clasificado en un tipo de fuente, es decir, lo que identifica su fuente son las unidades mínimas de las que está compuesto un vídeo, denominadas fotogramas. Por lo tanto, el primer paso para averiguar la fuente del vídeo digital consiste en determinar los fotogramas más representativos del vídeo, debido a que cada uno de ellos posee una "huella digital" que sirve para identificar dicha fuente. Este trabajo utiliza el algoritmo propuesto en [1] para realizar la extracción de fotogramas claves, debido a que este algoritmo presta especial atención a la naturaleza del vídeo teniendo en cuenta que el algoritmo coge los fotogramas que tienen mayor variación de escena. En [2] se demuestra que la mayoría de las técnicas de análisis forense de identificación que se pueden aplicar a una imagen pueden ser aplicadas a los diferentes fotogramas de un vídeo.

Para realizar cualquier tipo de clasificación de vídeos digitales ya sea tanto en escenarios abiertos como en escenarios cerrados es necesario obtener ciertas características que permitan a las técnicas de clasificación realizar su tarea. Estas clasificaciones se centran principalmente en el diseño de técnicas que pueden identificar y extraer propiedades de los componentes de una cámara para poder realizar la identificación.

En [3] se realiza una comparación detallada de los principales grupos de técnicas de fuente de adquisición que son divididas en cinco grupos principales: basadas en metadatos, defectos de la matriz (CFA del inglés *Color Filter Array*), características de la imagen, interpolación cromática, imperfecciones del sensor junto con transformadas de Wavelet. Este trabajo utiliza las técnicas basadas en las imperfecciones del sensor, concretamente aquellas basadas en el patrón de ruido del sensor (SPN del inglés *Sensor Pattern Noise*) el cual es originado por las imperfecciones en el proceso de fabricación de los semiconductores del sensor.

El objetivo del agrupamiento es relacionar una colección de objetos en clases representativas llamadas grupos, sin información a priori, de forma que los objetos pertenecientes a cada grupo guarden una mayor similitud con respecto a los objetos de otros grupos. La agrupación de fotogramas claves puede realizarse mediante técnicas de aprendizaje supervisadas o sin supervisión. En el primer caso es indispensable conocer la información del dispositivo a priori al igual que ocurre en los escenarios cerrados.

En [1] se realiza la identificación de vídeo utilizando una máquina de soporte vectorial SVM. Esta técnica requiere tener vídeos procedentes del mismo modelo para poder realizar dicha clasificación. El problema que ocurre en un escenario real es que puede llegar a resultar muy difícil contar con la videocámara en cuestión, por ese motivo se hace necesario utilizar técnicas de aprendizaje sin supervisión o escenarios abiertos.

En la literatura existe una amplia variedad de técnicas de agrupamiento no supervisado. En [4] se propone un algoritmo de agrupamiento basado en la combinación de agrupamiento jerárquico con momentos de Hu. El algoritmo propuesto se compone de dos pasos. El primer paso llamado etapa de agrupamiento grueso se basa en la combinación de agrupamiento jerárquico y el cálculo del coeficiente de silueta. Para mejorar el rendimiento de la clasificación y poder solucionar problemas como los valores de coeficientes de silueta muy pequeños que obstaculizan la capacidad de la etapa de agrupamiento grueso se propone una segunda etapa denominada agrupamiento de refinamiento que consiste en aplicar el vector de momento de Hu a las imágenes que fallan en la etapa de agrupación gruesa, para lograr una clasificación mejor.

En [5] se propone un algoritmo sin supervisión utilizando el ruido del sensor (PRNU del inglés *Photo Response Non Uniformity*). Se demuestra que las estimaciones del ruido PRNU mejoran constantemente con el número de residuos de una misma cámara por cada grupo, la técnica propuesta es un enfoque recursivo aglomerativo, utilizando el algoritmo de agrupamiento del vecino mas cercano (PNN del inglés *Pairwise Nearest Neighbor*).

En [6] se realiza un estudio del efecto que tiene la de compresión de vídeo en la identificación de la cámara fuente basada en el patrón de ruido PRNU, demostrando que los fotogramas tipo I obtienen mejores resultados que los fotogramas tipo P para realizar el cálculo de la estimación del patrón PRNU. Con esta observación, se propone una nueva técnica para estimar el patrón de ruido PRNU.

En [7] se presenta un algoritmo de agrupamiento denominado *OClustR*, se trata de un nuevo algoritmo de agrupamiento basado en gráficos para crear grupos superpuestos. El algoritmo propuesto introduce una nueva estrategia de cobertura de gráficos y una nueva estrategia de filtrado, que en conjunto permiten construir agrupaciones superpuestas con mayor precisión que otros algoritmos propuestos con anterioridad en la literatura. Con esta técnica, se comprueba que se construyen menos grupos que con otros algoritmos utilizados con anterioridad.

En la literatura, existen investigaciones que se centran en mejorar los tiempos de ejecución cuando se utiliza un gran volumen de datos: En [8] se realiza una técnica para deducir el tiempo de procesamiento cuando tratan bases de datos de gran tamaño, de esta manera pueden analizar bases de datos complejas, sin recurrir a la utilización de supercomputadoras.

En [9] Se evalúa una técnica novedosa basada en el algoritmo de Cortes Normalizados para realizar el agrupamiento sin supervisión de imágenes. La técnica propuesta logra resultados buenos en términos de precisión y complejidad computacional.

III. Técnica de Agrupamiento de Vídeos

La técnica de agrupamiento de vídeos propuesta se basa en el algoritmo de agrupación sin supervisión presentado en [10].

La técnica propuesta se divide en 2 fases: (1) Extracción del patrón del ruido del sensor de los vídeos bajo investigación y (2) clasificación de los patrones extraídos basada en un enfoque no supervisado.

Como se ha comentado en la introducción, los vídeos contienen formato de compresión más altos que comprometen sus huellas. Se considera que algunos patrones de ruido extraídos contaminan las escenas. Por tanto, la técnica propuesta utiliza uno de los modelos de mejoramiento propuesto en [11]. Existen otros trabajos donde también se utilizan funciones de mejora para el patrón del ruido [12] pero los resultados no mejoraron la tasa de acierto en [11].

El Algoritmo 1 describe las fases de la técnica de agrupamiento propuesta.

Algorithm 1: Algoritmo de Agrupamiento

1: Extraer fotogramasClaves de un vídeo V
2: Selección aleatoria del número de fotogramas $f_i \in V$ por cada vídeo V
3: Calcular el patrón de ruido n_i de cada fotograma $(i \in 1, ..., N)$
4: Generar la matriz de similitud con función mejorada $H \in R^{NxN}$.
5: Para cada valor de $q \in 0, ..., N$ hacer:
6: Encontrar el par de grupos $H(l, k)$ con mayor similitud
7: Eliminar las filas y columnas correspondientes a los grupos k y l
8: Calcular los valores del nuevo grupo k y l con la la función criterio enlace promedio y agregar las filas y las columnas.
9: Determinar el coeficiente silueta global SC_q
10: Almacenar partición P_q.
11: Encontrar la partición con menor coeficiente silueta $min_q(SC_q)$.

III-A. Extracción del Patrón del Ruido del Sensor

En esta fase se siguen los siguientes pasos:

Primero, dado un vídeo de entrada de 12 a 30 fotogramas por segundo (fps) es necesario dividirlo en fotogramas claves mediante la correlación de histograma de color. A continuación se realiza la extracción de un conjunto de fotogramas claves, esto es, con cambios de escena significativos. Para iniciar este proceso se marca como fotograma clave el primer fotograma, calculando la diferencia de histogramas entre cada fotograma y el último fotograma clave considerado mediante la correlación del histograma de color. Si la diferencia de histogramas satisface un cierto umbral, entonces el fotograma actual se selecciona como fotograma clave. Este proceso se repite para todos los fotogramas que componen el vídeo, hasta que se extraen el conjunto de fotogramas claves. Esta etapa es crucial para el resto del proceso. En la siguiente etapa se extrae el patrón de ruido de sensor de cada fotograma perteneciente al conjunto de fotogramas claves. Las características se obtienen utilizando la Transformada de Wavelet. Esta primera parte del algoritmo está basado en la propuesta [1].

En segundo lugar, es necesario obtener los patrones de ruido del sensor del conjunto de fotogramas I^i, $i = 1 \dots N$ utilizando el algoritmo de extracción propuesto en [13].

$$n^{(i)} = I^{(i)} - F(I^{(i)}) \tag{1}$$

donde n es el patrón de ruido de cada fotograma i, I es el conjunto de fotogramas con ruido y F es el filtro de extracción de ruido basado en la transformada de Wavelet.

III-B. Clasificación de Vídeos

Cada uno de los N ruidos n_1,...,n_N se obtiene el valor de correlación usando la *Ec.(1)* y junto con la ecuación *Ec.(2)* se genera una matriz de similitud *H* de dimensiones *NxN*. Esta matriz es simétrica y está compuesta de 1's en su diagonal principal.

$$n_e(i,j) = \begin{cases} e^{-0{,}5\frac{n^2(i,j)}{\alpha^2}} & , 0 \leq n(i,j) \\ -e^{-0{,}5\frac{n^2(i,j)}{\alpha^2}} & , \text{otro caso} \end{cases} \tag{2}$$

donde $n(i,j)$ y $n_e(i,j)$ son los componentes de la matriz de ruido para cada uno de los fotogramas que componen el vídeo. El parámetro α determina la atenuación del ruido es un valor que se debe ajustar en cada caso. En nuestro caso particular, se ha utilizado valores diferentes de $\alpha \in [-0{,}1, 20, 200, 5000]$ y tras realizar varios experimentos se ha llegado a la conclusión que el valor óptimo es $\alpha = 20$.

El algoritmo de agrupamiento jerárquico implementado consiste en seleccionar dentro de la matriz *H* el par de ruidos n_1,...,n_2 cuyo valor de correlación sea más alto sin tener en cuenta el valor de correlación de la diagonal principal (que siempre vale 1). Las filas y las columnas correspondientes a n_1,...,n_2 son eliminadas y se generan una nueva fila y columna con la unión de ambos. Dicha unión se realiza usando una función de criterio de enlace. Para este trabajo se realizaron pruebas con varias funciones criterio de enlace que son criterio simple, completo o criterio promedio. Tras realizar varios experimentos la función criterio que mejores resultados mostró fue la de criterio promedio al igual que en el caso de imágenes [10]. La ecuación *Ec. (3)* muestra la función criterio de enlace promedio entre dos grupos A y B.

$$H(A,B) = \frac{1}{|A||B|} \sum_{n_i \in A, n_j \in B} corr(n_i, n_j) \tag{3}$$

|A|y |B|es la cardinalidad de los grupos de A y B respectivamente. El valor de corr(n_i, n_jn) determina la similitud que existe entre dos objetos de un mismo grupo alcanzando su valor más alto cuanto más similares sean Xi y Xj. En la siguiente ecuación *Ec. (4)* se calcula la función de correlación normalizada.

$$corr(X_i, X_j) = \frac{(X_i - \overline{X_i}) \odot (X_j - \overline{X_j})}{||X_i - \overline{X_i}|| \cdot ||X_j - \overline{X_j}||} \tag{4}$$

Para este trabajo se ha determinado utilizar la función de correlación normalizada en lugar de otras como distancia Manhattan o Chebychev puesto que esta función es una de las medidas más utilizadas en la identificación de la fuente de imágenes [12], [10].

Cada iteración del algoritmo selecciona los dos grupos con el valor de correlación más elevado en la matriz H, mezcla los objetos contenidos en éstos para crear un grupo nuevo, y almacena el estado de los distintos grupos en particiones P_0, ...,$P_N - 1$.

En el agrupamiento jerárquico tradicional el resultado final es que todos los elementos pertenezcan al mismo grupo pero en este trabajo necesitamos que cada grupo represente un dispositivo al final de la ejecución. Para poder validar los grupos se ha utilizado como medida de validación de grupos el coeficiente silueta. El coeficiente de silueta mide la distancia promedio de un punto a todos los demás puntos en el mismo grupo (cohesión a_j), por otro lado también mide la distancia promedio de un punto de uno de los grupos a todos los demás grupos cercanos (separación b_j) como se muestra en la siguiente ecuación:

$$s_j = max(b_j) - a_j \tag{5}$$

El coeficiente de Silhoutte para todo el agrupamiento es:

$$SC_q = \frac{a}{N} \sum_{i=1}^{K} s_j \tag{6}$$

Para cada iteración q del algoritmo se obtiene una media global de todos los coeficientes siluetas calculados a partir de los K grupos. Una vez concluido el agrupamiento jerárquico se procede a buscar el SC_q con el valor mínimo, esto indica que los grupos de la partición P_q tienen un nivel de correlación mayor. En el caso ideal el número de grupos debería corresponderse con el número real de videocámaras.

IV. Experimentos y Resultados

Los experimentos se realizaron con un total de 5 dispositivos de diferentes fabricantes. Los vídeos utilizados en cada uno de los experimentos fueron capturados sin ningún tipo de restricción temporal o espacial, grabados tanto en espacios interiores como exteriores puesto que deben representar casos reales. La Tabla *I* muestra las especificaciones de cada uno de los dispositivos utilizados.

Tabla I
Configuraciones Utilizadas en las Videocámaras.

Marca	Modelo	Formato	Códec	Resolución
Apple	iPhoneX	mov	H.264	1920x1080
Lenovo	K5	3gpp	H.264	640x352
OnePlus	One	3gpp	H.264	640x352
Samsung	Galaxy A3	mov	H.264	1920x1080
Samsung	Galaxy S8	mov	H.264	1920x1080

En la Tabla II se resumen las condiciones experimentales.

Tabla II
Condiciones Experimentales

Parámetros	Valor
Método de extracción de fotogramas	Histograma de color
Vídeos por dispositivo	1
Máximo número de fotogramas por vídeo	100
Duración por cada vídeo(segundos)	10

Para realizar el primer grupo de experimentos, se han utilizado los dispositivos de las marcas Apple y Samsung con una resolución de 1920x1080 píxeles. El tamaño de recorte seleccionado para cada uno de los fotogramas es de 1024x1024 centrado. Para la extracción del patrón de ruido de cada uno de los fotogramas se ha utilizado el promedio a cero *(zero-mean)* de filas y columnas, los tres canales de color RGB se han convertido a una sola matriz en escala de grises. El primer experimento se ha realizado sin aplicar la función de mejora explicada en la *Ec.(2)*, con la idea de comprobar la efectividad de la técnica propuesta de este trabajo. Se han realizado diferentes pruebas variando el número de fotogramas por cada vídeo 5, 20 y 100 respectivamente. En la tabla *III* se muestran los resultados obtenidos. Cuando no se aplica la función de mejora la tasa máxima de acierto que se logra es del 77 %.

Tabla III
Tasa de Acierto. Sin Aplicar Función de Mejora

Número Fotogramas	Tasa Acierto
5	72 %
20	74 %
100	77 %

Se ha repetido el experimento aplicando la función de mejora con diferentes valores del parámetro $\alpha \in [-0{,}1, 20, 200, 5000]$.

La Tabla IV se puede comprobar que la mayor tasa de acierto ocurre cuando $\alpha = 20$, al aumentar este valor de $\alpha = 200$ o $\alpha = 5000$ no se obtiene ninguna mejoría en los resultados llegando a obtener las mismas clases de agrupamiento. Por lo tanto tras la realización de este experimento se puede apreciar que el valor óptimo es $\alpha = 20$ y que la diferencia de aplicar o no la función de mejora aumenta la tasa de acierto del 77 % al 92 %.

Tabla IV
Tasa de Acierto. Función de Mejora

Fotogramas por vídeo	Valores de α			
	α=-0.1	α=20	α=200	α=5000
5	68 %	67 %	67 %	67 %
20	56 %	67 %	67 %	67 %
100	81 %	92 %	92 %	92 %

Aunque aplicando la función de mejora se obtienen mejores resultados, el tiempo de procesamiento es mucho más elevado que si no se aplica, como se puede apreciar en la Tabla V.

Tabla V
Tiempo Procesamiento.

Fotogramas por vídeo	Tiempo de Ejecución(seg)	
	Sin Mejora	Con Mejora
5	12	258
20	101	1050
100	2518	7448

El resto de experimentos que se detallan a continuación se han realizado aplicando la función de mejora con el valor de $\alpha = 20$.

En todos los experimentos realizados, se han obtenido más clústeres que dispositivos, la ventaja es que de esos clústeres en más del 80 % de los casos corresponden a fotogramas del mismo dispositivo, llegando incluso al 95 %. Por lo tanto en este primer experimento podemos obtener como conclusión

que con un tamaño de recorte de 1024x1024 centrado en cada uno de los fotogramas, se produce un acierto superior al 80 %. En la Tabla VI se muestra el porcentaje de acierto por dispositivo.

Asimismo, se ha realizado un conjunto de experimentos seleccionando al azar los diferentes fotogramas de cada vídeo, en este conjunto de experimentos no se ha podido obtener conclusiones categóricas sobre utilizar unos fotogramas u otros llegando a la conclusión que estos fotogramas se pueden seleccionar de manera aleatoria a pesar que en la literatura existen otras técnicas que indican que seleccionado unos fotogramas concretos se obtienen mejores resultados [6].

Tabla VI
PORCENTAJE DE ACIERTO. TAMAÑO DE RECORTE 1024X1024

Fotogramas por vídeo	Número de Dispositivos	
	2	3
5	93 %	95 %
20	91 %	92 %
100	85 %	80 %

Un segundo grupo de experimentos se realizaron con un tamaño de recorte 320x320 centrado en cada uno de los fotogramas, en este experimento se utilizaron los siguientes modelos de dispositivos Iphone X , OnePlus One, Samsung Galaxy S3 y Lenovo K5. Al igual que en el caso anterior, los fotogramas son (5, 20 y 100) respectivamente. Cuando la cantidad de los fotogramas es inferior a 100 éstos se seleccionaron de forma aleatoria. La Tabla VII muestra los resultados obtenidos del experimento.

Tabla VII
PORCENTAJE DE ACIERTO. TAMAÑO DE RECORTE 320X320

Fotogramas por vídeo	Número de Dispositivos		
	2	3	4
5	64 %	65 %	66 %
20	34 %	37 %	36 %
100	61 %	66 %	68 %

Se realizó otro experimento con los mismos dispositivos que en el caso anterior, y se observó que a menor tamaño de recorte la tasa de acierto disminuye de un 80 % a un 66 %.

Mediante los resultados de los experimentos realizados se observa que existe una mejora en la tasa de acierto para un tamaño de recorte de 1024x1024 con respecto al tamaño de recorte de 320x320. También hay que tener en cuenta que a mayor tamaño de recorte mayor es el tiempo de ejecución del algoritmo de extracción de características.

V. CONCLUSIONES

En este trabajo se ha realizado una técnica de identificación de la fuente de adquisición producido por un dispositivo móvil en escenarios abiertos, es decir, el analista forense no conoce a priori el conjunto de videocámaras a las que pertenece el vídeo a identificar su fuente. Se debe tener en cuenta que en este tipo de clasificación, en la que no se tienen datos de cámaras a priori, el objetivo no es identificar la marca y modelo de la cámara, sino poder agrupar los diferentes vídeos en grupos disjuntos en los que todos sus fotogramas pertenecen al mismo dispositivo. Este planteamiento es muy cercano a situaciones de la vida real, ya que en muchos casos el analista desconoce por completo el conjunto de videocámaras a las que pueden pertenecer un conjunto de vídeos y es prácticamente imposible tener un conjunto de vídeos para entrenar un clasificador con todas las videocámaras de dispositivos móviles existentes en el mundo. En este caso es de gran utilidad poder agrupar los vídeos en conjuntos que pertenecen a un mismo dispositivo, ya que esto puede aportar información muy válida y en casos concluyente a los investigadores judiciales. Una vez presentada la técnica y realizados los experimentos variando los distintos parámetros, se llega a la conclusión general de que esta técnica es válida y obtiene buenos resultados.

AGRADECIMIENTOS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 700326.

REFERENCIAS

[1] L. J. García Villalba, A. L. Sandoval Orozco, R. Ramos López, and J. C. Hernández Castro, "Identification of Smartphone Brand and Model via Forensic Video Analysis," *Expert Systems with Applications*, vol. 55, no. 15, pp. 59–69, August 2016.

[2] P. Bestagini, K. Fontani, S. Milani, M. Barni, A. Piva, M. Tagliasacchi, and K. Tubaro, "An Overview on Video Forensics," in *Proceedings of the 20th European Signal Conference (EUSIPCO)*, Bucharest, Romania, August 2012, pp. 1229–1233.

[3] A. L. Sandoval Orozco, D. M. Arenas González, J. Rosales Corripio, L. J. García Villalba, and J. C. Hernández-Castro, "Techniques for Source Camera Identification," in *Proceedings of the 6th international conference on information technology*, Amman, Jordan, May 2013, pp. 1–9.

[4] O. Famhy, "An Efficient Clustering Technique for Cameras Identification Using Sensor Pattern Noise," in *Proceedings of International Conference on Systems Signals*, London, UK, September 2015, pp. 249–252.

[5] F. Marra, G. Poggi, C. Sansone, and L. Verdoliva, "Correlation Clustering for PRNU-Based Blind Image Source Identification," in *Proceedings of International Workshop on Information Forensicis and Security (WIFS)*, Abu Dhabi, United Arab Emirates, December 2016, pp. 1–6.

[6] W. Chuang, H. Su, and M. Wu, "Exploring Compression Effects for Improved Source Camera Identification Using Strongly Compressed Video," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, Brussels, Belgium, September 2011, pp. 1953–1956.

[7] A. Pérez-Suárez, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and J. E. Medina-Pagola, "OClustR: A new graph-based algorithm for overlapping clustering," *Neurocomputing*, vol. 121, pp. 234–247, December 2013.

[8] F. Gisolf, P. Barens, E. Snel, A. Malgoezar, M. Vos, A. Mieremet, and Z. Geradts, "Common source identification of images in large databases," *Forensic science international*, vol. 244, pp. 222–230, November 2014.

[9] I. Amerini, R. Caldelli, P. Crescenzi, A. Del Mastio, and A. Marino, "Blind Image Clustering Based on the Normalized Cuts Criterion for Camera Identification," *Signal Processing: Image Communication*, vol. 29, no. 8, pp. 831–843, September 2014.

[10] L. J. García Villalba, A. L. Sandoval Orozco, and J. Rosales Corripio, "Smartphone Image Clustering," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1927–1940, March 2015.

[11] L. Chang-Isun, "Source Camera Identification Using Enhanced Sensor Pattern Noise," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 2, pp. 280–287, September 2010.

[12] R. Caldelli, I. Amerini, F. Picchioni, and M. Innocenti, "Fast Image Clustering of Unknown Source Images," in *Proceedings of International Workshop on Information Forensicis and Security (WIFS)*, Seattle, WA, USA, December 2010, pp. 1–5.

[13] J. Lukas, J. Fridrich, and M. Goljan, "Digital Camera Identification from Sensor Pattern Noise," *IIEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 205–214, June 2006.

Evaluación de un sistema forense para BitTorrent

Alberto Caravaca
Dep. Teoría de la Señal, Telemática y Comunicaciones
Universidad de Granada - España
albertocp@correo.ugr.es

Gabriel Maciá-Fernández
Dep. Teoría de la Señal, Telemática y Comunicaciones - CITIC
Universidad de Granada - España
gmacia@ugr.es

***Resumen*—El protocolo BitTorrent es uno de los más utilizados para la compartición de archivos. Debido a su naturaleza distribuida, monitorizar la red BitTorrent es una tarea difícil. Este déficit de monitorización crea una sensación de anonimato que incita a que prolifere la compartición de archivos ilegales tales como información clasificada,violaciones de la ley de propiedad intelectual, contenidos pedófilos, etc. Este trabajo se centra en desarrollar un sistema que permite identificar si un fichero dado está siendo compartido o no en la red BitTorrent. El desarrollo del sistema se ha dividido en dos módulos, el primero encargado de la monitorización y la recolección de información de los archivos compartidos y el segundo dedicado al procesamiento y búsqueda de un fichero dado con el fin de detectar si ha sido compartido en la red. El sistema ha sido implementado y se realizan experimentos sobre la red real para determinar su correcto funcionamiento.**

***Index Terms*—BitTorrent, peer-to-peer, sistema forense**

I. Introducción

La compartición de información a través de redes como Internet se ha convertido en una de las claves del progreso de nuestras sociedades modernas. El desarrollo de tecnologías que han permitido poner en común una ingente cantidad de información y contenidos ha sido un factor clave en la denominada revolución TIC o revolución digital. En este proceso, algunos servicios de Internet, entre los que destaca el servicio Web como actor principal, juegan un papel fundamental.

Lamentablemente, la compartición masiva de información se ha producido sin el desarrollo paralelo de adecuados mecanismos de control, lo que ha generado algunos efectos secundarios no deseables. Uno de los más relevantes ha sido la proliferación de contenidos de naturaleza ilegal que se comparten en las redes, tales como contenidos pedófilos, recursos que violan la ley de propiedad intelectual, información clasificada y privada, traspaso de información relacionada con el robo industrial, etc.

Así, en torno a esta problemática se ha desarrollado en los últimos años un tejido considerable de empresas que han focalizado su negocio en la monitorización de los contenidos en Internet para ofrecer servicios tales como la protección de propiedad intelectual, persecución de bandas criminales, protección de datos e información personal, protección de la reputación en Internet, garantía del derecho al olvido, etc. Estas empresas han desarrollado sistemas basados en el rastreo (crawling) de los contenidos en Internet, en busca de determinados patrones que les permiten ofrecer sus servicios.

Debido a la gran cantidad de contenidos existentes en Internet, las empresas de este sector han focalizado su negocio en el rastreo del servicio más utilizado para la compartición de contenidos, esto es, el servicio Web. Sin embargo, en la Internet actual existen otros muchos servicios en los que se produce una elevada compartición de contenido sensible o ilegal. Entre ellos destacan los que utilizan una infraestructura basada en redes Peer-to-Peer (P2P).

Las redes Peer-to-peer (P2P) constituyen un sistema muy popular para la compartición de información en Internet. Aunque su popularidad ha caído en los últimos años, el número de usuarios es aún muy relevante, habiendo un tráfico por mes de 6 EB en global, según CISCO [1]. Frente a otros servicios en Internet como la navegación Web o el correo electrónico, estas redes se caracterizan por su naturaleza distribuida, lo que hace que sean robustas frente a problemas que pudieran afectar a determinados nodos críticos. Generalmente, se utilizan para compartir contenidos digitales de muy diversa naturaleza, como películas, música, libros e, incluso, software, aunque tienen otras muchas aplicaciones. Sin embargo, sus características hacen que también sean redes idóneas para compartir contenido de tipo ilegal.

En la actualidad, el servicio de tipo P2P con mayor utilización es BitTorrent. Este servicio se utiliza principalmente para la compartición de recursos (archivos) y se caracteriza, entre otras cosas, por un rendimiento muy eficiente en la descarga y compartición de los mismos. La amplia difusión del servicio BitTorrent ha motivado que muchos trabajos de investigación se hayan dirigido hacia este tópico, con el objetivo de estudiar diversos aspectos, tales como la naturaleza de los recursos que se están compartiendo, o el desarrollo de metodologías y sistemas de monitorización para comprender la evolución del tráfico en esta red [2][3][4].

Específicamente, en este artículo se plantea el diseño y evaluación previa de un sistema que, partiendo de un recurso dado como entrada (fichero con contenidos tales como películas, música, texto, ejecutables, etc.) pudiera servir de apoyo a un equipo judicial o policial para determinar si un contenido (sensible o ilegal) ha sido compartido en BitTorrent. Los retos de investigación a resolver están relacionados con determinar hasta qué punto se puede responder a las siguientes preguntas:

- ¿Se está compartiendo dicho recurso en la red BitTorrent?
- ¿Quién está realizando la compartición de dichos recursos?
- ¿Cómo ha sido la evolución temporal de la compartición de estos recursos? Esto es, cuántos usuarios lo han descargado y sus perfiles.
- ¿Se ha respetado a lo largo del proceso de diseño y en su funcionamiento la normativa en vigor?
- ¿Podría esta información servir como evidencia electrónica y digital en un procedimiento judicial o en cualquier otro marco jurídico?

En este artículo se plantea el diseño inicial de un sistema

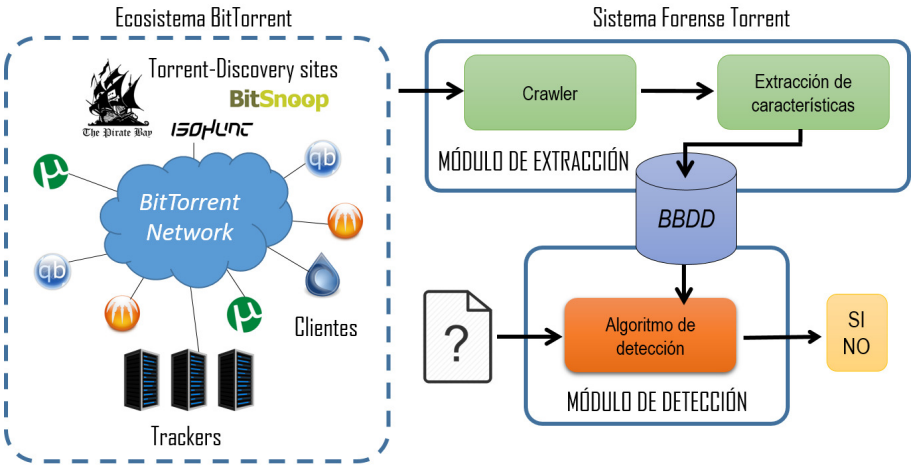


Figura 1. Arquitectura del sistema forense propuesto.

que permita resolver estas preguntas. Se describe dicho diseño y se realiza una evaluación experimental preliminar de su funcionamiento. Así, en el Apartado II se presentan los conceptos básicos sobre BitTorrent para comprender completamente la propuesta. En la Sección III se presenta el diseño del sistema, mientras que los resultados de su evaluación preliminar se muestra en el Apartado IV. Dado que el sistema presenta unas limitaciones, estas se clarifican en la Sección V. Finalmente, en la Sección VI se recogen las principales conclusiones de este artículo.

II. Conceptos básicos sobre BitTorrent

Tal y como se ha comentado anteriormente, el protocolo BitTorrent es un protocolo de tipo P2P (*peer-to-peer*) que sirve para la compartición de ficheros (*recursos*). Los denominados *peers* son instancias de clientes BitTorrent que comparten un recurso en la red. Con el fin de agilizar las transferencias de los recursos, estos se dividen en *piezas* o *trozos* de igual tamaño para cada recurso, exceptuando el último trozo que tendrá un tamaño menor al estar truncado debido a que el tamaño total del recurso no es normalmente múltiplo del tamaño de la pieza. A la hora de compartir un recurso concreto en la red, los diferentes *peers* que participan en esta compartición se pueden comportar como *seeders* o como *leechers*. Los *seeders* tienen todas las piezas de dicho recurso, mientras que los *leechers* solamente poseen algunas de ellas y están tratando de descargar el resto.

Un *enjambre* (*swarm*) se define para un recurso como el conjunto de *peers* (*seeders* y *leechers*) que están participando en la compartición de dicho recurso en la red. Existen unos nodos especiales en la red denominados *trackers*, que son servidores centralizados que mantienen, para cada recurso, la lista de los *peers* que forman el enjambre correspondiente.

Cada recurso compartido en la red está asociado a un fichero denominado *.torrent* que tiene dos partes describiendo las características de dicho recurso. Por un lado, existe una sección `announce` que contiene las URLs de los *trackers* responsables de mantener la información sobre el enjambre del recurso. Por otro, existe otra sección denominada `info`, que contiene la información que describe al recurso: (*i*) *name*: nombre del fichero o del directorio que contiene múltiples ficheros en el recurso; (*ii*) *length*: longitud en bytes (cuando existe un único fichero en el recurso) o la lista de ficheros (cuando el recurso está formado por varios ficheros); (*iii*) *longitud de las piezas*: en número de bytes; (*iv*) *hashes de las piezas*: que es una cadena con los SHA-1 (20 bytes) de las diferentes piezas que conforman el recurso; y (*v*) un parámetro opcional denominado *private* para indicar si el torrent es privado o no. El hash SHA-1 de la sección `info` completa se denomina *infohash*, y permite diferenciar un fichero `.torrent` de otro de forma unívoca.

Para unirse a un enjambre, el *peer* se anuncia a uno o varios de los *trackers* anunciados en la lista del fichero `.torrent`. Después de ello, los *trackers* responden con su lista de *peers* conocidos que en dicho momento están descargando el recurso. Finalmente, el *peer* se conecta a dichos *peers* utilizando el protocolo BitTorrent e inicia el intercambio de piezas con ellos.

Desde 2005, el protocolo BitTorrent implementa también un mecanismo descentralizado para la funcionalidad de *tracker*, denominado DHT (*distributed hash table*). Fue inicialmente implementado en la red BitTorrent Azureus (actualmente conocida como Vuze). La DHT, al igual que los *trackers*, almacena los *peers* que pertenecen a un enjambre, y esto se hace en los propios nodos de la red BitTorrent. En la actualidad, existen dos implementaciones de la DHT para BitTorrent: Vuze y Mainline. Ambas utilizan la implementación denominada Kademlia [5]. En este trabajo nos centraremos en la implementación Mainline, de la que se obtendrán los datos de análisis para el sistema forense.

III. Diseño del sistema propuesto

Una ilustración del sistema propuesto se puede observar en la Figura 1. En ella se aprecia a la izquierda el ecosistema BitTorrent, compuesto por los diferentes agentes, como los clientes, los *trackers*, utilizados para localizar a los *peers* que están compartiendo los recursos, y los *torrent-discovery sites*, que son sitios web que se emplean para descargar los ficheros `.torrent` con la descripción de los recursos que se están compartiendo en la red.

El sistema forense propuesto está representado en la parte derecha de la figura. Podemos observar dos módulos diferenciados. En primer lugar, un módulo de extracción que se

utiliza para interaccionar directamente con la red BitTorrent y extraer la información que el sistema precisa. El objetivo de este módulo es interaccionar con la red a través de un *crawler* o rastreador de la red para obtener los datos precisados por el sistema, en este caso, contenidos en ficheros `.torrent`. La recolección de estos datos se ha realizado insertando 8 nodos en la DHT de la red BitTorrent. Estos nodos tienen implementado el protocolo y el funcionamiento que se indica en el estándar DHT de BitTorrent [6]; sin embargo, además de las funciones de los nodos DHT, se les ha añadido una función adicional propia de los clientes BitTorrent [7], de modo que son capaces de solicitar a un *peer* los metadatos (sección `info` del fichero `.torrent`) correspondientes a un *infohash*.

Estos nodos trabajan con normalidad en la red BitTorrent, siendo propagados por la DHT sin crear sospechas en el resto de nodos. Así, los *peers* les mandan mensajes de tipo `announce_peer`, cuya función es la de anunciar los nuevos recursos que están compartiendo. Cada vez que uno de nuestros nodos recibe un mensaje de este tipo, hace uso de la función adicional y solicita al *peer* los metadatos correspondientes a ese recurso, obteniendo así toda la información relevante que necesitamos sobre ese `.torrent`.

Previamente a almacenar la información en una base de datos, los ficheros `.torrent` se preprocesan en un submódulo de extracción de características, obteniendo las siguientes:

- *Infohash* del fichero torrent.
- Nombre del recurso (tal y como aparece en el fichero `.torrent`). En caso de estar compuesto de varios ficheros, se almacena el nombre de cada uno de ellos.
- Lista de *hashes* de las diferentes partes en que está dividido.
- Longitud de los trozos o partes.
- Tamaño del recurso o de cada uno de los ficheros que lo componen (en caso de múltiples ficheros).
- Indicador sobre si el `.torrent` está marcado como privado o no.

Por otro lado, el módulo de detección será el que reciba como entrada el recurso al que se va a realizar el análisis forense. Dicho recurso pasará por un submódulo que implementa un algoritmo de detección, el cual realizará operaciones optimizadas de correlación con la información contenida en la base de datos con el fin de obtener información sobre las preguntas planteadas anteriormente. En concreto, el recurso a analizar se divide en partes de diferentes tamaños estándar y se calcula el *hash* a dichas partes, con el fin de comprobar si en la base de datos existen o no recursos con dichos valores de *hash*. A continuación, en la Sección IV se muestra algún experimento que ilustra la capacidad de detectar que tendría este sistema simplificado.

IV. Validación experimental

En esta sección se describe en primer lugar el entorno experimental sobre el que se ha trabajado, describiendo los datos analizados y sus características, y en segundo lugar los experimentos realizados para analizar las capacidades de detección del sistema propuesto.

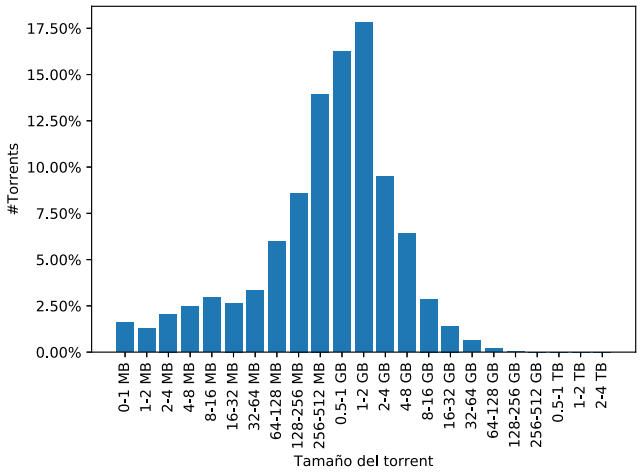


Figura 2. Número de torrents en función del tamaño.

IV-A. *Descripción de los datos analizados*

Para la evaluación del funcionamiento del sistema planteado, se ha recolectado la información de un total de 6 millones[1] de torrents y se han preprocesado con el submódulo de extracción de características. Tras procesar toda la información, el conjunto de datos obtenido ha sido almacenado en una base de datos, ocupando un total de 381 GB. Para determinar si los datos obtenidos son realmente significativos para extraer conclusiones, se ha realizado un análisis exploratorio sobre los mismos.

En la Figura 2 se observa la distribución de tamaños de los torrents analizados. Se puede observar cómo dicha distribución sigue una normal con media centrada en 2.29GB. El intervalo con mayor frecuencia de aparición es el de [1GB-2GB). El máximo valor en este intervalo es algo esperable, dado que la compartición de películas haciendo uso de la red BitTorrent es muy común y el tamaño de algunos de los formatos más utilizados suelen estar comprendidos en ese intervalo. En cuanto a los valores anómalos detectados en el conjunto de torrent, hemos observado la aparición de un torrent cuyo tamaño es de 3.54TB, lo que es inusualmente grande. El porqué de haber encontrado solamente un fichero en este intervalo se debe a que la compartición de ficheros tan grandes es muy lenta y, además, gran parte de los usuarios no disponen de tanta capacidad de almacenamiento. En este caso, el fichero consistía en una tabla *rainbow* de MD5 [8], es decir, *hashes* precalculados para una serie de palabras.

Respecto al número de piezas que componen los torrents, en la Figura 3 se ha representado la cantidad de piezas en intervalos de 200 piezas, desde 0 hasta 9000. Asociado a cada intervalo se muestra el porcentaje de los torrents cuyo número de piezas se encuentra comprendido en dicho intervalo. La cota superior ha sido fijada en 9000 piezas en la figura para mejorar la visualización, ya que el porcentaje de torrents que tienen un número de piezas mayor a esa cota es tan solo el 2.24 % y son valores muy dispersos. El intervalo modal es el de [600-800) piezas, y contiene al 23,82 % de los torrents. A partir de este intervalo, conforme aumenta el número de piezas, el porcentaje de torrents disminuye considerablemente.

[1]El número exacto es 5.979.893.

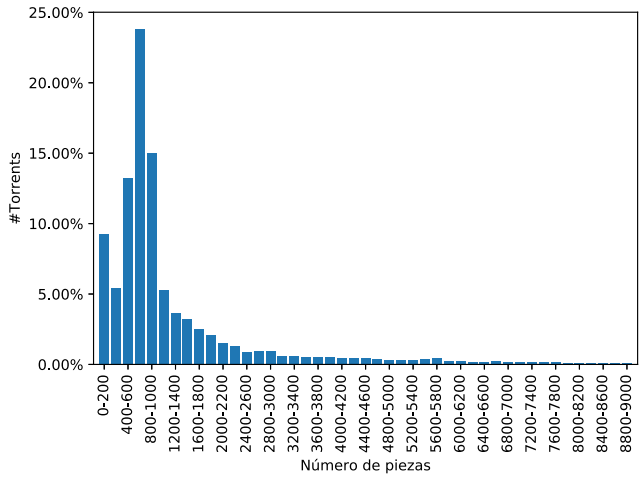


Figura 3. Número de torrents en función del número de piezas.

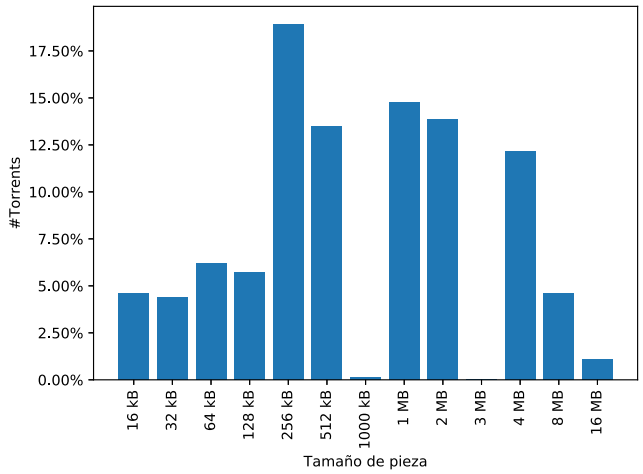


Figura 4. Número de torrents en función del tamaño de pieza.

Según la recomendación establecida en [9] sobre el número de piezas óptimo para un torrent, se esperaba que el intervalo con mayor número de piezas fuera el de [1000-1500] piezas; sin embargo, comprobamos que al tratarse tan solo de una recomendación y al no ser de obligado cumplimiento, los diferentes clientes BitTorrent utilizan algoritmos distintos para asignar de manera automática el tamaño de trozo a cada torrent y por ende, el número de trozos para un mismo fichero puede variar en función del cliente desde el cual haya sido generado. Además, muchos clientes permiten asignar el tamaño de pieza de manera manual sin importar que el resultado obtenido sea o no el óptimo.

Analizando los tamaños de las piezas que componen cada uno de los torrents se han encontrado un total de 766 tamaños diferentes. De ellos, 644 tamaños tan solo aparecen una vez. En la Figura 4 se han representado los tamaños de pieza que aparecen en más de 1000 torrents, dando un total de 13 tamaños diferentes. Estos 13 tamaños abarcan el 99.89 % de los torrents analizados. De ellos, tan solo dos no son múltiplos de 16 KB (1000 KB y 3 MB). Estos tamaños, además de no ser múltiplos de 16 KB, destacan visualmente en la gráfica debido a que son los que tienen el menor porcentaje de aparición. El tamaño de pieza más común es de 256 KB con un 18.92 % y le sigue el tamaño de 1 MB con un 14,75 %. Estos dos tamaños son también los indicados como más comunes en la especificación del protocolo BitTorrent [10]. Además, cabe destacar que de los tamaños encontrados, tan solo 9 de ellos son tamaños impares. Esto enfatiza el hecho de que la aparición de torrents en los que el tamaño se ha seleccionado manualmente es común en la red.

En resumen, se verifica que el conjunto de datos obtenido es representativo de múltiples casos diversos y tipologías de torrent en la red. Es por ello que, en una segunda fase experimental, se ha abordado el análisis de las capacidades de detección del sistema propuesto utilizando este conjunto de datos como partida.

IV-B. Capacidad de detección

El algoritmo de detección que se ha implementado es un algoritmo sencillo que trata de evaluar preliminarmente las capacidades básicas de detección del sistema. En la Sección V se discutirán las limitaciones de esta detección y algunas líneas de trabajo que se están iniciando.

El algoritmo se basa en la comparación de los *hash* de los trozos que componen un torrent, con los almacenados en la base de datos que describe a los torrent observados en la red. Tal y como se ha explicado anteriormente, los distintos hashes de un torrent se calculan aplicándole la función SHA-1 a cada uno de los trozos. Esta función nos permite afirmar con exactitud que si dos hashes son idénticos se debe a que han sido generados a partir del mismo trozo y, por tanto, si los hashes de los trozos de un fichero aparecen en un torrent, podremos afirmar que ese torrent contiene al fichero.

Estamos, por tanto, interesados en conocer hasta qué punto un torrent puede tener trozos en común con otros torrent que se comparten en la red, y también hasta qué punto la existencia de trozos en común podría llevar a detectar que un torrent se está compartiendo cuando en realidad no es así (falso positivo).

Para llevar a cabo esta evaluación se ha extraído una muestra aleatoria con el 10 % del conjunto de torrents previamente recolectados y procesados. El motivo de extraer esta muestra en vez de trabajar con el total de datos recogidos es el de agilizar los análisis realizados, ya que requieren de una capacidad de cómputo elevada.

En primer lugar, se han considerado todas las piezas de cada uno de los torrents pertenecientes a la muestra, buscando si aparecen en algún otro torrent de la muestra, y obteniendo así el número de trozos comunes que cada torrent tiene con cada uno de los otros torrents. Definimos entonces dos medidas:

- *Redundancia* de un torrent, R: es el porcentaje de sus trozos que aparecen en algún otro torrent de la muestra. Esta medida nos indica si hay trozos en un torrent que se repiten en otros torrent cualesquiera, lo que es una medida de redundancia.
- *Máxima similitud* de un torrent, S: si se compara un torrent T_i con otro torrent de la muestra T_k podremos obtener el porcentaje de trozos de T_i que aparecen en T_k. Tomando esta medida con todos los T_k pertenecientes a la muestra, obtenemos la máxima similitud, S, como el máximo porcentaje de coincidencia entre T_i y alguno de los torrent de la muestra. Esta medida nos permitirá

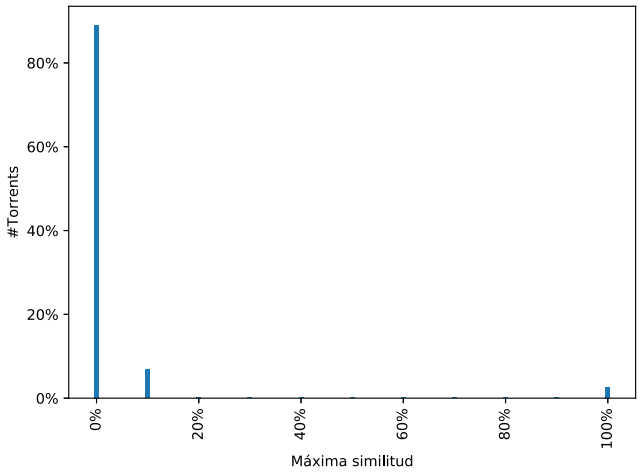


Figura 5. Número de torrents en función de la similitud máxima.

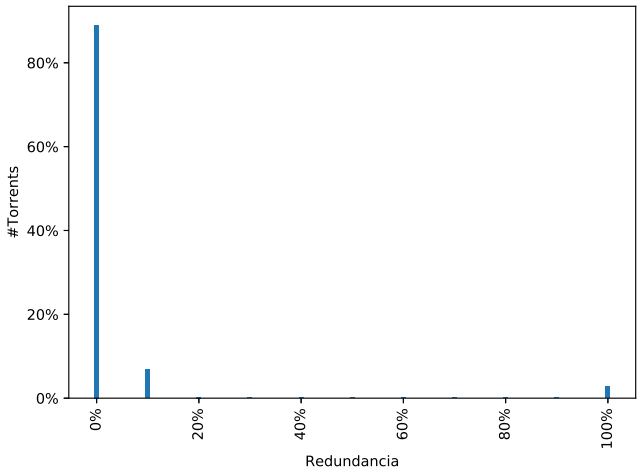


Figura 6. Número de torrents en función de la redundancia.

discernir si el sistema podrá detectar de forma unívoca la presencia de un torrent o si, por el contrario, podría confundirse con la existencia de otros (falso positivo). Por ejemplo, si dos torrent coinciden en un $S = 80\,\%$ de las piezas, podríamos pensar que son el mismo torrent.

En la Figura 5 se presenta la máxima similitud de todos los torrent de la muestra. Se observa que el 89.06 % de los torrents analizados son totalmente únicos para ese tamaño de pieza, no habiéndose producido ninguna coincidencia de sus hashes en otro torrent. Sin embargo, el 6.84 % de los torrents han tenido un porcentaje de aparición entre el 0 % y 10 %, y tan solo un 2.7 % ha tenido un 100 % de similitud, es decir, disponen de todos sus hashes en otro torrent, lo que significa que los ficheros que lo componen se encuentran contenidos en al menos otro torrent.

Es notable que el porcentaje de torrents con hasta un máximo del 10 % de similitud es relativamente elevado. Esto nos ha llevado a realizar un análisis del contenido de los trozos con mayor número de apariciones, ya que podría parecer extraño que torrents diferentes tengan trozos comunes. Tras analizar el contenido de estos trozos, hemos concluido que en la mayoría de los casos estos trozos han resultado estar formados por información de relleno como bytes a 0 o la palabra PADDING concatenada repetidas veces hasta completar el tamaño del trozo. Sin embargo, otros trozos proporcionan información más relevante sobre el contenido del torrent; por ejemplo, hemos detectado un trozo de 16 kB que es común a la cabecera de gran parte de los ficheros de audio de tipo m4a. Otro de los trozos encontrados frecuentemente se trata de un trozo de tamaño 32 kB localizado en la cabecera de algunos archivos con extensión .exe. Además, también hemos detectado trozos comunes que no señalaban a un tipo de fichero en concreto sino que nos permiten identificar los torrents subidos por un mismo usuario, ya que tiene en común alguna imagen o descripción que aparece en todos los torrents subidos por él/ella.

Este análisis nos indica que es posible descartar que se trate del mismo fichero cuando dos torrent tiene una similitud de hasta un 10 %. A partir de este 10 % de similitud es cuando realmente la información que tienen en común el fichero con el torrent es importante y no es simplemente información común al tipo de fichero.

En la Figura 6 hemos analizado la *redundancia* de los trozos de cada torrent, es decir, el porcentaje de los trozos que se repiten en cualquier otro torrent independientemente de si cada trozo se localiza en un torrent distinto o todos en el mismo. Como ya se ha indicado, esto nos permite ver la cantidad de información única. En el caso de un 100 % de redundancia, todas las piezas han sido localizadas en otros torrent; por el contrario, si el porcentaje es bajo nos indica que los trozos que posee son realmente significativos ya que no son compartidos con otros torrents. Observando la figura, los resultados muestran que el 89.06 % de los torrents poseen una redundancia del 0 %. Sin embargo, encontramos una redundancia en el intervalo (0 %,10 %] para el 6,87 % de los torrents y con un 100 % de redundancia tenemos el 2.7 %. El resto de intervalos de redundancia representados tienen una aparición muy baja, siendo todos inferiores al 0.25 %. Si comparamos estos resultados con los resultados de la máxima similitud podemos comprobar que son muy similares. Estos datos nos confirman que en la mayoría de situaciones, los torrents poseen bien una redundancia muy baja o por el contrario muy alta, lo que quiere que los *hashes* de los ficheros son realmente únicos a excepción de cuando existe un torrent que contiene esos mismos ficheros.

V. Limitaciones del sistema

La implementación realizada es una demostración de que el sistema planteado es realmente funcional. Sin embargo, para su aplicación práctica surgen diversas limitaciones que debemos de tener en cuenta. Estas limitaciones son:

- La imposibilidad de detectar modificaciones en los recursos analizados: en el caso de que un fichero haya sido modificado y posteriormente compartido, nuestro sistema actual lo podría detectar como un fichero totalmente distinto y descartaría la similitud; por ejemplo, si un vídeo se vuelve a codificar haciendo uso de un codec diferente, la nueva codificación modificará en gran parte el contenido del fichero y, como consecuencia, al trocearlo, los trozos obtenidos serán totalmente distintos a los del fichero original, dando como resultado que

nuestro sistema actual no sería capaz de detectar las similitudes. Actualmente, estamos trabajando en algoritmos que tienen en cuenta esta problemática.

- El sistema precisa monitorizar la compartición de todos los recursos en la red BitTorrent al completo. La red BitTorrent es muy grande y se encuentra en constante expansión. Para que nuestro sistema sea aplicable a casos reales es necesario analizar en tiempo real todos y cada uno de los torrent compartidos en la red, debiendo de ser capaces de detectar y analizar los nuevos torrents publicados en los *torrent-discovery sites* y los anunciados por los *peer* en cualquier parte de la red. Aunque existen diversos estudios [11][12][13][14] que consideran esta problemática y proponen soluciones y estimaciones para hacer posible esta monitorización, es cierto que la capacidad de cómputo y almacenamiento necesarias son considerables.

VI. Conclusiones

En este trabajo, hemos diseñado un sistema que nos permite detectar si un recurso dado ha sido compartido en la red BitTorrent. Este sistema se encuentra dividido en dos módulos distintos: (*i*) un módulo encargado de la monitorización de la red BitTorrent con el fin de recolectar los torrents compartidos y procesar la información que contienen para su posterior búsqueda; (*ii*) el segundo módulo se encarga de procesar un fichero dado con el fin de poder buscar coincidencias entre los torrents de la red y detectar si este fichero ha sido compartido.

Se ha analizado la capacidad de detección de este sistema, obteniendo una muestra de 6 millones de torrents en la red BitTorrent. Se ha encontrado que un 6,8 % de los torrent analizados tiene trozos redundantes con otros torrents compartidos en la red, debido principalmente a elementos vacíos, cabeceras de tipos de archivo, o elementos que pudieran identificar a usuarios que generaron los torrents. En base a estos resultados, se ha parametrizado el funcionamiento del sistema definiendo un mínimo de un 10 % de similitud en dos torrents para poder considerar probable que fueran el mismo recurso.

Como trabajo actual y futuro, se están llevando a cabo el desarrollo y la implementación de mejoras en ambos módulos. El primer módulo se optimizará para poder analizar la red BitTorrent al completo de una manera eficiente, mientras que el segundo módulo incorporará algoritmos mejorados que permitan detectar si un fichero dado ha sido modificado y posteriormente compartido en la red BitTorrent.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el Gobierno Español-MINECO (Ministerio de Economía y Competitividad) y fondos FEDER, a través de los proyectos TIN2014-60346-R y TIN2017-83494-R.

Referencias

[1] CISCO. Cisco visual networking index: Forecast and methodology, 2016–2021. [Online]. Available: https://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf

[2] R. A. Rodríguez-Gómez, G. Maciá-Fernández, L. Sánchez-Casado, and P. García-Teodoro, "Analysis and modelling of resources shared in the bittorrent network," *Transactions on Emerging Telecommunications Technologies*, vol. 26, no. 10, pp. 1189–1200, 2015.

[3] L. Wang and J. Kangasharju, "Measuring large-scale distributed systems: case of bittorrent mainline dht," in *Peer-to-Peer Computing (P2P), 2013 IEEE Thirteenth International Conference on*. IEEE, 2013, pp. 1–10.

[4] M. Kryczka, R. Cuevas, C. Guerrero, A. Azcorra, and A. Cuevas, "Measuring the BitTorrent ecosystem: Techniques, tips, and tricks," *IEEE Communications Magazine*, vol. 49, no. 9, pp. 144–152, 2011.

[5] P. Maymounkov and D. Mazieres, "Kademlia: A peer-to-peer information system based on the xor metric," in *International Workshop on Peer-to-Peer Systems*. Springer, 2002, pp. 53–65.

[6] A. Loewenstern and A. Norberg. (2017) Bittorrent specification. dht protocol. [Online]. Available: http://www.bittorrent.org/beps/bep_0005.html

[7] G. Hazel and A. Norberg. (2017) Bittorrent specification. extension for peers to send metadata files. [Online]. Available: http://bittorrent.org/beps/bep_0009.html

[8] H. Kumar, S. Kumar, R. Joseph, D. Kumar, S. K. S. Singh, and P. Kumar, "Rainbow table to crack password using md5 hashing algorithm," in *Information & Communication Technologies (ICT), 2013 IEEE Conference on*. IEEE, 2013, pp. 433–439.

[9] VuzeWiki. (2018) Torrent piece size. [Online]. Available: http://wiki.vuze.com/w/Torrent_Piece_Size

[10] B. Cohen. (2018) The bittorrent protocol specification. [Online]. Available: http://www.bittorrent.org/beps/bep_0003.html

[11] R. A. Rodríguez-Gómez, G. Maciá-Fernández, P. García-Teodoro, M. Steiner, and D. Balzarotti, "Resource monitoring for the detection of parasite p2p botnets," *Computer Networks*, vol. 70, pp. 302–311, 2014.

[12] C. Zhang, P. Dhungel, D. Wu, and K. W. Ross, "Unraveling the bittorrent ecosystem," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 7, pp. 1164–1177, 2011.

[13] S. L. Blond, A. Legout, F. L. Fessant, W. Dabbous, and M. A. Kaafar, "Spying the world from your laptop–identifying and profiling content providers and big downloaders in bittorrent," *arXiv preprint arXiv:1004.0930*, 2010.

[14] Z. Xinxing, T. Zhihong, and Z. Luchen, "A measurement study on mainline dht and magnet link," in *Data Science in Cyberspace (DSC), IEEE International Conference on*. IEEE, 2016, pp. 11–19.

Improving a smart metering system using elliptic curves and removing the trusted dealer

Ricard Garra
Dept. Matemàtica,
Univ. de Lleida.
C. Jaume II, 69, Lleida
garra@matematica.udl.cat

Santi Martínez
Dept. Matemàtica,
Univ. de Lleida.
C. Jaume II, 69, Lleida
santi@matematica.udl.cat

Josep M. Miret
Dept. Matemàtica,
Univ. de Lleida.
C. Jaume II, 69, Lleida
miret@matematica.udl.cat

Francesc Sebé
Dept. Matemàtica,
Univ. de Lleida.
C. Jaume II, 69, Lleida
fsebe@matematica.udl.cat

***Abstract*—Fine-grained electricity consumption information has been proven to allow to infer people's habits from their consumption patterns. Hence, smart metering systems need a secure way to regularly transmit the electricity consumption to the supplier in such a way that the privacy of customers is preserved. Several proposals can be found in the literature. Some of them, classified as aggregative solutions, organize the meters into groups and employ homomorphic cryptography in such a way that the supplier only obtains the aggregated consumption values of the meters in each group.**

In this paper we propose a technique which allows to remove the presence of a trusted dealer in one of the most efficient current proposals. Furthermore, we increase the overall performance of the resulting system by employing elliptic curve cryptography. The resulting system only requires a single message sent from each smart meter to its substation, without needing the meters to communicate among them. With the use of elliptic curve cryptography, the only relevant computation that the smart meters need to perform is a single elliptic curve point multiplication by a scalar. The system has been proven to be feasible and practical even for large neigborhoods including several thousands of meters.

***Index Terms*—cryptography, elliptic curve, privacy, smart meter**

I. Introduction

Smart meters are household devices that transmit information about electricity consumption to the energy supplier in short intervals. Such information allows an accurate prediction of consumption trends so that the production can be properly adjusted avoiding electricity surplus. On the other side, customers can benefit from an accurate billing, reduced prices, and better knowledge of consumption habits. The EU aims to replace at least 80% of electricity meters with smart meters by 2020. Spain is expected to be the first in Europe to replace all analogue electricity meters by the end of 2018.

Unfortunately, smart meters have been proven to be privacy invasive. Fine-grained information about electricity consumption allows to infer sensitive information such as the time a customer leaves or arrives at home, watches TV, or goes to bed. Hence, appropriate solutions have to be deployed in order to preserve customers'privacy.

Privacy-preserving proposals for smart metering can be classified into three main classes:

- *Perturbative:* Meters add some random noise to electricity consumption readings before transmitting them to the energy supplier. In this way, the latter only obtains a noisy version of household consumption profiles. Such solutions [1] require to adjust the noise magnitude for trading-off consumer privacy and data accuracy.
- *Anonymous:* Consumption values are transmitted so that the energy supplier can not determine the identity of transmitters. In proposals [2], [3], each meter has a unique pseudonym which is attached to metering data. In proposal [4] data are transmitted without a pseudonym, while, in [5], each pseudonym is shared among several meters.
- *Aggregative:* Meters are clustered into groups that add their readings prior to transmitting them to their substation. Data are usually aggregated by a trusted party [6] or by making use of some homomorphic cryptosystem [7], [8], [9], [10], [11].

In the smart metering literature it is common to find hybrid proposals combining both perturbative techniques with aggregative ones [8], [10]. In those cases, the meters add some random noise to their data before transmitting them so that the substation obtains a noisy aggregation of consumption readings. The noise added by the meters is usually chosen in such a way that *differential privacy* is provided [12].

A. Aggregative solutions

Aggregative solutions are designed so that the substation can only obtain the addition of meters consumption readings. Most proposals make use of homomorphic encryption.

Proposal [7] requires each meter to have a public key of some additive homomorphic cryptosystem. The transmission protocol has a rather elevated $O(n^2)$ communication cost for a neighborhood composed of n meters.

The approach described in [8] makes use of the Paillier cryptosystem. There exists a neighborhood public key whose private key is distributed among all the meters. The protocol for consumption readings transmission has an efficient linear cost, but the key setup for a distributed Paillier is very complicated when a trusted dealer is not desirable.

The protocol described in [9] uses a distributed threshold ElGamal cryptosystem in order to create a distributed group key y among the meters by themselves. When the meters are requested to send their readings m_i, each of them generates a random integer z_i and encrypts and sends

$$E_y(g^{m_i+z_i}) = (c_i, d_i).$$

The substation then aggregates all messages together, obtaining $(c, d) = (\prod c_i, \prod d_i)$. The value c is sent back to all the meters which employ their private keys x_i to compute a partial decryption

$$T_i = c^{x_i} \cdot g^{z_i}$$

which is transmitted to the substation. The substation then computes

$$D = g^m = d \cdot (\prod T_i)^{-1}.$$

It finally solves an easy instance of the discrete logarithm problem, obtaining the sum of all the readings $m = \log_g D$.

The proposal [10] involves a trusted dealer that generates a set of random values $\{s_0, \ldots, s_n\}$ whose addition is zero. Then, the substation receives s_0 while the remaining values are secretly distributed among the meters. At time step t, each meter M_i, $1 \leq i \leq n$, encrypts its reading m_i by computing

$$c_i = g^{m_i} \cdot H(t)^{s_i},$$

and sends c_i to the substation which aggregates all the received ciphertexts as

$$V = H(t)^{s_0} \prod_{i=1}^{n} c_i,$$

so that the resulting value is $V = g^{\sum m_i}$. Finally, the added readings are obtained by solving the discrete logarithm of V to the base g.

This basic proposal is extended in [10] to provide differential privacy. The mentioned proposal is very lightweight and only requires unidirectional communications from the meters to the substation. Its drawback is the presence of a trusted dealer for key distribution. A similar approach which requires a trusted dealer that generates an RSA modulus is described in [11].

B. Trusted dealer avoidance

The aggregation-based proposal from [10] has been proven to be secure, and has very reduced computational and communication costs. Unfortunately, it requires a trusted dealer for key setup. As mentioned in [13], the requirement of a trusted dealer can be replaced with a TTP-free protocol permitting full-length cleartext homomorphic addition.

The proposals [8], [11] allow homomorphic full-length cleartext addition but require complicated protocols for a distributed generation of hard-to-factor modulus in a trusted dealer-free deployment.

On the other side, the proposal described in [9], which is based on a distributed ElGamal cryptosystem, offers a very simple distributed setup. Its drawback comes from the fact that the resulting aggregated cleartext is obtained after solving a discrete logarithm. Hence, it can only be employed for obtaining the addition of reduced length cleartexts.

C. Contribution and plan of this paper

The objective of this paper is the design of an efficient and easy to configure system able to be employed during the key establishment phase of [10] so as to avoid the presence of a trusted dealer. We propose a method which takes advantage of the simple setup provided by [9] while addressing its limitation of not being able to cope with full-length aggregated cleartexts.

The paper has been structured as follows. Section I has provided an overview to the privacy challenges to be addressed by smart metering systems and has also introduced the objective of our research. Next, our proposal is described in Section II. The security of the proposal is analized in Section III while Section IV proves its feasibility by means of experimental results. Finally, the paper is concluded in Section V.

II. Our proposal

Along the paper we assume the use of elliptic curve cryptography [14] which significantly reduces the size of the involved messages reducing also the computational cost.

The required trusted dealer in [10] will be avoided by employing [9] at the key establishment phase.

The system in [9] allows the transmission of short cleartexts by the smart meters, whose addition is obtained by the substation. We will assume the maximum bitlength of cleartexts to be 13 bits so that, in a neighborhood of up to 128 smart meters, the aggregated message can be obtained by the substation after solving a 20 bits discrete logarithm problem, which took less than half a second in all our experiments.

By using elliptic curve cryptography, the bitlength of the secret values s_i, $0 \leq i \leq n$, to be distributed during the key establishment in [10] can be reduced to 256 bits (down from the original 2048, even obtaining a slightly higher security level). We propose to split such secrets into twenty 13-bits fragments, and transmit the aggregation of those short fragments using [9] in such a way that the substation is able to obtain $\sum_{i=1}^{n} s_i$ by combining the received aggregated fragments without obtaining any information about the individual secrets. By setting s_0 to be $-\sum_{i=1}^{n} s_i$, we get that $\sum_{i=0}^{n} s_i = 0$.

A. Setup

The setup needs to be done before deploying the system.

- Choose an elliptic curve E defined over a prime field $\mathbb{F}_q$, and a point $P \in E(\mathbb{F}_q)$ with prime order p of at least 256 bits.
- Choose a hash function H that returns a point belonging to the subgroup of $E(\mathbb{F}_q)$ generated by P given an input t representing the current time step.

B. Key establishment

The key establishment phase has to be performed before the first execution of the protocol and each time that a smart meter is added or removed from the neighborhood. The secret s_i generated by each smart meter should be similar in size to p.

- Each smart meter M_i, $1 \leq i \leq n$, generates a random secret $s_i < p$ which is then bitwisely split into l chunks s_{ij} of at most 13 bits each such that

  $$s_i = (s_{il}|| \ldots ||s_{i2}||s_{i1}).$$

  It is required that $13 \cdot l$ is equal or bigger than the size (in bits) of p.
- The meters and the substation execute the protocol in [9] l times (which can be run in parallel). After each execution j, for $1 \leq j \leq l$, the substation obtains $\sum_{i=1}^{n} s_{ij}$. Then the substation computes

  $$s'_0 = \sum_{j=1}^{l} \left(2^{13 \cdot (j-1)} \cdot \sum_{i=1}^{n} s_{ij} \right) = \sum_{i=1}^{n} s_i,$$

  and finally sets its secret s_0 to be $s_0 = -s'_0 \pmod p$. Note that $\sum_{i=0}^{n} s_i = 0 \pmod p$ as required in [10].

C. Consumption transmission

In a smart metering neighborhood, the meters regularly communicate with their substation. At each time step t, every smart meter M_i sends its consumption reading m_i to the substation SSt. The time step may be simply represented as an integer that is incremented by one after each round.

The consumption readings at time step t are transmitted as follows:

- Each meter M_i (storing its secret s_i) transmits its reading m_i by computing an elliptic curve point C_i as
$$C_i = m_i \cdot P + s_i \cdot H(t), \quad (1)$$
which is then sent to the substation.
- The substation receives and aggregates all the points C_i by computing:
$$C = \sum_{i=1}^{n} C_i = m \cdot P + \sum_{i=1}^{n} s_i \cdot H(t),$$
with $m = \sum_{i=1}^{n} m_i$.
- Then, the substation computes:
$$D = C + s_0 \cdot H(t) = m \cdot P + s \cdot H(t) = m \cdot P,$$
since $s = \sum_{i=0}^{n} s_i = 0$.
- Finally, the substation computes the discrete logarithm of D to the base P to get the aggregated readings of all the meters:
$$m = \sum_{i=1}^{n} m_i = \log_P D.$$

Since the computed discrete logarithm is known to be short, it can be efficiently computed by means of Pollard's Lambda algorithm [15].

III. Security analysis

The security of [9], [10] holds on the assumed difficulty of solving the Computational Diffie-Hellman (CDH) problem. Given a multiplicative group G of prime order p with generator g, and given g^a and g^b (a and b are unknown), the computation of g^{ab} is assumed to be hard. The elliptic curve definition of that problem is straightforward by replacing the multiplicative notation with an additive one.

In [9], it is shown that obtaining a partial aggregation of the values sent by the smart meters is as hard as solving the CDH problem. Since we employ [9] during the key establishment phase, no information about the individual secrets s_i is leaked during that phase.

Regarding the transmission of consumption values, we can use [10, Theorem 1] adapted to elliptic curves which proves that the construction in Eq. 1 satisfies aggregator oblivious security (the aggregator is incapable of obtaining any information other that the desired sum of values) in the encrypt-once model.

IV. Experimental results

In this section we test the feasibility of our proposal through experimentation.

The most time-consuming part of the proposal are the computations of the discrete logarithms required both in 'key establishment' and 'consumption transmission'. We have first considered neighborhoods composed of 128 meters, each of them sending a 13 bits number as a cleartext so that the overall aggregated cleartext requires 7 additional bits. Hence, these discrete logarithms have a 20 bits solution.

By using 13 bits, a transmitted cleartext can accomodate the consumption reading of a client during a 30 minutes period. These figures were already proposed in [9].

We used a custom elliptic curve cryptography library and an implementation of Pollard's Lambda algorithm [15] for solving the discrete logarithm problem. Pollard's Lambda is a generic algorithm for solving the discrete logarithm that can be implemented on any cyclic group and, in particular, on the group of points of an elliptic curve as required in our work. Pollard's Lambda algorithm is efficient at solving discrete logarithms when the solution is known to be in a small interval $\{a, \ldots, b\}$. Its time complexity is $O(\sqrt{b-a})$.

The implementation was done in Java (version 1.8.0_92) and executed on a computer with an Intel® Core™ i5-4460 processor running at 3.2 GHz with 16 GB of RAM, using only one of the four cores.

We solved 1000 instances of the problem obtaining an average time of 0.1 seconds per computation. Note that due to the random nature of the algorithm, it can take much more (or much less) time in some instances. In our experiments, the maximum time never surpassed 0.5 seconds.

We also considered the possibility of assuming larger groups of meters (several buildings, entire streets, neighborhoods or towns). Doubling the size of the group requires one additional bit for representing the aggregated consumption. In order to test the feasibility over larger groups, we solved larger instances of the problem.

Meters	Bits	Average	Maximum
128	20	0.10 s	0.48 s
512	22	0.21 s	0.87 s
2048	24	0.42 s	1.48 s
8192	26	0.81 s	4.29 s
32768	28	1.57 s	7.08 s

TABLE I
AVERAGE AND MAXIMUM TIMES FOR SOLVING THE DISCRETE LOGARITHM PROBLEM OVER ELLIPTIC CURVES

Table I shows, for several group sizes, the bitlength of the discrete logarithm to be computed, and the average and maximum running times.

It can be seen that taking 4 times larger groups (which adds 2 bits to the logarithm) doubles the average running time. This is the expected behavior considering the time complexity of Pollard's Lambda algorithm. Due to its random nature, however, the maximum times are more difficult to predict, but our experiments show that they roughly double every two bits. Nevertheless, these extreme times are very unusual.

Our experiments show that larger groups with even 8192 meters are still affordable, since the average time takes less than a second and, in the worst case, it never takes more than 5 seconds. If larger groups were needed, we could employ a more powerful computer, or implement a multicore version of the algorithm.

The cost of the key establishment phase, which needs to be performed each time a smart meter is added or removed

from the group, is mostly determined by the computation of l discrete logarithms, each of size the bitlength of $2^{13} \cdot n$. The cost of the remaining computations is negligible. In our experiments, in the largest considered neighborhood (32768 meters), the execution of a key establishment phase would take around 30 seconds. Note, however, that a smart metering neighborhood is a rather static scenario so that such executions would be quite infrequent.

V. Conclusion

In this paper, we have proposed a system that allows to remove the need for a trusted third party during the key establishment phase of an existing privacy-preserving smart metering system. We also propose to implement it employing elliptic curve cryptography which allows the use of much shorter keys. The system is based on combining two existing proposals, which were already proven to be secure and private.

In the resulting system, the link between electricity consumption values and the customers is broken by homomorphically adding the readings. Since there is no trusted party, even assuming a corrupted substation and some malicious smart meters, the individual readings are still kept private: the only information such a corrupted coalition can obtain is the addition of the readings of the remaining honest meters. The resulting protocol does not require communication among the smart meters, and reading consumptions are transmitted by sending only one message per round. For the meters, the only relevant computation per round is an elliptic curve point multiplication by a full-length scalar. The substation is required to solve an easy instance of the discrete logarithm problem at each round. With an appropriate choice of parameters, this computation has been shown to be solvable in less than one second in a standard computer.

The elliptic curve point multiplication by a scalar performed by the meters, when using projective coordinates and an appropriate curve, can be achieved in general with 12 multiplications, 2 squares, 6 additions, and 1 doubling in the base field $\mathbb{F}_q$, so that only basic modular operations are needed.

Acknowledgments

Research of the authors was supported in part by the European Regional Development Fund of the European Union in the scope of the "Programa Operatiu FEDER de Catalunya 2014–2020" (project COMRDI16-1-0060), and by the Spanish Ministry of Science, Innovation and Universities (project MTM2017-83271-R). The Cryptography & Graphs Research Group of Universitat de Lleida has been recognized as a Consolidated Research Group by Generalitat de Catalunya (2017 SGR 1158).

References

[1] P.V. Barbosa, A. Brito, and H. Almeida: "A technique to provide differential privacy for appliance usage in smart metering", *Information Sciences*, vol. 370-371, pp. 355–367, 2016.

[2] C. Efthymiou and G. Kalogridis: "Smart grid privacy via anonymization of smart metering data", *Proc. of First IEEE International Conference on Smart Grid Communications*, pp. 238–243, 2010.

[3] S. Finster and I. Baumgart: "Pseudonymous smart metering without a trusted third party", *Proc. of 12th IEEE Intl. Conf. on Trust, Security and Privacy in Computing and Communications*, 2013.

[4] R. Petrlic: "A privacy-preserving concept for smart grids", *Sicherh. vernetzten Syst.*, vol. 18, B1-B14, 2010.

[5] M. Stegelmann and D. Kesdogan: "GridPriv: A smart metering architecture offering k-anonymity", *IEEE 11th Intl. Conf. on Trust, Security and Privacy in Computing and Communications*, pp. 419–426, 2012.

[6] J.M. Bohli, C. Sorge, and O. Ugus: "A privacy model for smart metering", *IEEE 1st Intl. Workshop on Smart Grid Communications*, 2010.

[7] F. García and B. Jacobs: "Privacy-friendly energy-metering via homomorphic encryption", *Proc. of 6th Intl. Conf. on Security and Trust Management*, LNCS vol. 6710, pp. 226–238, 2011.

[8] V. Rastogi and S. Nath: "Differentially private aggregation of distributed time-series with transformation and encryption", *Proc. of SIGMOD'10*, pp. 735–746, 2010.

[9] N. Busom, R. Petrlic, F. Sebé, C. Sorge, and M. Valls: "Efficient smart metering based on homomorphic encryption", *Computer Communications*, vol. 82, pp. 95–101, 2016.

[10] E. Shi, R. Chow, T.-H. H. Chan, D. Song, and E. Rieffel: "Privacy-preserving aggregation of time-series data", *Proc. of NDSS' 2011*, The Internet Society, 2011.

[11] M. Joye and B. Libert: "A scalable scheme for privacy-preserving aggregation of time-series data", *Proc. of FC'2013*, Springer–Verlag Berlin Heidelberg, LNCS vol. 7859, pp. 111–125, 2013.

[12] D. Dwork: "Differential privacy: a survey of results", *Proc. of TAMC'2008*, Springer–Verlag Berlin Heidelberg, LNCS vol. 4978, pp. 1–19, 2008.

[13] T. Jung and X. Li: "Collusion-tolerable privacy-preserving sum and product calculation without secure channel", *IEEE Transactions on Dependable and Secure Computing*, vol. 12, no. 1, pp. 45–57, 2015.

[14] D. Hankerson, A. Menezes, S. Vanstone: Guide to elliptic curve cryptography, Springer, New York, 2004.

[15] J.M. Pollard: "Monte Carlo methods for index computation (*mod p*)", *Mathematics of computation*, vol. 32, no. 143, pp. 918–924, 1978.

Identificación de polinomios primitivos sobre cuerpos extendidos mediante análisis de secuencias entrelazadas

Guillermo Cotrina Cuenca
Universidad de Málaga
Andalucía Tech
E.T.S.I.Telecomunicación
gcotrinacuenca@uma.es

Andrés Ortiz García
Universidad de Málaga
Andalucía Tech
E.T.S.I.Telecomunicación
aotriz@ic.uma.es

Alberto Peinado Domínguez
Universidad de Málaga
Andalucía Tech
E.T.S.I.Telecomunicación
apeinado@ic.uma.es

***Resumen*—In this article, a method to discriminate primitive polynomials over extended Galois Field is described. The method employs the generation of a small number of elements of the associated LFSR and its equivalent model based on interleaved binary sequences. The consideration of the sequence generated by the associated LFSR as an interleaving sequence allows one to move the analysis to the decimated binary sequences. In this way, one can improve the running time of the previous methods.**

***Index Terms*—Cryptography, primitive polynomials, extended Galois Fields, LFSR, symmetric cryptography, m-sequences.**

I. INTRODUCCIÓN

Los registros de desplazamiento realimentados linealmente (LFSR) constituyen la base de un gran número de sistemas criptográficos simétricos debido a su simplicidad, a la longitud y aleatoriedad de las secuencias que generan y a la posibilidad de predecirlas mediante modelos lineales [1], [2], [3]. A pesar de la eficiencia de su funcionamiento, en los últimos años se ha propuesto sustituir la utilización de los LFSR tradicionales, esto es, definidos sobre $GF(2)$ por LFSR definidos sobre cuerpos extendidos, con el fin de aumentar la velocidad a la que se genera la secuencia cifrante, pasando de 1 bit a un word en cada generación. Un ejemplo de ello se puede encontrar en el algoritmo de cifrado simétrico empleado en las redes móviles LTE [4], donde tales secuencias se generan de 32 en 32 bits. Este cambio, propiciado por la evolución en la longitud de palabra de los procesadores, se refleja en la utilización de LFSRs definidos en $GF(2^n)$, siendo n la longitud de palabra del procesador.

Como ocurre con los LFSR en $GF(2)$, para garantizar que las secuencias generadas satisfacen unos requisitos mínimos de seguridad, aleatoriedad y complejidad, es necesario, o al menos recomendable, que la realimentación aplicada esté caracterizada por un polinomio primitivo, lo que asegura una longitud de secuencia máxima. Si bien la búsqueda de polinomios primitivos en $GF(2)$ no es un problema, la búsqueda en un cuerpo extendido $GF(2^n)$ no es sencilla y requiere de algoritmos eficientes. La solución a este problema se ha planteado, como suele ser habitual desde dos perspectivas. Por una parte, se han propuesto métodos de construcción o generación de polinomios primitivos en $GF(p^n)$ a partir de otros polinomios, como el publicado por Cardell y Climent [5] en 2017, que utilizan polinomios de grado b y r definidos sobre $\mathbb{F}_q$ y $\mathbb{F}_{q^b}$, respectivamente. Por otra parte, se han desarrollado tests que permiten únicamente comprobar si un determinado polinomio es o no primitivo. En esta última aproximación es en la que se centra este trabajo, que tiene como antecendentes las propuestas de Wang, Zheng and Li [6] y J.A. Gordon [7] que basan su desarrollo en la generación de la secuencia en su totalidad, lo cual resulta en una implementación bastante lenta; y las más recientes de Michiel [8] en 2001 y Menezes [9] en 2004, que utilizan métodos basados en el producto matricial.

La propuesta que se presenta en este trabajo está basada en el tratamiento de las secuencias generadas por los LFSRs en $GF(2^n)$ como secuencias binarias entrelazadas, lo que permite realizar cálculos mucho más sencillos en $GF(2)$, reduciendo así el tiempo de computación.

En la sección II se presentan la notación y los fundamentos matemáticos sobre los que se desarrolla la propuesta. La sección III describe el funcionamiento y comportamiento de los LFSRs definidos en el cuerpo extendido $GF(2^n)$. El modelo equivalente basado en secuencias entrelazadas se describe en la sección IV. Todo ello permite presentar en la sección V el test de polinomios primitivos junto con un ejemplo de su implementación. Finalmente, las secciones VI y VII muestran los resultados del análisis del tiempo de ejecución y las conclusiones.

II. NOTACIÓN Y FUNDAMENTOS

Llamaremos cuerpo de Galois $GF(q)$ de orden q a un cuerpo finito de $q = p^n$ elementos, donde p es primo y $n \in \mathbb{Z}^+$. Cualquier cuerpo finito de $q = p^n$ elementos es isomorfo a $GF(q)$ [10], por lo que, si denotamos un cuerpo mediante $\mathbb{F}$, denotaremos a cualquier cuerpo finito de q elementos o cuerpo de Galois de orden q como $\mathbb{F}_q$ o $GF(q)$,indistintamente.

De acuerdo con esta notación y definiciones un polinomio $f \in \mathbb{F}_q[x]$ de grado $k \geq 1$ se dirá que es primitivo sobre $\mathbb{F}_q$ si es un polinomio minimal sobre $\mathbb{F}_q$ de un elemento primitivo de $\mathbb{F}_{q^k}$. Equivalentemente, un polinomio $f \in \mathbb{F}_q[x]$ de grado k es primitivo si $min_{n\in\mathbb{N}}\{n : f(x)|x^n - 1\} = q^k - 1$.

Como ya es conocido, este valor, $q^k - 1$, supone la máxima longitud de las secuencias generadas por recurrencias polinómicas, que se alcanzan si y sólamente si el polinomio

característico $f \in \mathbb{F}[x]$ en el que $f(0) \neq 0$, es un polinomio primitivo sobre $\mathbb{F}_q$.

Por último, y con el fin de poder realizar operaciones en $\mathbb{F}_{q^m}$, se utilizará un polinomio primitivo $q(x)$ de grado m para poder construir $\mathbb{F}_{q^m}$, el cuerpo de Galois de q^m elementos.

III. LFSR EN EL CUERPO EXTENDIDO

Un LFSR es un registro compuesto por n celdas de memoria cuyo contenido es desplazado a la celda vecina, siempre en la misma dirección, y actualizado mediante una función de realimentación lineal como se ilustra en la figura 1. Todo el proceso está controlado por un reloj único, de manera que el desplazamiento de todas las celdas y la actualización de la celda más externa se realizan simultáneamente. El estado actual $(s_{j-1}, s_{j-2}, \ldots, s_{j-n})$, del LFSR junto con los coeficientes de realimentación c_j determinan el valor del bit s_j de la secuencia mediante la siguiente relación

$$s_j = c_1 s_{j-1} + c_2 s_{j-2} + \ldots + c_n s_{j-n}, \tag{1}$$

De este modo, la secuencia pseudoaleatoria generada por el LFSR queda constituida por los sucesivos valores de cualquiera de las celdas, típicamente s_{j-n} como se observa en la figura 1. Habitualmente los LFSR se definen en $GF(2)$, lo que determina que tanto el contenido de las celdas como los elementos de la secuencia son binarios y, en consecuencia, cada elemento generado proporciona únicamente un bit a la secuencia final. Con el fin de generar un mayor número de bits a cada pulso de reloj, se definen los LFSR directamente sobre $GF(2^n)$, haciendo coincidir en la medida de lo posible la n con la longitud de palabra del procesador utilizado. Este cambio supone una mejora en la eficiencia puesto que cada nuevo elemento generado aporta n bits a la secuencia. Además, dado que los procesadores están diseñados para trabajar con operandos de n bits, la eficiencia de las operaciones aumenta al eliminar la necesidad de realizar operaciones binarias, que suponen mayor complejidad para este tipo de procesadores.

Considerando que el LFSR trabaja en $GF(2^m)$, y que dicho cuerpo ha sido generado por un polinomio primitivo $U(x)$ de grado m, cada elemento s_j del cuerpo queda representado por una $m-$ upla $(a_0, a_1, \cdots, a_{m-1})$ con $a_i \in GF(2)\ \forall i = \{1, 2, \cdots, m-1\}$ que corresponden a los coeficientes del polinomio:

$$n(x) = (a_0, a_1, \cdots, a_{n-1}) \cdot \left(1, x, \cdots, x^{n-1}\right) \tag{2}$$

donde $\cdot$ representa el producto escalar de vectores y cada uno de los coeficientes se obtiene módulo 2 y módulo el polinomio $q(x)$. De tal forma que se obtiene una operación interna que dota de estructura de cuerpo a $GF(2^m)$.

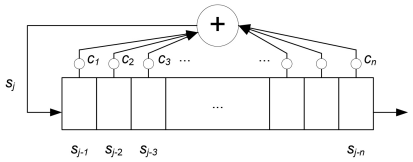


Figura 1. Esquema gerneral de un LFSR

Por otra parte, los coeficientes de realimentación $c_i \in GF(2^m)$ determinan el polinomio $p(x) = c_n x^n + \cdots + c_2 x^2 + c_1 x + 1$, que es el que permite estudiar las características de las secuencias sin necesidad de ser generadas. De este modo, si el polinomio $p(x)$ es primitivo, el período de las secuencias será máximo, de valor $2^n - 1$ sobre $GF(2^m)$, lo que supone un total de $2^{nm} - 1$ bits, con independencia de la semilla escogida, a excepción de la semilla nula que se mantiene como elemento invariante debido a la linealidad de la realimentación. Es decir, la recurrencia lineal definida por el polinomio primitivo contiene a todos los elementos no nulos.

IV. MODELO EQUIVALENTE DE SECUENCIAS ENTRELAZADAS

En [12] y en [13], tanto John J . Komo y Maurice como S. Lam Gong presenta las secuencias entrelazadas mediante la siguiente definición.

Definición IV.1. Sea A una secuencia sobre $GF(q^n)$, sea $d \in \mathbb{Z}^+$, se define la secuencia diezmada A sobre d y la representaremos A^d como, $A^d(t) = A(td)\ \forall t \geq 0$.

Definición IV.2. Sea $f(x)$ un polinomio sobre $GF(q)$ de grado n con $f(0) = 0$, y sea l un entero positivo. Para cualquier secuencia $u = (u(0), u(1), u(2), ...) = (u(t))$ sobre $GF(q)$, si las secuencias diezmadas $u_j = (u(tl + j)), t \geq 0$ son generadas por $f(x)$ para todo j, entonces u se denomina secuencia entrelazada sobre $GF(q)$ de tamaño l asociada a $f(x)$.

Al mismo tiempo, en [12], se presenta el siguiente lema.

Lema 1. *Sea u una secuencia entrelazada definida sobre $GF(q)$ de tamaño l asociada a $f(x)$, y sea $h(x)$ el polinomio minimal de u sobre $GF(q)$. Entonces, se cumple que*

1. $h(x) | f(x^l)$
2. $T(u) | per(f)^l$*, donde $T(u)$ es el período de la secuencia u y $per(f)$ el exponente del polinomio f.*

Por otra parte, en [13] se establece la relación entre las m-secuencias generadas sobre $GF(q^m)$ y las m-secuencias generadas sobre $GF(q)$, de tal modo que éstas últimas se pueden obtener como secuencias diezmadas de la primera, o recíprocamente, la primera es una secuencia entrelazada compuesta de m-secuencias sobre $GF(q)$. Esta relación, es la que permite analizar la secuencia en el cuerpo extendido a partir de sus componentes diezmadas en el cuerpo base.

V. IDENTIFICACIÓN DE POLINOMIOS PRIMITIVOS

Consideremos un polinomio $p(x)$ de grado m sobre $GF(2^n)$, del que se quiere comprobar si es primitivo y un polinomio $U(x)$ de grado n con coeficientes binarios, utilizado para construir $GF(2^n)$. Los siguientes teoremas proporcionan la base del método que se describe a continuación.

Teorema 2. *Dadas la secuencia $\{A_{n,m}\}_{n,m\in\mathbb{N}} = \{\{a_{1,1}, a_{1,2}, \cdots, a_{1,n}\}, \{a_{2,1}, a_{2,2}, \cdots, a_{2,n}\}, \cdots\}$ generada por el polinomio $f \in \mathbb{F}_{q^n}[x]$. Sea $a^1 = \{a_{1,1}, a_{2,1}, \cdots\}$, $a^2 = \{a_{1,2}, a_{2,2}, \cdots\}$, $\cdots$, $a^q = \{a_{1,q}, a_{2,q}, \cdots\}$, el conjunto de secuencias diezmadas, todas sobre $GF(2)$. En estas condiciones el polinomio f es primitivo si y solamente si las secuencias diezmadas $\{a^1, a^2, \cdots, a^q\}$ son secuencias maximales sobre $GF(2)$ y están todas generadas por el mismo polinomio. Lo que es más, en estas condiciones si*

el grado de f es m entonces este único polinomio que genera las secuencias $\{a^1, a^2, \cdots, a^q\}$ tiene grado $m \cdot n$ y las secuencias están generadas son maximales y tienen complejidad lineal $m \cdot n$

Proof: Si aplicamos el **Teorema 1** dado en [12] podemos afirmar que dado un polinomio primitivo de grado $m \cdot n$ sobre $GF(q)$ donde α es una raíz que genera una secuencia maximal c y donde f_1 es un polinomio primitivo que divide a $g(x)$ sobre $GF(q^n)$ y que genera una secuencia maximal d. Por otro lado sea $h(x)$ un polinomio de grado m sobre $GF(q)$ con raíz γ donde $\gamma = \alpha^{z\omega}$ y pertenece a la base $\{1, \gamma, \gamma^2, \cdots, \gamma^{m-1}\}$ para expresar $GF(q^m)$ sobre $GF(q)$, dado $z = \left(\frac{q^{mn}-1}{q^m-1}\right)$ y donde ω y $q^m - 1$ son relativamente primos, entonces se verifica que para algún e_0

$$d = \sum_{j=0}^{m-1} \gamma^j T^{e_0 + z k_j \omega_c} \tag{3}$$

donde T^{i_C} indica un desplazamiento a la izquierda de la secuencia c de i elementos.

Del mismo modo como consecuencia inmediata el **teorema 1** dado en [12] podemos utilizar los **corolario 1 y 2** dado en el mismo artículo [12], todos los elementos de los registros de desplazamiento de las secuencias maximales de longitd N sobre $GF(q^m)$ se pueden obtener mediante el diezmado de esta secuencia maximal, dada una base de polinomios de $GF(q^m)$ sobre $GF(q)$. ■

Teorema 3. *En las mismas condiciones del **teorema 2**, supongamos que el polinomio obtenido es $a^1(x) = a_0^1 + a_1^1 x + a_2^1 x^2 + \cdots a_{m\cdot n}^1 x^{m\cdot n}$. Consideremos la secuencia plana generada como la concatenación de las secuencias $\{A_{n,m}\}_{n,m\in\mathbb{N}} = \{\{a_{1,1}, a_{1,2}, \cdots, a_{1,n}\}, \{a_{2,1}, a_{2,2}, \cdots, a_{2,n}\}, \cdots\}$. Entonces esta secuencia plana tienen complejidad lineal $m \cdot n^2$ y está generada por el polinomio $A(X) = a_0^1 + a_1^1\ x^{1\cdot n} + a_2^1\ x^{2\cdot n} + \cdots a_{m\cdot n^2}^1\ x^{m\cdot n^2}$.*

Proof: John J . Komo and Maurice S. Lam establecen en [13] establecen la relación existente entre las secuencias maximales sobre $GF(q^n)$ en función de las secuencias maximales de los polinomios sobre $GF(q)$ ■

Vamos a ilustrar de manera resumida cómo se efectúa el método. Lo primero que realizaremos será generar $2\cdot m\cdot n+1$ elementos utilizando el polinomio que se desea analizar. Una vez obtenidos estos elementos pararemos al modelo equivalente de secuencias entrelazadas diezmando adecuadamente la secuencia. Posteriormente se analizan las secuencias entrelazadas y se les aplicará el algoritmo de Berlekamp-Massey. Finalmente se analizará si estas cumplen las condiciones pedidas en los **teoremas 2 y 3**.

Ilustremos el modelo paso a paso.

Test de primitividad. Input: polinomio $p(x)$ de grado m sobre $GF(2^n)$. Output: Primitivo/No primitivo.

- *Paso 1*. Se construye un LFSR sobre $GF(2^n)$ con m celdas, utilizando $p(x)$ como realimentación y una semilla $A_0 = \{(a_0^1, a_1^1, \cdots, a_{n-1}^1), \cdots, (a_0^m, \cdots, a_{n-1}^m)\}$, donde no todos los elementos son nulos. Se ha de tener en cuenta que no se pierde generalidad, por lo que el método es independiente de la semilla escogida.
- *Paso 2*. Se generan $2\cdot m\cdot n+1$ elementos de la secuencia, lo que permite disponer de $2mn^2$ bits.
- *Paso 3*. Se obtienen las secuencias entrelazadas $\{B^i\}$.
- *Paso 4*. Se analizan las secuencias binarias entrelezadas para calcular su complejidad lineal y obtener el polinomio de realimentación $\{p_1, p_2, \cdots p_n\}$ con coeficientes binarios de cada una de ellas, aplicando, por ejemplo, el algoritmo de Berlekamp-Massey.
- *Paso 5*. El polinomio p será primitivo si y solamente si $p_i = p_j$, $\forall i, j$ y además el grado de p_i es $m \cdot n$, $\forall i, j$.

A continuación, se presenta un ejemplo de la ejecución del método aplicado a un polinomio $p(x)$ de grado 6 con coeficientes en $GF(2^4)$, esto es, $p(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_6 x^6$, con $c_i \in GF(2^4)$. En este caso, el cuerpo queda definido por el polinomio $U(x) = 1 + x + x^4$ sobre $GF(2)$. Utilizando la representación en m- uplas de los elementos del cuerpo, podemos considerar como ejemplo los siguientes coeficientes de $p(x)$:

$$\begin{array}{c} c_0 = \{0,1,0,1\} \\ c_1 = \{1,1,0,1\} \\ c_2 = \{0,0,1,1\} \\ c_3 = \{1,1,0,1\} \\ c_4 = \{1,0,0,1\} \\ c_5 = \{0,0,0,1\} \end{array} \tag{4}$$

o de manera equivalente,

$$\begin{array}{c} c_0 = 1 + \alpha \\ c_1 = 1 + \alpha + \alpha^3 \\ c_2 = \alpha^2 + \alpha^3 \\ c_3 = 1 + \alpha + \alpha^3 \\ c_4 = 1 + \alpha^3 \\ c_5 = \alpha^3 \end{array} \tag{5}$$

donde α es una raiz del polinomio $U(x)$ que define el cuerpo. En consecuencia, se tiene que $\alpha^4 = \alpha + 1$, quedando todos los elementos del cuerpo definidos del siguiente modo.

$$\begin{array}{l} 1 \\ \alpha \\ \alpha^2 \\ \alpha^4 = \alpha + 1 \\ \alpha^5 = \alpha^2 + \alpha \\ \alpha^6 = \alpha^3 + \alpha^2 \\ \alpha^7 = \alpha^4 + \alpha^3 = \alpha^3 + \alpha + 1 \\ \alpha^8 = \alpha^4 + \alpha^2 \\ \alpha^9 = \alpha^3 + \alpha \\ \alpha^{10} = \alpha^4 + \alpha^2 = \alpha^2 + \alpha + 1 \\ \alpha^{11} = \alpha^3 + \alpha^2 + \alpha \\ \alpha^{12} = \alpha^4 + \alpha^3 + \alpha^2 = \alpha^3 + \alpha^2 + \alpha + 1 \\ \alpha^{13} = \alpha^4 + \alpha^3 + \alpha^2 + \alpha = \alpha^3 + \alpha^2 + 1 \\ \alpha^{14} = \alpha^4 + \alpha^3 + \alpha = \alpha^3 + 1 \\ \alpha^{15} = \alpha^4 + \alpha = 1 \end{array} \tag{6}$$

Definido el LFSR, seleccionamos la semilla

$$\begin{array}{c} \text{Semilla} = (\{1,0,0,1\}, \{0,1,0,0\}, \{0,1,0,1\}, \\ \{1,0,0,1\}, \{0,1,0,1\}, \{0,0,0,1\}) \end{array} \tag{7}$$

A continuación, se generan $2 \cdot 6 \cdot 4 + 1$ elementos que se corresponden con las filas de la siguiente matriz

$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \tag{8}$$

Las columnas de esta matriz conforman las secuencias entrelazadas asociadas a este registro de desplazamiento. El análisis de estas secuencias, mediante el algoritmo de Berlekamp–Massey da como resultado el siguiente polinomio de realimentación para todas las secuencias.

$$r(x) = x^{24} + x^{23} + x^{21} + x^{19} + x^{18} + x^{17} + x^{13} + \\ + x^{11} + x^{10} + x^8 + x^6 + x^5 + x^3 + x + 1 \tag{9}$$

Además, como se puede observar, el polinomio es de grado $6 \cdot 4 = 24$, con lo que se concluye que el polinomio $p(x)$ es primitivo en $GF(2^4)$.

VI. TIEMPOS DE EJECUCIÓN

Para la realización de las pruebas se ha utilizado el software Mathematica en su versión 10,0,0,0 sobre una plataforma de Microsoft Windows de $64-$bits. Se han programado cada una de las funciones realizadas al igual que cada uno de los bucles en el proceso de iteración. La máquina sobre la cual se han ejecutado los procesos posee un procesador Intel(R) Core (TM) $i7-4510U$ CPU @ 2,00 Ghz con una memoria RAM de $16,0$ GB y un sistema operativo Windows 10 de 64 bits.

Para realizar la medida del tiempo de ejecución se ha tenido en cuenta la temperatura del procesador y el control de los procesos que en ese momento se estaban ejecutándo, con el fin de reproducir las mismas condiciones en cada experimento. Las siguientes figuras muestras la evolución comparada de los tiempos de ejecución en función del grado del polinomio, considerando las extensiones $n = 8$ y $n = 16$, en las que se puede apreciar la mejora del método propuesto.

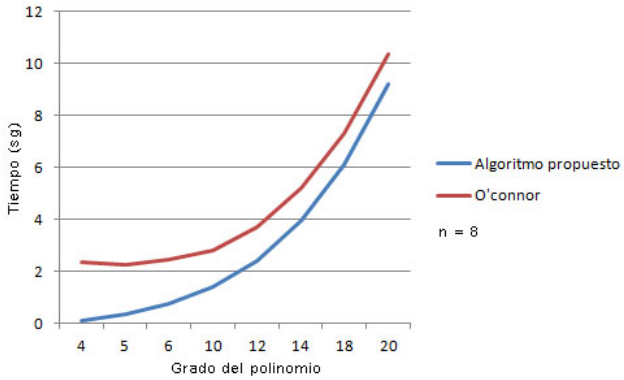


Figura 2. Tiempos de ejecución en $GF(2^8)$

Para poder comprobar si el método presentado mejoraba los resultados anteriores se ha realizado un estudio comparativo con el método presentado por [8]. Es evidente que para

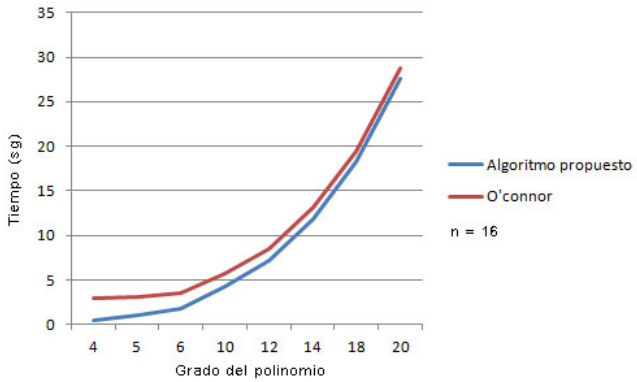


Figura 3. Tiempos de ejecución en $GF(2^{16})$

ambos métodos el número de operaciones depende del grado de los polinomios y del cardinal utilizado para la extensión del cuerpo finito. La principal ventaja presentada en nuestra propuesta es que sólo se han de generar $2 \cdot m \cdot n + 1$ bits mientras que en las propuestas que existen en la literatura se ha de generar la secuencia entera o de manera alternativa se han de efectuar tantas divisiones como elementos tenga la secuencia maximal sobre el cuerpo extendido. Este proceso hace que la evolución de los tiempos de ejecución para un determinado cardinal del cuerpo finito en nuestro método sean lineales en lugar de geométricos/exponenciales.

VII. CONCLUSIONES

En el presente trabajo se ha descrito un método computacional para realizar un test que permite identificar si un polinomio con coeficientes en $GF(2^n)$ es primitivo sobre dicho cuerpo. La implementación del método analiza las secuencias binarias diezmadas a partir de la secuencia global generada por un LFSR definido sobre el cuerpo extendido, aprovechando la relación que existe entre ambos tipos de m-secuencias, que permite analizar la secuencia en el cuerpo extendido como secuencia entrelazada. El método propuesto reduce el tiempo de ejecución con respecto a las propuestas existentes.

AGRADECIMIENTOS

Este trabajo ha sido financiado parcialmente por el Campus de Excelencia Internacional Andalucía Tech de la Universidad de Málaga y por el Ministerio de Economía, Industria y Competitividad (MINECO), la Agencia Estatal de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER, UE), a través del proyecto COPCIS, referencia TIN2017-84844-C2-1-R.

REFERENCIAS

[1] S. Babbage, M. Dodd, The MICKEY Stream Ciphers, LNCS 4986, (2008) pp. 191– 209, Springer-Verlag.
[2] S. Kiyomoto, J. Kurihara, The RAKAPOSHI Stream Cipher. Information and Communications Security, LNCS 5927[C], 2009, pp. 32–46.
[3] S. Kiyomoto, T. Tanaka, K. Sakurai, K2: A stream cipher algorithm using dynamic feedback control. SECRYPT, pp. 204–213. INSTICC Press (2007).
[4] ETSI, SNOW 3G Specification, Specification of the 3GPP Confidentiality and Integrity Algorithms UEA2 & UIA2, 2006.

[5] Sara D. Cardell and Joan-Josep Climent (2017) A construction of primitive polynomials over finite fields, Linear and Multilinear Algebra, 65:12, 2424-2431, DOI: 10.1080/03081087.2016.1275507
[6] Jiantao Wang, Dong Zheng, and Qiang Li, School of Information Security Engineering, Shanghai Jiao Tong University, Shanghai 200240, China National Engineering Laboratory for Wireless Security, Xi'an University of Posts and Telecommunications, Xi'an 710121, Shaanxi Province, China
[7] J.A. Gordon, Very simple method to find the minimum polynomial of an arbitrary nonzero element of a finite field, Electronic Letters, vol. 12, pp. 663–664, 1976
[8] Sean E. O'Connor Hazewinkel, Michiel, Primitive Polynomial Testing in Mathematica, ed. (2001) [1994], Berlekamp-Massey algorithm, Encyclopedia of Mathematics, Springer Science+Business Media B.V. / Kluwer Academic Publishers, ISBN 978-1-55608-010-4
[9] O. Ahmadi and A. Menezes, On the number of trace-one elements in polynomial bases for $GF(2^m)$, Designs, Codes and Cryptography, vol. 37, pp. 493–507, 2005.
[10] Lidl R, Niederreiter H. Introduction to finite fields and their applications. New York (NY): Cambridge University Press; 1986.
[11] Golomb, S.W., Shift-Register Sequences, *Aegean Park Press*, Laguna Hill, California, 1982.
[12] G. Gong, Theory and applications of q-ary interleaved sequences, IEEE Trans. Informat. Theory, 41 (1995) 400-411.
[13] J.J. Komo, M.S. Lam, Primitive Polynomials and M-Sequences Over $GF(q^m)$, IEEE TRANSACTIONS ON INFORMATION THEORY, VOL 39 (2), pp.643-647, 1992.

Criptografía Basada en Identidad aplicada a una MANET formada por robots

Jonay Suárez-Armas, Alexandra Rivero-García, Pino Caballero-Gil y Cándido Caballero-Gil
Departamento de Ingeniería Informática y Sistemas, Universidad de La Laguna
Tenerife, España
{jsuarear,ariverog,pcaballe,ccabgil}@ull.edu.es

***Resumen*—En situaciones en las que las comunicaciones pueden verse comprometidas, las redes MANET juegan un papel importante, las cuales permiten definir un punto de comunicaciones estableciendo comunicaciones descentralizadas entre los nodos de la red. En este trabajo se propone la implementación de una MANET basada en robots, de modo que cada nodo de la red sea un robot que se puede mover remotamente gracias al uso de streaming de audio y vídeo. Para facilitar el movimiento de los robots, estos son capaces de realizar seguimiento de objetos, de modo que controlando un robot el resto sean capaces de seguirlo. Para proteger las comunicaciones entre los nodos de la red se utiliza criptografía basada en identidad, utilizando un esquema que permite firmar y cifrar al mismo tiempo.**

***Index Terms*—MANET, Robótica, Criptografía basada en Identidad**

I. Introducción

En un mundo en el que la tecnología se ha vuelto esencial, tanto en el ámbito empresarial como en el personal, las comunicaciones son imprescindibles. Debido a la centralización de la información y al almacenamiento de los datos en la nube, un corte en las comunicaciones puede provocar el caos. Si en vez de producirse un corte en las comunicaciones de una empresa, se produce en las comunicaciones de un núcleo poblacional las consecuencias pueden ser mucho más graves, sobre todo si se debe a algún tipo de catástrofe.

Cuando se da una situación de este tipo se trata de buscar una solución, estableciendo un punto de comunicaciones temporal, que permita al menos, realizar las comunicaciones básicas para mantener el control de la situación.

Para conseguir establecer un punto de comunicaciones en un lugar donde se han roto las comunicaciones o se desee establecer un punto de comunicaciones temporal, en este documento se propone una MANET (Mobile Ad Hoc Network) basada en el uso de robots, la cual podría adoptar las siglas RANET (Robotic Ad Hoc Network). En dicha red, los robots son los encargados de formar una red multisalto [1] para finalmente dotar de un punto de comunicaciones al lugar, o los lugares, donde se establezcan los nodos (robots) finales de la red.

La robótica ha ido adentrándose cada vez más en los hogares, hasta el punto de que tenemos robots de cocina [2], o robots que limpian el piso [3] por nosotros. Estos últimos, los robots aspiradores, poseen diferentes ventajas como la posibilidad de volver automáticamente a la estación de carga de batería cuando es necesario, por lo que están siendo usados para otras tareas que no están relacionadas con la limpieza. Por ello, la propuesta de este trabajo se basa en el uso de robots aspiradores como nodos de la red.

La seguridad es un componente esencial del sistema propuesto. Se ha trabajado en la forma de autenticación de los diferentes nodos de la red, verificando que cada uno de ellos están autorizados por la misma. También se propone una forma de cifrado de los datos, garantizando que las comunicaciones sean seguras y la información esté protegida desde el primer nodo hasta el último. Para conseguir confidencialidad, autenticidad e integridad en las comunicaciones, tanto entre pares como entre la estación base y los robots, se usa un modelo que consigue en el mismo paso cifrar y firmar la información, conocido en inglés como IDentity-Based Signcryption (IDBS) [4].

En el siguiente apartado se presentan una serie de trabajos relacionados con la propuesta. La sección III explica el sistema propuesto entrando en más detalle, así como los elementos de los nodos de la red. La sección IV muestra el esquema de seguridad usado para proteger el sistema. Finalmente, el documento finaliza con una breve conclusión.

II. Trabajos relacionados

La aparición del Internet de las cosas [5], en inglés Internet of Things (IoT), puede tener una gran importancia en las comunicaciones descentralizadas. Cada vez hay más dispositivos conectados, y puede ser interesante que entre ellos formen redes de sensores. Muchas veces, este tipo de redes puede ser la solución para enviar datos a la nube cuando el dispositivo no dispone directamente de una conexión a Internet, pero saltando entre dispositivos es capaz de enviar la información a uno que sí dispone de conexión a Internet.

En IoT, diferentes trabajo se basan en el uso de Raspberry Pi para la recogida de datos mediante sensores que forman redes para transmitir la información entre ellos [6]. Por ejemplo, se pueden montar redes MANET utilizando drones que incorporan una Rapsberry Pi para poder establecer comunicaciones entre ellos [7]. Los microcontroladores Arduino son otro tipo de dispositivo que se usa para formar redes de sensores, como en [8] que se utilizan estos dispositivos formando una red de sensores para la monitorización de temperatura y así ayudar a los sistemas de aire acondicionado a trabajar de una manera más eficiente.

El aumento de la capacidad de cómputo de los smartphones de hoy en día también ha hecho que estos se utilicen para crear redes descentralizadas. La tecnología Bluetooth es una de las elegidas para formar redes MANET entre dispositivos móviles [9]. Estos también son utilizados para crear VANETs que permiten el intercambio de información entre vehículos en carretera [10], sin tener que realizar una comunicación a través de la nube. Además, la capacidad de los smartphones

para obtener datos mediante los sensores que incorporan, hace que estos dispositivos puedan ser montados sobre otros para desplazarlos por diferentes lugares y poder realizar mediciones [11], en ocasiones formando una MANET para poder intercambiar algún tipo de información. Por ejemplo, se pueden montar redes MANET utilizando drones que incorporan una Rapsberry Pi para poder establecer comunicaciones entre ellos.

En las redes descentralizadas es muy importante tener en cuenta la seguridad. Se debe estudiar la forma de autenticación en la misma así como la protección de los datos que viajan entre los diferentes nodos de la red. Algunos autores proponen modelos para la autenticación y autoconfiguración de nodos en redes MANET [12], e incluso algún trabajo para la prevención de ataques de impersonación utilizando doble factor de autenticación [13]. Aparte de esto, en [14] se estudian diferentes protocolos criptográficos a nivel de consumo de energía para el uso en MANETs, concluyendo que el estándar AES es el que mejor rendimiento produce.

III. Sistema propuesto

El sistema que se propone es una MANET en la que los diferentes nodos de la red son robots. Con ello, se pretende situar cada nodo en el lugar necesario para poder formar una red en la que se establezca un punto final de comunicaciones. Esto da la posibilidad de tener una conexión de voz y datos en un lugar incomunicado por alguna razón, y en el cual es necesario tener comunicaciones. El uso de robots como nodos de la red permite desplazar remotamente cada nodo hasta la ubicación deseada, lo cual puede ser importante si se trata de una zona peligrosa para que las personas transporten los equipos.

Para empezar a desplegar los robots, primero se debe estudiar la zona y calcular el número de nodos necesarios para extender la red. Dicho cálculo se hace de acuerdo al rango máximo de la tecnología inalámbrica usada para la conexión entre nodos, en este caso Wi-Fi [15], la cual cuenta con un rango máximo de 100 metros (ver Fig. 1).

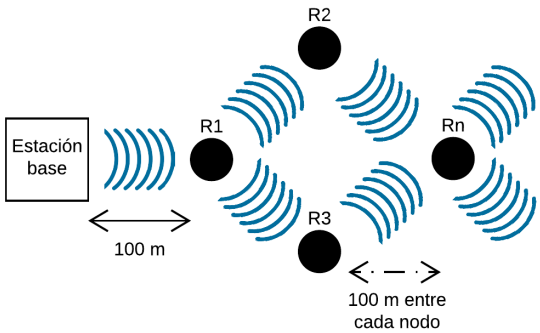


Figura 1. Esquema de la red.

El sistema propuesto se compone de una estación base desde la cual se controla todo el sistema y los distintos nodos. Además, reúne varias funcionalidades: control remoto de robots, streaming de audio y vídeo, seguimiento de objetos y despliegue de redes inalámbricas.

En primer lugar, el control remoto de robots permite mover remotamente cada nodo de la red hasta el punto que previamente se ha decidido. Desde la estación base es posible seleccionar uno de los robot y enviarle órdenes de movimiento en tiempo real.

En segundo lugar está el streaming [16] de audio y vídeo. Para poder mover los robots de forma remota es necesario saber el punto en el que se encuentran en cada momento, por lo que esta funcionalidad es indispensable. Gracias al streaming es posible visualizar el entorno a la vez que se controlan los robots desde la estación base.

En tercer lugar tenemos el seguimiento de objetos [17], usado para la parte más interesante de la propuesta. Dirigir cada robot hasta la zona deseada puede llegar a ser una tarea tediosa, por lo que se propone que controlando solamente un robot el resto sean capaces de seguirlo de forma autónoma. Para ello se selecciona un robot en modo maestro, y el resto en modo esclavo. Cada robot tiene en la parte trasera un código QR que lo identifica en la red. Cada robot esclavo tiene que hacer reconocimiento de imágenes, buscando algún código QR en las imágenes capturadas por la cámara. Cuando se produce la detección de un código de este tipo, el robot comprueba si se corresponde con el robot que debe seguir. Una vez que un robot esclavo detecta el código QR del robot que debe seguir, este realiza los cálculos para realizar los movimientos adecuados, ayudándose de la posición del QR en las imágenes y de un sensor de ultrasonido que ayuda a mantener la distancia (ver Fig. 2).

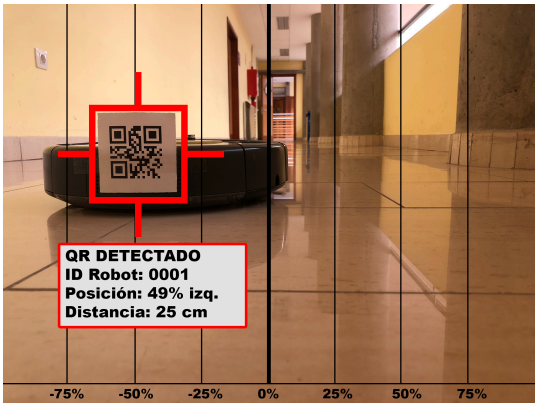


Figura 2. Seguimiento de QR.

La última funcionalidad es la encargada de desplegar una red inalámbrica en cada nodo de la red. Cada robot recibe una señal Wi-Fi de su predecesor, de manera que pueda comunicarse con la estación base saltando entre nodos. A su vez, es necesario que los nodos despliegen una nueva red Wi-Fi de modo que el siguiente nodo pueda conectarse a él para mantener la cadena de comunicaciones o para dar conexión a dispositivos cliente si se trata de un nodo final.

Cada nodo es construido sobre un robot aspirador iRobot Roomba conectado a una Raspberry Pi, la cual se puede considerar como el núcleo del nodo. En la Fig. 3 se muestra el esquema de conexionado de los diferentes elementos que componen cada nodo con la Raspberry Pi.

IV. Seguridad

La seguridad de las comunicaciones es un punto esencial del sistema. Por ello, se usa un esquema de cifrado y firma basado en identidad, en inglés ID-Based Signcryption (IBSC). Este tipo de criptosistema es una combinación de un cifrado

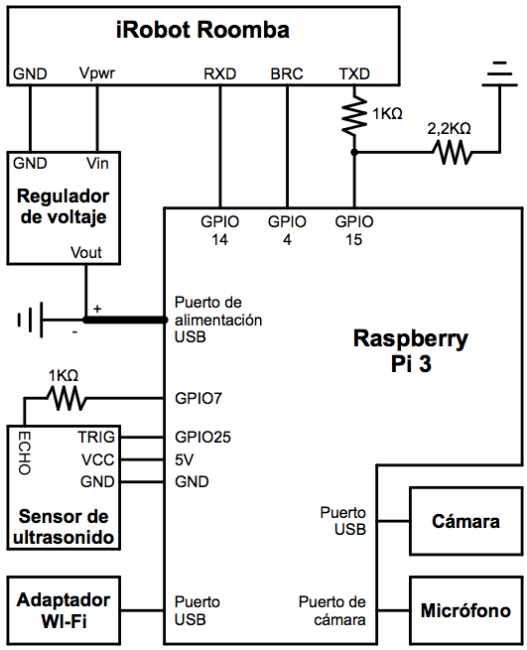


Figura 3. Esquema de conexionado.

basado en identidad, en inglés ID-Based Encryption (IBE), y de la firma basada en identidad, en inglés ID-Based Signature (IBS), donde todos los mensajes compartidos son cifrados y firmados mediante el uso de identificadores públicos. El esquema propuesto, basado en [4], ofrece la ventaja de simplificar la gestión de las claves al no tener que definir una infraestructura de clave pública. A parte de esto, este esquema es bastante flexible y es eficiente en consumo de memoria.

En la RANET propuesta, la identificación de los nodos es fundamental. Cada robot tiene un código QR en la parte trasera que lo identifica con un ID público único, el cual es usado para el IBSC.

Antes que nada, se describen las herramientas matemáticas usadas. Consideremos que $(G,+)$ y $(V,\cdot)$ son dos grupos cíclicos del mismo orden primitivo q. P es generador de G y existe un mapa bilineal de emparejamiento $\hat{e} : G \times G \to V$ que satisface las siguientes condiciones:

- Bilineal: $\forall P, Q \in G$ y $\forall a, b \in \mathbb{Z}$, $\hat{e}(aP, bQ) = \hat{e}(P, Q)^{ab}$
- No degenerado: $\exists P_1, P_2 \in G$ que $\hat{e}(P_1, P_2) \neq 1$. Esto significa que si P es generador de G, entonces $\hat{e}(P, P)$ es un generador de V.
- Computable: Existe un algoritmo para calcular $\hat{e}(P, Q), \forall P, Q \in G$

Se necesitan algunas funciones hash como las siguientes: $H_1 : \{0,1\}^* \to G^*, H_2 : \{0,1\}^* \to \mathbb{Z}_q^*, H_3 : \mathbb{Z}_q^* \to \{0,1\}^n$, donde el tamaño del mensaje es definido por n. A parte de esto, $a \xleftarrow{r} N$ denota la selección de un elemento a aleatoriamente de un conjunto N, $b \leftarrow a$ la asignación del valor a a b y $||$ se usa para la concatenación.

Los pasos utilizados en este esquema criptográfico son los siguientes:

- **SETUP:** Este paso se ejecuta en la estación base para para establecer los parámetros básicos y para generar la clave pública (pk_{bs}) y la clave privada (sk_{bs}) de la estación base. Para ello, se genera un valor primo q basándose en algún dato privado $k \in \mathbb{Z}$ y el sistema escoge dos grupos cíclicos $(G,+)$ y $(V,\cdot)$ de orden q. Se selecciona aleatoriamente un mapa de emparejamiento bilineal $\hat{e} : G \times G \to V$, teniendo en cuenta que P es generador de G ($P \in G$). Finalmente se eligen las funciones hash H_1, H_2 y H_3.

$$sk_{bs} \xleftarrow{r} \mathbb{Z}_q^*$$
$$pk_{bs} \leftarrow sk_{bs} \cdot P$$

- **EXTRACT (ID_R):** En este paso, se generan las claves de cada robot, en la estación base, basándose en sus ID. El robot R con la identidad ID_R envía su información a la estación base. Se genera por un lado la clave pública $Q_{ID_R} \in G^*$, y la clave secreta $S_{ID_R} \in G$ teniendo en cuenta sk_{bs}. Este intercambio de información se realiza mediante un canal seguro.

$$Q_{ID_R} \leftarrow H_1(ID_R)$$
$$S_{ID_R} \leftarrow sk_{bs} \cdot Q_{ID_R}$$

- **SIGNCRYPTION ($S_{ID_{Ra}}$, ID_{Rb}, m):** Si un robot A con la identidad ID_{Ra} envía un mensaje $m \in \{0,1\}^n$ al robot B con la identidad ID_{rb}, la información será cifrada y firmada. La clave pública del robot receptor es generada teniendo en cuenta el ID_{rb}. El mensaje es firmado con $S_{ID_{ra}}$ y cifrado con $Q_{ID_{Rb}}$ dando como resultado σ (una 3-upla de tres componentes: c, T, U).

$$Q_{ID_{Rb}} \leftarrow H_1(ID_{Rb})$$
$$x \xleftarrow{r} \mathbb{Z}_q^*$$
$$T \leftarrow x \cdot P$$
$$r \leftarrow H_2(T||m)$$
$$W \leftarrow x \cdot pk_{bs}$$
$$U \leftarrow r \cdot S_{ID_{Ra}} + W$$
$$y \leftarrow \hat{e}(W, Q_{ID_{Rb}})$$
$$k \leftarrow H_3(y)$$
$$c \leftarrow k \oplus m$$
$$\sigma \leftarrow (c, T, U)$$

- **UNSIGNCRYPTION (ID_{Ra}, $S_{ID_{Rb}}$, σ):** El robot B recibe σ y analiza la información. Si la verificación que se describe posteriormente es correcta, se devuelve el mensaje $m \in \{0,1\}^n$. De otra manera, si hay algún problema en la firma o el cifrado de m, se devuelve $\perp$. La clave pública del emisor es generada teniendo en cuenta su identificador ID_{Ra} y el mensaje es descifrado con $S_{ID_{Rb}}$.

$$Q_{ID_{Ra}} \leftarrow H_1(ID_{Ra})$$
$$split \quad \sigma \quad as \quad (c, T, U)$$
$$y \leftarrow \hat{e}(S_{ID_{Rb}}, T)$$
$$k \leftarrow H_3(y)$$
$$m \leftarrow k \oplus c$$

$$r \leftarrow H_2(T||m)$$

Verificación:

$$\hat{e}(U,P) = \hat{e}(r \cdot S_{ID_{Ra}} + W, P) =$$
$$= \hat{e}(r \cdot sk_{bs} \cdot Q_{ID_{Ra}} + W, P) =$$
$$= \hat{e}(r \cdot sk_{bs} \cdot Q_{ID_{Ra}}, P) + \hat{e}(x \cdot pk_{bs}, P) =$$
$$= \hat{e}(Q_{ID_{Ra}}, sk_{bs} \cdot P)^r + \hat{e}(x \cdot P, pk_{bs}) =$$
$$= \hat{e}(Q_{ID_{Ra}}, pk_{bs})^r + \hat{e}(T, pk_{bs})$$

V. Conclusiones

Las redes MANET permiten la conexión entre diferentes dispositivos e intercambiar información de forma descentralizada. Como característica principal, estas tienen que los nodos de la red están en movimiento. Construyendo una red MANET sobre robots se añade la capacidad de mover los nodos de la misma de forma remota, de modo que no sea necesario desplazarse hasta los diferentes puntos donde se van a ubicar, sino que es posible moverlos remotamente gracias a la visualización de vídeo mediante streaming. Además, la propuesta incorpora seguimiento de objetos, el cual permite que al mover uno de los robots, el resto se mueva detrás de él de forma autónoma.

El uso de Raspberry Pi como dispositivo central de cada nodo da muchas posibilidades al sistema, ya que tener un dispositivo que ejecuta un sistema operativo Linux proporciona un valor añadido al sistema debido a la posibilidad de añadir diferentes servicios ejecutándose con algún fin o de desarrollar nuevas funcionalidades. Los puertos USB, GPIO, DISPLAY y CAMERA permiten la conexión de periféricos, sensores para la obtención de datos o incluso actuadores. Además, el bajo consumo de batería de estos dispositivos da la posibilidad de usarlos con una pequeña batería externa durante un largo periodo de tiempo.

En el sistema propuesto, con el fin de aprovechar la interfaz inalámbrica incorporada en la Raspberry Pi, se usa Wi-Fi como tecnología inalámbrica. A pesar de ello, es posible utilizar cualquier otra tecnología inalámbrica que permita un mayor rango de cobertura, siempre y cuando se disponga de un adaptador compatible con Raspberry Pi a través de alguno de sus puertos.

A parte de para la funcionalidad para la que ha sido diseñado este sistema, se puede utilizar para otras tareas. Aprovechando que cada uno de los nodos incorpora una cámara y un micrófono, estos pueden utilizarse para realizar tareas de videovigilancia, con el añadido que da la posibilidad de mover las cámaras remotamente gracias a los robots a las cuales están ancladas. Esto puede ser interesante si se desea inspeccionar una zona de difícil acceso, o en la cual corra peligro la integridad física de las personas.

La incorporación de mecanismos de seguridad proporciona un valor añadido al sistema, ya que se puede garantizar que los dispositivos cliente que se conecten a la red generada están seguros. Usando criptografía basada en identidad somos capaces de cifrar y firmar la información de una forma bastante flexible.

Para mejorar el seguimiento de objetos, en este caso códigos QR, en el futuro se puede mejorar el diseño incluyendo una pequeña pantalla en la parte trasera de cada robot. De esta manera, el QR aparecería en dicha pantalla, pudiendo ser cambiado cada poco tiempo. Esto aportaría un extra de seguridad al sistema, ya que cada poco tiempo el código quedaría obsoleto, evitando que cualquiera pudiera hacer que un robot se mueva utilizando un QR demasiado antiguo.

Agradecimientos

Investigación apoyada por el Ministerio de Economía y Competitividad Español, la TESIS2015010102, el Fondo Europeo FEDER y la Fundación CajaCanarias, bajo los proyectos TEC2014-54110-R, RTC-2014-1648-8, MTM2015-69138-REDT y DIG02-INSITU.

Referencias

[1] Xi, J., & Bettstetter, C. (2002). Wireless multihop Internet access: gateway discovery, routing, and addressing. Proceedings of 3GWireless, 2.

[2] Sugiura, Y., Sakamoto, D., Withana, A., Inami, M., & Igarashi, T. (2010, April). Cooking with robots: designing a household system working in open environments. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2427-2430). ACM.

[3] Bartsch, E. R., Fisher, C. W., France, P. A., Kirkpatrick, J. F., Heaton, G. G., Hortel, T. C., ... & Stigall, J. R. (2002). U.S. Patent No. 6,459,955. Washington, DC: U.S. Patent and Trademark Office.

[4] Malone-Lee, J. (2002). Identity-Based Signcryption. IACR Cryptology ePrint Archive, 2002, 98.

[5] Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. Future generation computer systems, 29(7), 1645-1660.

[6] Oda, H., Kulla, E., Ozaki, R., & Nishihara, N. (2016, November). Design of an adhoc testbed for IoT and WSAN applications using Raspberry Pi. In International Conference on Broadband and Wireless Computing, Communication and Applications (pp. 535-546). Springer, Cham.

[7] Kim, G. H., Nam, J. C., Mahmud, I., & Cho, Y. Z. (2016, July). Multi-drone control and network self-recovery for flying ad hoc networks. In Ubiquitous and Future Networks (ICUFN), 2016 Eighth International Conference on (pp. 148-150). IEEE.

[8] Boonsawat, V., Ekchamanonta, J., Bumrungkhet, K., & Kittipiyakul, S. (2010, May). XBee wireless sensor networks for temperature monitoring. In the second conference on application research and development (ECTI-CARD 2010), Chon Buri, Thailand.

[9] Thomas, J., Robble, J., & Modly, N. (2012, November). Off grid communications with android meshing the mobile world. In Homeland Security (HST), 2012 IEEE Conference on Technologies for (pp. 401-405). IEEE.

[10] Caballero-Gil, P., Caballero-Gil, C., & Molina-Gil, J. (2013). Design and implementation of an application for deploying vehicular networks with smartphones. International Journal of Distributed Sensor Networks, 9(12), 834596.

[11] Lodeiro-Santiago, M., Santos-González, I., Caballero-Gil, P., & Caballero-Gil, C. (2017). Secure system based on UAV and BLE for improving SAR missions. Journal of Ambient Intelligence and Humanized Computing, 1-12.

[12] Buiati, F., Puttini, R., de Sousa, R., Abbas, C. B., & Villalba, L. G. (2004, August). Authentication and autoconfiguration for MANET nodes. In International Conference on Embedded and Ubiquitous Computing (pp. 41-52). Springer, Berlin, Heidelberg.

[13] Glynos, D., Kotzanikolaou, P., & Douligeris, C. (2005, April). Preventing impersonation attacks in MANET with multi-factor authentication. In Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, 2005. WIOPT 2005. Third International Symposium on (pp. 59-64). IEEE.

[14] Umaparvathi, M., & Varughese, D. K. (2010, December). Evaluation of symmetric encryption algorithms for MANETs. In Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on (pp. 1-3). IEEE.

[15] Lindskog, J., Rydnell, G., & Larsson, P. (2008). U.S. Patent No. 7,443,822. Washington, DC: U.S. Patent and Trademark Office.

[16] Misra, S., Reisslein, M., & Xue, G. (2008). A survey of multimedia streaming in wireless sensor networks. IEEE communications surveys & tutorials, 10(4).

[17] Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A survey. Acm computing surveys (CSUR), 38(4), 13.

Prácticas de desarrollo de extensiones en el mercado de Mozilla tras Firefox Quantum

Félix Brezo Fernández
ElevenPaths, Telefónica CyberSecurity Unit
Ronda de la Comunicación s/n,
Edificio Oeste 1, Planta 1
felix.brezo@11paths.com

Yaiza Rubio Viñuela
ElevenPaths, Telefónica CyberSecurity Unit
Ronda de la Comunicación s/n,
Edificio Oeste 1, Planta 1
yaiza.rubio@11paths.com

***Resumen*—Con la llegada de Firefox 57, la fundación Mozilla introdujo WebExtensions, una nueva API para el desarrollo de extensiones de navegador La alta compatibilidad de esta nueva API con las ya existentes para los navegadores Google Chrome y Opera ha ampliado las funcionalidades que están también disponibles para el navegador de Mozilla ya que estas extensiones pueden ser adecuadas con mayor rapidez. Ante la ausencia de estudios similares que afronten las implicaciones de estos cambios y para analizar su impacto, en este documento se estudia el nuevo ecosistema de extensiones surgido en el mercado oficial de complementos de Mozilla Firefox, por medio de un análisis estático de los ficheros contenidos en las muestras distribuidas por este método, poniendo especial atención a los permisos requeridos actualmente por las extensiones distribuidas por este medio y a los posibles usos maliciosos que han emergido recientemente de la mano del *cryptojacking*.**

***Index Terms*—complementos, *cryptojacking*, Mozilla Firefox, navegación web, Neto, WebExtensions**

I. Introducción

Una parte signficativa del tráfico web tiene lugar a través de los navegadores que se han conformado en la herramienta por excelencia para acceder al contenido disponible *online*. Como tal, las extensiones de navegador llevan existiendo desde que surgen los navegadores. En el caso de Firefox, la gestión y el uso de extensiones y *add-ons* está disponible desde Firefox 0.2 (entonces Phoenix) [1] publicado el 1 de octubre de 2002. La comunidad de desarrolladores ha venido elaborando una gran cantidad de documentación y tutoriales para ampliar las funcionalidades del navegador con el objetivo de implementar soluciones que afrontan problemas concretos o que aportan aplicaciones no estándar del navegador.

Con Firefox 57 (a. k. a., Quantum) se dio entrada a WebExtensions, un cambio importante en las API que estandarizan el proceso de creación de complementos de una forma más compatible con otros navegadores como Google Chrome, Opera o Edge. Sin embargo, a pesar de la existencia de análisis previos fechados en 2014 sobre las extensiones disponibles para Google Chrome [2], los que hacen referencia a las extensiones para Firefox [3] están desactualizados y, en ningún caso se han realizado tras la adopción de WebExtensions. Por este motivo, este artículo tiene el objetivo de cubrir ese hueco analizando el estado actual de las extensiones disponibles en el mercado oficial de Mozilla a partir de entonces, identificando las prácticas de desarrollo más habituales y algunos de los potenciales usos maliciosos que pueden derivarse a partir de un análisis estático del contenido presente en las decenas de miles de extensiones oficiales.

El resto de este documento está estructurado como sigue. En la sección 2, se analizan los tipos de complementos de navegador existentes. En la sección 3, se detallan los detalles de la arquitectura de extensiones adoptada por Mozilla a partir de Firefox 57. En la sección 4, se recoge la metodología de análisis estático llevada a cabo sobre toda la población de extensiones de navegador disponibles a principios de 2018. En la sección 5, se definen posibles líneas de trabajo futuras relacionadas con las distintas tecnologías web. Por último, la sección 6 resume las conclusiones de este trabajo.

II. Tipos de complementos de navegador

La Fundación Mozilla define diferentes tipos de complementos oficiales para sus aplicaciones como Mozilla Firefox o Mozilla Thunderbird en función de su funcionalidad y objetivos [4]. Los complementos están desarrollados empleando tecnologías Web como HTML, Javascript o CSS y permiten a sus desarrolladores incorporar funcionalidades específicas que no fueron incorporadas en el producto de serie. Por ejemplo, pueden ser utilizadas para modificar la interfaz de usuario de un determinado sitio web o incorporar nuevas funcionalidades de conexión con otras plataformas. A continuación, se analizan los diferentes tipos de complementos de navegador existentes, desde las extensiones hasta los temas, pasando por los *scripts* de usuario y los *plugins* para tener una visión general del tipo de complementos existentes.

II-A. Las extensiones

Las extensiones de navegador son pequeños complementos que se integran en el navegador del usuario para añadir algunas funcionalidades avanzadas. Aunque las extensiones también se pueden distribuir de forma independiente (usando los ficheros .xpi o .crx distribuidos individualmente), la mayor parte de los desarrolladores optan por utilizar los canales de distribución oficiales de cada navegador. Este tipo de plataformas que también se usan en Google Chrome u Opera, tienen el objetivo de centralizar la distribución y actualización de las extensiones al ofrecer un punto de encuentro común para desarrolladores y usuarios, en donde estos últimos tienen la posibilidad de explorar las funcionalidades de la extensión y consultar la popularidad de las mismas entre otros usuarios.

En las extensiones se utiliza HTML, CSS y Javascript para la creación de las interfaces de usuario de la extensión (como el menú desplegable de las mismas o los menús contextuales por ejemplo) y Javascript para interactuar con los contenidos web visualizados por el usuario así como para implementar

la lógica de las nuevas funcionalidades en segundo plano. En el caso de Mozilla Firefox, las extensiones son archivos comprimidos en formato .zip renombrados a .xpi que incluyen una estructura de archivos concreta y que, desde Firefox 57, gira en torno al archivo *manifest.json* y, anteriormente, al fichero *package.json*.

Hasta noviembre de 2017, WebExtensions había estado solo parcialmente disponible en las versiones más recientes del navegador Firefox. Una de las ventajas que presentan las nuevas API es su alto grado de compatibilidad con las API para extensiones [5] desarrolladas por Google Chrome y Opera junto con W3C Draft Community Group. De hecho, la propia fundación Mozilla ofrece una página para testear la compatibilidad de las extensiones ya implementadas para estos navegadores (en formato .crx) con la nueva WebExtensions API [6] así como una lista de llamadas incompatibles o parcialmente soportadas [7].

II-B. Otros complementos que añaden funcionalidades

La funcionalidad de los navegadores también se puede ver ampliada con otro tipo de utilidades y complementos.

- Los *user scripts* son pequeñas piezas de código Javascript que permiten al usuario añadir comportamientos concretos al navegador en determinadas páginas web. Permiten a los desarrolladores y usuarios ejecutar tareas sencillas que amplíen las funcionalidades de una página web determinada, como resaltar un determinado contenido o incorporar nuevas opciones de usabilidad.
- Los temas son pequeños paquetes que ayudan a personalizar la interfaz de presentación del navegador.
- Los *plugins* por su parte hacen referencia a funcionalidades adicionales que incorporar al navegador.
- En esta categoría entrarían los *plugins* de Flash o Java tan populares hasta principio de década.

III. Estructura del sistema de extensiones de Firefox a partir de Firefox Quantum

En las versiones más recientes de Firefox, una extensión fundamental podría contener apenas dos o tres ficheros, siendo en todos los casos obligatoria la presencia del archivo *manifest.json*. Este fichero contiene detalles relevantes sobre la extensión como el nombre de la misma y la versión, el idioma por defecto, los permisos que requiere para su correcto funcionamiento o las estructuras de las direcciones URL sobre las que la extensión operará. Asimismo, este fichero también contiene apuntadores a los ficheros Javascript que se deberán ejecutar tanto en segundo plano (*background scripts*) como con cada página que cargue el propio navegador (*content scripts*).

- Los *content scripts* se corresponden con acciones que se ejecutan en los procesos correspondientes al contenido web que se carga en una pestaña y pueden interactuar con la página en sí misma manipulando el contenido directamente como cualquier otra porción de código Javascript que opere sobre ella. En el fichero *manifest.json* se puede especificar diferentes *content scripts* a cargar en función del patrón que cumpla cada dirección URL.
- Los *background scripts* se corresponden con *scripts* que se ejecutan junto con el proceso principal del navegador. Se pueden utilizar para modificar parte de la interfaz del navegador o interactuar con partes del sistema que han de ser persistentes a lo largo de una sesión.

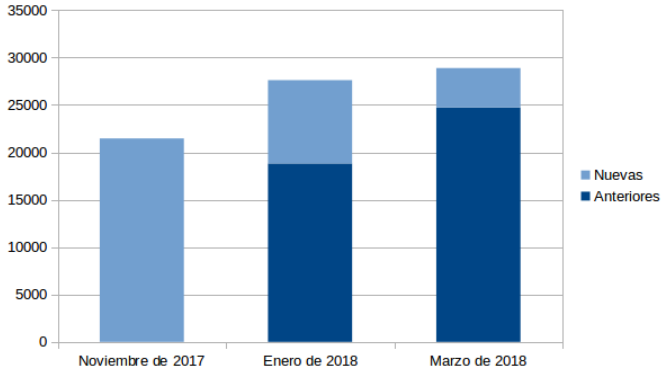


Figura 1. Número de extensiones identificadas en las sucesivas iteraciones.

Es habitual que la distribución oficial de las nuevas extensiones se realice a través de los mercados oficiales. Para que la extensión sea publicada también debe cumplir con unas políticas publicadas por Mozilla [8] y que hacen referencia a la prohibición de publicar extensiones que creen o expongan vulnerabilidades, que permitan la ejecución remota de código o que realicen determinadas acciones potencialmente intrusivas como la descarga masiva de datos o el envío de SMS o la intervención de llamadas sin el conocimiento del usuario. Este rol de supervisión recae hoy en día en la comunidad de revisores de Mozilla.

IV. Análisis estático de extensiones disponibles en el mercado oficial de Mozilla Firefox

Para analizar el ecosistema de extensiones surgido a partir de los cambios propuestos por Mozilla para Firefox 57 y superiores se ha procedido a descargar el total de extensiones de navegador disponibles en el mercado oficial de extensiones.

IV-A. Metodología

La población objeto de estudio ha sido recopilada del mercado oficial de complementos disponible para Mozilla Firefox. Con el fin de poder ilustrar la evolución de las prácticas observadas en el tiempo se ha efectuado la recolección de extensiones en tres instantes diferentes, noviembre de 2017 (antes del lanzamiento oficial de Quantum), enero de 2018 y marzo de 2018. En este proceso se han recuperado más de 34 386 extensiones diferentes a lo largo de estos períodos, distribuidas como se muestra en la Fig. 1. Como se puede comprobar, no todas las extensiones descargadas sufrieron actualizaciones entre los diferentes períodos.

Dado que los ficheros .xpi de las extensiones son en realidad ficheros comprimidos, se ha interpuesto una capa intermedia de descompresión de las extensiones descargadas para facilitar el procesamiento automático de sus ficheros. Para realizar el análisis se ha desarrollado Neto [9], una herramienta que integra el proceso de descompresión y análisis de extensiones de navegador liberada bajo licencia GPLv3+. Desde los módulos de análisis, Neto permite al analista configurar los procedimientos de análisis estático que considere oportuno para la extracción de expresiones regulares, el análisis de los ficheros de configuración o la búsqueda de código malicioso

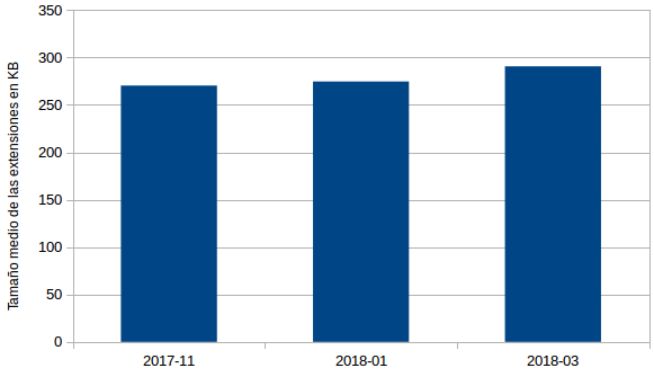


Figura 2. Tamaño promedio en KB de las extensiones disponibles.

concreto. En cualquier caso, Neto ha sido concebida como una herramienta diseñada para ser extensible y ampliada por la comunidad y permite la creación de nuevos plugins que amplíen sus funcionalidades. Gracias a esta herramienta se ha podido se ha podido llevar a cabo diferentes procedimientos de análisis estático sobre todo el corpus de extensiones como el estudio del tamaño del fichero y del tipo de extensión, los permisos utilizados, las estadísticas de localización de las mismas y la identificación de prácticas de programación potencialmente sospechosas vinculadas a la minería de criptodivisas.

Si atendemos al número de extensiones compatibles con el nuevo formato, se puede observar una actualización progresiva a lo largo de este proceso. Una forma sencilla de comprobar si la extensión implementa la versión más actualizada de las extensiones pasa por identificar la existencia de un archivo *manifest.json* en contraposición con el archivo *package.json* presente en formatos anteriores.

El proceso de migración a las nuevas API está bien documentado por Mozilla [10]. Entre otras cosas, se hace referencia a las modificaciones que es necesario aplicar en la definición del archivo *manifest.json* . Pese a ello, a fecha de marzo de 2018 todavía existían 2837 extensiones que no habían sido portadas con lo que ya no podrían ser utilizadas en navegadores posteriores a Firefox 57.

IV-B. Prácticas de desarrollo identificadas

Aunque la muestra no es excesivamente amplia, sí que se ha identificado un ligero crecimiento en el tamaño medio de las extensiones disponibles en el *marketplace* de Firefox. En cuatro meses han pasado de los 270,34 KB a los 290,71 KB, lo que supone un incremento promedio superior al 7,5 %. Sin embargo, el tamaño medio de las extensiones solo ha de emplearse como una referencia dado que existe una gran desviación en cuanto al tamaño de las mismas dado que se han encontrado veinticuatro extensiones con tamaños superiores a los 20 MB de tamaño, hasta incluso alcanzar los 100 MB[1].

Atendiendo a la naturaleza de los ficheros identificados en las extensiones, se puede observar que los ficheros de tipo Javascript (*.js*) son los más habituales (más del 30 % de los ficheros son de este tipo), lo cual es razonable teniendo en cuenta que la mayor parte de las funcionalidades de las extensiones se implementan a través de ellos. Asimismo, en la Tabla I también destacan algunos formatos de imágenes (principalmente, en formato PNG), seguidos de los ficheros en formato Json (propios de archivos *manifest* y también de la localización como se verá más adelante) y de otros vinculados con el desarrollo de interfaces Web usando CSS y HTML.

Este análisis de los ficheros también ha desvelado la existencia de archivos que no deberían estar presentes en aplicaciones disponibles en producción. Entre ellos destacan una gran cantidad de archivos vinculados a la gestión de versiones como Git y varios centenares de archivos temporales (fácilmente identificables por estar terminados en el carácter «˜») o con referencias a archivos de respaldo. Algunos de estos ficheros, especialmente los asociados a *software* de control de versiones, podrían contener información de los entornos de pruebas que no debiera ser pública. Esta circunstancia pone de manifiesto que no todos los desarrolladores cumplen los mismos estándares y que el hecho de que la extensión haya sido publicada en un *marketplace* oficial no descarta la posibilidad de que queden expuestos datos por error u omisión de los equipos de desarrollo y de control de la calidad del código. Paralelamente, la existencia de ficheros multimedia podría ser utilizada para identificar a los autores en el transcurso de una investigación realizando análisis de metadatos de dichos archivos o tratando de identificar copias de los mismos disponibles en internet.

IV-C. Estadísticas de localización

Aunque las extensiones intentarán ajustar el idioma de la interfaz de forma automática a la del navegador del usuario, en el fichero *manifest.json* también es posible definir los ajustes de localización que se utilizarán por defecto en caso de que no existan los correspondientes al idioma del usuario. Más del 90 % de las 4216 extensiones que tienen configurado un idioma por defecto optan por el inglés como idioma principal (el 84,25 % como inglés y el 6,62 % en sus versiones británica,

Tabla I
TIPOS DE FICHEROS MÁS HABITUALES ENLAS EXTENSIONES ANALIZADAS.

Tipo de fichero	2017-11	2018-01	2018-03
Javascript	249 784	305 749	309 712
PNG	217 054	245 273	254 381
Json	35 846	48 550	54 679
CSS	33 060	38 116	39 258
DTD	32 545	32 143	31 696
SVG	26 942	28 210	30 374
HTML	21 107	25 184	26 693
MF	19 487	25 152	26 316
SF	19 485	25 150	26 311
RSA	19 485	25 150	26 311
JPG	17 977	19 156	21 814
Properties	22 033	21 292	20 995
RDF	15 852	18 150	17 905
XUL	16 410	17 938	17 691
Gif	11 964	12 783	12 702
Manifest	8053	9201	9098
Markdown	4117	5066	5267
TXT	4570	4882	4899
Control de versiones	3286	2567	2830
Temporales	240	323	307
Otros	45 457	55 444	57 143

[1]enable_screen_capturing_with_websocket_in_firefox-1.0.007-fx.xpi ocupa 124,71 MB mientras que lolbyte_league_of_legends_summoner_search-2.10.15fx.xpi alcanza los 99,44 MB.

canadiense o estadounidense), seguidos, muy de lejos, por el ruso (2,66 %), el francés (1,54 %) y el alemán (1,02 %).

Los idiomas más utilizados por los actuales desarrolladores de extensiones para Firefox son el inglés (el más habitual también por omisión), alemán, francés, ruso y español, tal y como se recoge en la Tabla II que agrupa el total de extensiones diferentes analizadas en este período. La relación de dicha tabla tiene en cuenta la existencia de archivos de localización de extensiones en dichos idiomas.

IV-D. Permisos más populares

El sistema de permisos se basa en la asunción de que generalmente las aplicaciones requieren menos privilegios que la totalidad de ellos. De forma similar a como ocurre en ecosistemas relacionados con el desarrollo de aplicaciones para *smartphones*, las extensiones basan su funcionamiento en un sistema de permisos que son aprobados explícitamente por el usuario en el momento de instalar una nueva extensión o cuando esta recibe una actualización que requiere un permiso nuevo. Un esquema de seguridad basado en permisos en lugar de en modelo de usuarios presenta las siguientes ventajas [11]:

- *Consentimiento del usuario*. El modelo abre la puerta a que los usuarios con conocimientos tengan capacidad para detectar usos abusivos en la propia instalación.
- *Defensa en profundidad*. El control de los permisos limita la explotación de posibles fallos de una aplicación al ámbito para el que ha recibido permisos.
- *Facilidad de revisión*. La limitación del ámbito de una extensión en base a permisos también permite acotar las potenciales extensiones implicadas en una determinada actividad sospechosa si se conocen los permisos que requeriría, lo que también facilita la labor del analista.

Para mitigar los efectos potencialmente maliciosos que pudiera tener un uso erróneo o no autorizado de una determinada funcionalidad del navegador el desarrollador deberá definir los permisos que requiere una extensión en el archivo *manifest.json*. De forma paralela, en las extensiones también se fuerza la Content Security Policy (CSP) para prevenir la realización de ataques de *cross-site scripting*, *clickjacking* u otras prácticas de inyección de código malicioso.

Se pueden identificar dos tipos de permisos: los que hacen referencia a funcionalidades que deben estar disponibles para la extensión (como las *cookies*, el historial, las notificaciones o los menús contextuales) y los que especifican bajo qué direcciones URL debe activarse la extensión. Aunque se puede activar de forma global con *<all_urls>*, es preferible ser más selectivo para aligerar la carga del navegador.

En la Tabla III se listan los 20 permisos más populares identificados. Si descartamos los asociados a las direcciones URL sobre las que aplica cada extensión, el promedio de permisos requeridos por las extensiones que los demandan asciende a 2,88 por extensión. Como cada permiso tiene diferentes implicaciones desde el punto de vista de la seguridad, se describe el cometido de los más populares a continuación.

- *storage*. Este permiso permite a las extensiones al macenar y recuperar cierta cantidad de información así como escuchar posibles cambios en los objetos almacenados.
- *tabs*. Se utiliza para interactuar con el sistema de pestañas del navegador y recuperar una lista de las pestañas abiertas, así como abrir, actualizar, mover, recargar o eliminar las pestañas. Esta funcionalidad es muy potente pero también entraña sus riesgos.
- *activeTab*. Da un permiso temporal a la extensión para acceder al contenido de la pestaña activa actual. Esto es útil como alternativa al uso del permiso *tabs* siempre y cuando la interacción de la extensión con la pestaña activa ocurra bajo demanda del usuario (por ejemplo, al pulsar un atajo o hacer clic en el icono de la extensión).
- *menus*. También conocida como *contextMenus* en versiones anteriores a Firefox 55, esta funcionalidad permite al desarrollador añadir submenús al sistema de menús que se activan en el navegador con botón derecho.
- *webRequest*. Añade *listeners* a la extensión para que pueda monitorizar peticiones HTTP, lo que puede conllevar ciertos problemas de privacidad.
- *notifications*. Se utiliza para mostrar notificaciones al

Tabla II
20 DE IDIOMAS MÁS EMPLEADOS PARA LOCALIZAR EXTENSIONES

Idioma	Total	Porcentaje
Inglés	27 695	80,54 %
Alemán	24 458	71,13 %
Francés	24 453	71,11 %
Ruso	23 121	67,24 %
Español	22 176	64,49 %
Italiano	21 782	63,35 %
Polaco	20 750	60,34 %
Chino (China)	20 637	60,02 %
Neerlandés	20 424	59,40 %
Japonés	20 409	59,35 %
Turco	19 364	56,31 %
Portugués (Brasil)	19 347	56,26 %
Chino (Taiwán)	18 492	53,78 %
Koreano	18 364	53,41 %
Árabe	17 361	50,49 %
Checo	17 150	49,87 %
Ucraniano	17 112	49,76 %
Finlandés	16 744	48,69 %
Sueco	16 690	48,54 %
Danés	16 662	48,46 %

Tabla III
PERMISOS MÁS HABITUALES PRESENTES EN EXTENSIONES DE FIREFOX.

Idioma	Total
storage	8931
tabs	8243
activeTab	4445
<all_urls>	3386
menus	3701
webRequest	2935
notifications	2621
webRequestBlocking	2050
cookies	1530
webNavigation	1404
downloads	1034
clipboardWrite	882
alarms	748
private-browsing	657
unlimitedStorage	640
multiprocess	625
nativeMessaging	548
management	534
bookmarks	531
history	387
Otros	3150

usuario empleando el mecanismo correspondiente implementado para cada sistema operativo.
- *webRequestBlocking*. Es un permiso complementario a *webRequest* que permite, por ejemplo, examinar y modificar el cuerpo de las peticiones HTTP tan pronto como son recibidas por el navegador.
- *cookies*. Permite a la extensión leer y escribir *cookies* así como notificarle cuando estas experimenten cambios.
- *webNavigation*. Funciona de forma similar a *webRequests* pero monitoriza eventos de navegación que tienen lugar en el contexto de una pestaña capturando timestamps de los eventos, tipos de transición, etc.

IV-E. La minería de criptodivisas y el criptojacking en extensiones de navegador

Desde finales de 2017, el uso no autorizado de la capacidad de cómputo para el minado de criptodivisas ha recobrado un interés que ya había sido previsto años atrás [12]. El proceso de minado de criptodivisas es una tarea que requiere que los nodos que conforman la red lleven a cabo una tarea matemática. Así, por *cryptojacking* se entiende a la práctica por medio de la cual una aplicación maliciosa hace uso no autorizado de la capacidad de cómputo del equipo infectado con el objetivo de minar diferentes criptodivisas aprovechando que el coste energético derivado del minado corre a cargo de la víctima. En este sentido, informes de empresas de seguridad como Check Point [13] ya situaban en octubre de 2017 la amenaza del minado no autorizado de criptodivisas como una de las más importantes del mes. Sin embargo, la existencia de *hardware* altamente especializado para la realización de estas tareas en el ámbito de Bitcoin ha motivado que muchos desarrolladores de *malware* migren a otras criptodivisas para cuyas tareas de prueba de trabajo es más difícil dar con *hardware* específico que sea eficiente.

Entre ellas se encuentra Monero, una criptodivisa que, además, ha sido concebida para ser anónima y privada. A diferencia de Bitcoin, en donde se puede consultar en la cadena de bloques las direcciones emisoras y receptoras de cada operación así como el importe de cada transacción, en Monero no es posible llevar a cabo esta tarea por la implementación del método de Ring Confidential Transactions definido por Noether *et al.* en 2016 [14]. Además, las características del algoritmo de prueba de trabajo utilizado para las labores de minería han impedido la existencia de dispositivos ASIC de minado especializados, de modo que se democratizan los esfuerzos de minado al tratar de mantener la rentabilidad de la capacidad de cómputo de los equipos domésticos.

Para aprovechar este fenómeno y de forma similar a como ya ocurriera con Bitcoin en 2011 a través del sitio web de Bitcoin Plus [15], empezaron a surgir minadores destinados a ser ejecutados en el contexto de un navegador. Plataformas como coinhive.com o crypto-loot.com entre otras muchas empezaron a ofrecer código Javascript de minado destinado a ser incrustado en páginas web. Bajo el pretexto de conformar un mecanismo de financiación alternativo a la publicidad, estos servicios han sido utilizados por algunas páginas web de forma intencionada o, en algunos casos, como consecuencia de infecciones en sus servidores o errores de configuración.

La repercusión mediática de la amenaza ha dado lugar a la proliferación de extensiones y aplicaciones que, principalmente, intentan detectar la existencia de este tipo de comportamientos en los sitios web que visita el usuario. Pese a que en noviembre de 2017 apenas existían seis aplicaciones que bloqueaban la minería (Nominers, Adnauseam, Mining Blocker, NoMiner, uBlock y Antiminer), aunque esta cifra ha ido en aumento conforme se hacían más populares.

La naturaleza de estas tecnologías originalmente concebidas para ser utilizadas en páginas web convencionales también las hace usables en el seno de una extensión que, básicamente, utiliza estas mismas tecnologías. Del total de extensiones identificadas en el mercado oficial vinculadas a prácticas de minería, apenas se han podido identificar unas pocas extensiones que implementen esta práctica. Para comprobarlo, en Neto se ha utilizado una lista de 115 indicadores [9] que incluyen dominios y expresiones regulares de direcciones URL empleadas por minadores conocidos en páginas web convencionales.

Una de las extensiones que implementaba el minado sin ponerlo en conocimiento del usuario es la versión 1.0.0 del juego *2048 in Pop Up*[2] que también contenía código de Javascript propio de Coinhive en su primera versión, pese a que ya no se incluye este código en versiones más recientes de la aplicación. También se han podido identificar otras que llevan a cabo esa labor de minado pero informando en su descripción de su cometido. Ejemplos de estos comportamientos son las dos versiones existentes de la extensión Cryplo [16], cuyas versiones 1.0 fueron identificadas en la primera iteración de 2018 como *cryplo-1.0-fx-linux.xpi*, *cryplo-1.0-fx-mac.xpi* y *cryplo-1.0-fx-windows.xpi*. Todas estas extensiones contienen código asociado a Coinhive al incluir el fichero de Javascript de minado *coinhive.min.js* en la carpeta local *./res/js*.

Para la identificación de este tipo de prácticas, no es suficiente con la enumeración de ficheros con nombres concretos o con el uso de algunos con un resumen criptográfico concreto ya que estos podrían ser renombrados o alterados ligeramente para modificar su huella. De hecho, es habitual que muchos minadores ofusquen el contenido de los correspondientes ficheros Javascript lo que es indicador de la necesidad de contar con soluciones con capacidad para analizar el comportamiento de una extensión una vez instalada. En este sentido, el procedimiento de revisión del contenido especifica explícitamente la búsqueda de contenido ofuscado [17] pero no existen datos publicados sobre el número de extensiones que han intentado hacer uso de dichas funcionalidades en el mercado de Firefox a pesar de que sí que se han conseguido identificar ejemplos que implementan estas tecnologías. La posición de Mozilla sobre la minería [18] pasa por prohibir explícitamente «el uso de minadores de criptodivisas o funcionalidades similares en una extensión» y es similar a la también adoptada por Apple con respecto al minado de criptodivisas desde aplicaciones para iOS a partir de junio de 2018 [19].

V. Líneas de trabajo futuras

Este estudio pone el foco en las extensiones disponibles en el *marketplace* de Mozilla para su navegador Firefox y en un análisis de los contenidos de las mismas. Sin embargo, todavía no existen estudios que comparen las diferentes prácticas seguidas por los desarrolladores de otros navegadores como

[2]SHA1: e28fa15a99f27f56132651450a6b21717c04eccb.

Google Chrome u Opera, que también ofrecen la posibilidad de distribuir sus complementos a través de sus correspondientes mercados oficiales. Aunque no era el objetivo de este documento, y dado que las extensiones se distribuyen a través de mercados en los que se registran opiniones y valoraciones de los usuarios, se podría utilizar dicha información como complemento al proceso de diagnóstico del riesgo que entraña la instalación de una extensión.

Asimismo, también sería conveniente analizar el uso de paquetes y extensiones para otras aplicaciones, especialmente de aquellas basadas en *frameworks* que se apoyan en tecnologías web como Electron. La existencia de mercados de este tipo de extensiones también entraña el riesgo de que estos sean utilizados para distribuir complementos maliciosos. Como tampoco existen estudios que analicen el funcionamiento de utilidades como los *user scripts*, convendría ampliar estos estudios al conjunto de *scripts* distribuidas en mercados como GreasyFork o UserScripts-Mirror.

En todo caso, el análisis llevado a cabo es un análisis estático del contenido distribuido por los desarrolladores de extensiones. Algunos estudios han afrontado el análisis de extensiones simulando conexiones a direcciones URL de sitios web que activaban la extensión como ya hicieron Kapravelos *et al.* [20]. Sin embargo, su ámbito se circunscribe al ecosistema de Google Chrome y data de 2014. Además, el auge de los métodos de ofuscación y el uso de técnicas de desempaquetamiento dinámico del código Javascript que ya se han observado en *scripts* de minería en páginas web convencionales podría ser utilizado también para enmascarar código malicioso en el seno de las mismas dificultando los procedimientos de análisis actuales. Para atajar este tipo de problemáticas sería necesario ampliar el ámbito del estudio incluyendo técnicas de análisis dinámico con el objetivo de tener una idea más aproximada del comportamiento real de las aplicaciones y de sus conexiones con aquellas aplicaciones con las que pudiera estar interactuando.

VI. Conclusiones

Afortunadamente para quienes deseen profundizar en el desarrollo de nuevas funcionalidades existe una buena documentación en inglés y multitud de tutoriales oficiales sobre la creación de extensiones y el uso de las distintas API que añaden cada una de las nuevas funcionalidades. Sin embargo, la gran cantidad de posibilidades que se ofrecen no han sido exploradas desde el punto de vista de la seguridad teniendo en cuenta que gran parte del uso diario que hacemos de la tecnología pasa por el navegador.

De forma similar a como ocurre en el desarrollo de aplicaciones móviles, son los desarrolladores quienes definen los permisos que una extensión requiere. Teniendo en cuenta el uso masivo que hacemos los usuarios de los navegadores para tareas cotidianas (comunicación, redes sociales, operaciones bancarias, etc.), la falta de transparencia en la gestión de los permisos puede suponer un riesgo porque los usuarios no sean conscientes de qué pueden hacer.

Así, el hecho de que algunas extensiones superan con creces las decenas de megabytes y de que existan otras con capacidad para disparar el consumo de recursos del equipo (e. g., el uso de CPU en aplicaciones de minería) son un sintoma más de la posibilidad de ampliar notablemente las capacidades de un navegador. La posibilidad de contar con herramientas que permitan monitorizar el comportamiento de las extensiones oficiales comporta, por tanto, una oportunidad para crear un sistema de detección temprana de posibles incidentes de seguridad, facilitando un seguimiento más transparente de las nuevas funcionalidades que se implementen en el futuro y ofreciendo, como consecuencia, un mayor control al usuario sobre qué ejecuta en su máquina. En este sentido, pese a que los procesos de revisión implementados por los mercados oficiales ya suponen un primer filtro para controlar la publicación de extensiones maliciosas que ha conseguido evitar gran parte de los problemas, no hay que perder de vista que la instalación manual seguirá siendo un vector de distribución de extensiones maliciosas de difícil control.

Referencias

[1] Mozilla, *Phoenix 0.2 (Santa Cruz) release Notes and FAQ*, Oct. 2002. Available online at https://website-archive.mozilla.org/www.mozilla.org/firefox_releasenotes/en-US/firefox/releases/0.2.html.

[2] L. Bauer, S. Cai, L. Jia, T. Passaro, and Y. Tian, "Analyzing the dangers posed by Chrome extensions," in *2014 IEEE Conference on Communications and Network Security*, pp. 184–192, Oct. 2014.

[3] J. Wang, X. Li, X. Liu, X. Dong, J. Wang, Z. Liang, and Z. Feng, "An Empirical Study of Dangerous Behaviors in Firefox Extensions," in *Information Security*, Lecture Notes in Computer Science, pp. 188–203, Springer, Berlin, Heidelberg, Sept. 2012.

[4] MDN Web Docs, *Add-ons*, Feb. 2018. Available online at https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons.

[5] Chrome Development Team, *What are extensions?* Available online at https://developer.chrome.com/extensions.

[6] Mozilla Firefox, *Extension Compatibility Tester*. Available online at https://www.extensiontest.com/.

[7] MDN Web Docs, *Chrome incompatibilities*, Apr. 2018. Available online at https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/Chrome_incompatibilities.

[8] MDN Web Docs, "Add-on review criteria," July 2016. Available online at https://developer.mozilla.org/en-US/docs/Mozilla/Marketplace/Add-on_submission/Review_criteria.

[9] E. I. Lab, *Neto — A tool to analyse browser extensions*, May 2018. Available online at https://github.com/ElevenPaths/neto.

[10] MDN Web Docs, *Porting a legacy Firefox extension*, Nov. 2017. Available online at https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/Porting_a_legacy_Firefox_add-on.

[11] A. P. Felt, K. Greenwood, and D. Wagner, "The effectiveness of application permissions," in *Proceedings of the 2Nd USENIX Conference on Web Application Development*, WebApps'11, (Berkeley, CA, USA), pp. 7–7, USENIX Association, 2011.

[12] F. Brezo and P. G. Bringas, "Issues and Risks Associated with Cryptocurrencies Such as Bitcoin," (Venecia (Italia)), pp. 20–26, Oct. 2012.

[13] C. P. R. Team, *October's Most Wanted Malware: Cryptocurrency Mining Presents New Threat*, Nov. 2017. Available online at https://blog.checkpoint.com/2017/11/13/octobers-wanted-malware-cryptocurrency-mining-presents-new-threat/.

[14] S. Noether, A. Mackenzie, and M. R. Lab, "Ring Confidential Transactions," *Ledger*, vol. 1, pp. 1–18, Dec. 2016.

[15] S. Eskandari, A. Leoutsarakos, T. Mursch, and J. Clark, "A first look at browser-based Cryptojacking," *arXiv:1803.02887 [cs, econ]*, Mar. 2018. arXiv: 1803.02887.

[16] Cryplo, *Cryplo 1.0*, Nov. 2017. Available online at https://addons.mozilla.org/en-US/firefox/addon/cryplo/.

[17] Mozilla, "FirefoxOS Add-on Review guide," May 2016. Available online at https://wiki.mozilla.org/Marketplace/Reviewers/FxOSAdd-ons/Testing.

[18] MDN Web Docs, *Add-on Policies*, Apr. 2018. Available online at https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/AMO/Policy/Reviews.

[19] Apple Inc., *App Store Review Guidelines*, 2018. Available online at https://developer.apple.com/app-store/review/guidelines/.

[20] A. Kapravelos, C. Grier, N. Chachra, C. Kruegel, G. Vigna, and V. Paxson, "Hulk: Eliciting malicious behavior in browser extensions," in *Proceedings of the 23rd USENIX Conference on Security Symposium*, SEC'14, (Berkeley, CA, USA), pp. 641–654, USENIX Association, 2014.

Seguridad aplicada a una herramienta para el Trastorno por Déficit de Atención e Hiperactividad

Nayra Rodríguez-Pérez, Alexandra Rivero-García, Pino Caballero-Gil and Josué Toledo-Castro
Departamento de Ingeniería Informática y de Sistemas
Universidad de La laguna
{mrodripe, ariverog, pcaballe, jtoledoc}@ull.edu.es

***Resumen*—Las tecnologías de la información están cada vez más integradas en la sociedad. Los dispositivos móviles son utilizados por personas entre un gran rango de edades, lo que permite desarrollar herramientas que faciliten su vida y que abordan una amplia variedad de problemas. Mediante el uso de m-health, término utilizado para referirse a la práctica de la medicina basada en el uso de dispositivos móviles, es posible desarrollar nuevas herramientas que ayuden a mejorar las habilidades cognitivas, con ayuda de la gamificación. En particular, se describe una aplicación segura de m-health diseñada para ayudar a las personas que padecen Trastorno por Déficit de Atención e Hiperactividad, aplicando una capa de seguridad que mantenga la confidencialidad de la información médica de los usuarios, así como la protección en las comunicaciones. La información y los datos médicos deben ser especialmente protegidos.**

***Index Terms*—Seguridad, m-health, Trastorno por Déficit de Atención e Hiperactividad, Android**

I. Introducción

El uso de herramientas tecnológicas para ayudar en el tratamiento de un determinado trastorno neurológico ayuda y facilita en gran medida la vida de las personas. El uso de técnicas de gamificación en este tipo de herramientas es un medio fundamental para proporcionar distintas soluciones que fomenten sus capacidades cognitivas. Una aplicación móvil que permita la integración de los distintos perfiles involucrados en el Trastorno por Déficit de Atención e Hiperactividad [1] facilita la relación entre el afectado y su entorno.

El desorden que nos ocupa, el Trastorno por Déficit de Atención e Hiperactividad (TDAH), es un trastorno neurobiológico asociado a un importante impacto funcional, personal y social [2]. Sus síntomas se manifiestan con muestras de inatención, impulsividad e hiperactividad, que producen el deterioro en múltiples áreas de la vida de la persona afectada, como en el ámbito académico, social o familiar. Se trata de uno de los trastornos infantiles más frecuentes, con una prevalencia aproximada de entre el 3-7 % de los niños a nivel mundial dependiendo de factores como la edad, el entorno sociocultural o las técnicas empleadas en su tratamiento. Con una prevalencia mayor en el rango de edad de 6-9 años [3].

La aplicación móvil que se propone tiene como objetivo integrar las diferentes figuras implicadas en el TDAH, generando una herramienta que permita compartir información entre ellos de forma segura. El sistema propuesto proporciona un seguimiento, diagnóstico y mejora de los pacientes con TDAH y a su vez un esfuerzo coordinado entre la familia, equipo médico y profesores. Asimismo, es importante establecer un alto grado de seguridad debido a que los datos médicos se definen como información especialmente protegida, lo cual debe cumplir unos principios fundamentales de protección.

El trabajo que se detalla a continuación se ha estructurado de la siguiente forma. En primer lugar se exponen aplicaciones existentes relacionadas con la temática tratada. Seguidamente se realiza una definición del sistema que se ha desarrollado, la aplicación móvil. En la siguiente sección se detallan los métodos de seguridad empleados en este sistema. Finalmente, se reseñan algunas conclusiones y líneas de trabajo futuras.

II. Estado del arte

Las herramientas m-health han demostrado ser útiles para proporcionar tratamientos para diferentes trastornos neurológicos. El uso de dispositivos móviles para tratamientos de salud (m-health) se engloba en el uso de la tecnología para el apoyo en tratamientos médicos (e-health). En los últimos años, se ha puesto especial énfasis en el desarrollo de aplicaciones m-health en dispositivos móviles para el aprendizaje y el apoyo en el tratamiento de diferentes tipos de trastornos mentales [4]. En este sentido, se ha llevado a cabo una evaluación de la calidad de un sistema en particular [5] y [6] para verificar la satisfacción de los niños al usarlo. Además, se presentan dos reseñas de uso de aplicaciones para dispositivos móviles en el tratamiento de trastornos alimenticios [7] y [8].

Por otra parte, existe también una revisión de las características y la calidad del contenido de las aplicaciones de m-health, para el trastorno bipolar [9]. Los efectos de la gamificación en las aplicaciones móviles para el aprendizaje y con niños en particular, se estudian también en otros trabajos. Los beneficios de una aplicación de gamificación basada en el lenguaje es una fuente de estudio para examinar el efecto que tiene en los niños y su efectividad académica [10].

En lo que se refiere a la protección y a la seguridad en aplicaciones móviles m-health, existen varios estudios que recomiendan el uso de diferentes herramientas criptográficas para proteger los datos médicos de los usuarios. Por lo tanto, los desarrolladores de aplicaciones móviles deben tener en cuenta las recomendaciones para cumplir con la legislación vigente en materia de privacidad y seguridad [11], ya que la falta de seguridad en las aplicaciones móviles puede causar graves problemas de privacidad de los datos, lo que puede tener un gran impacto en los usuarios [12]. Por este motivo, ha sido estudiada la detección de pérdidas de información en aplicaciones móviles [13] donde se destaca la importancia de analizar los principales problemas de seguridad existentes en aplicaciones móviles. Un enfoque centrado en el usuario se considera la base del desarrollo de aplicaciones, las soluciones de privacidad, seguridad y confianza deben incluirse para garantizar una funcionalidad sin problemas [14].

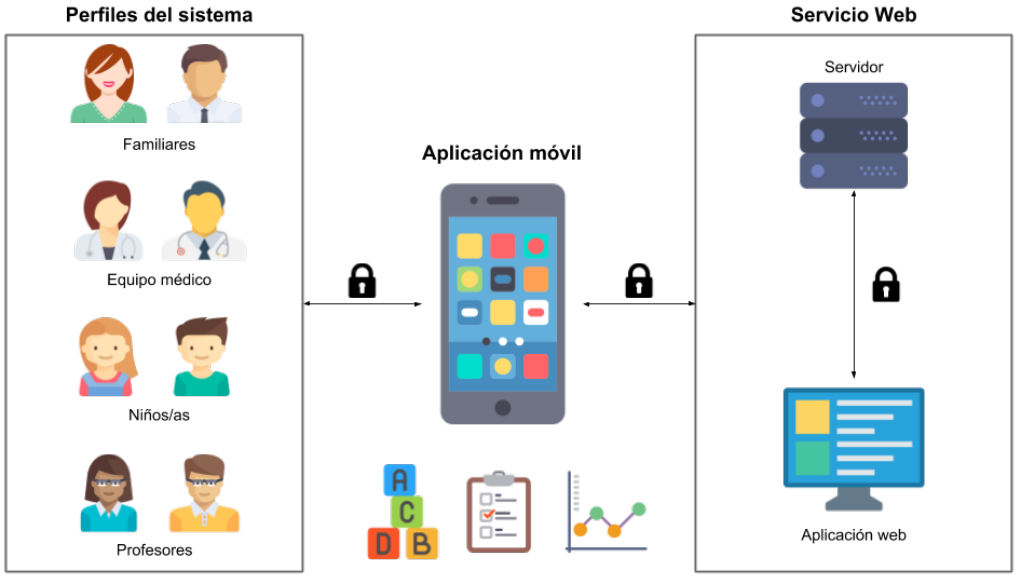


Figura 1. Esquema general

III. Marco de la herramienta

Se propone una herramienta que facilite la integración de los principales perfiles involucrados en el estudio del trastorno por TDAH (Figura 1). La comunicación de forma segura ha sido uno de los principales objetivos de este trabajo entre los perfiles implicados. Este sistema se ha elaborado en colaboración con la sección de trastornos infantiles del Hospital Universitario de Canarias, se ha desarrollado una aplicación móvil cuyo propósito es proporcionar una ayuda en el tratamiento del Trastorno por Déficit de Atención e Hiperactividad (TDAH), así como para seguir la evolución del progreso de las personas afectadas.

III-A. Perfiles del sistema

La herramienta que se propone está destinada a cuatro perfiles principalmente: la persona afectada, el entorno familiar, el equipo médico y un miembro del centro educativo. Estos perfiles engloban las figuras indispensables para proporcionar una ayuda en el tratamiento del TDAH.

En primer lugar, en el rol de las personas afectadas (niños/as o adolescentes) se incluyen aquellas con posibilidad de padecer el trastorno. Con este perfil podrán acceder a los juegos disponibles en la aplicación, en los cuales podrán desarrollar y estimular sus capacidades cognitivas, teniendo como objetivo reforzar el esfuerzo más que la obtención de resultados. En segundo lugar, un miembro del entorno familiar (padres, madres o tutores) a través de la aplicación, podrán realizar un prediagnóstico inicial y monitorear el progreso del niño. En tercer lugar, el equipo médico (médicos, psicólogos, psiquiatras o educadores) involucrados en el tratamiento de un paciente con este trastorno, tendrán acceso a los resultados obtenidos en las pruebas realizadas por los tutores y a un seguimiento de la evolución de los pacientes registrados. Finalmente el rol del centro educativo representado principalmente por el profesor tutor de un niño/a es fundamental como fuente de información acerca del comportamiento del niño en el contexto escolar. Es común que el docente encuentre alumnos con dificultades o problemas que pueden ser más fácil de resolver a través de una colaboración con el entorno familiar. De este modo ambos perfiles se implican en la evolución y el aprendizaje del niño a través de una colaboración entre centro educativo y familia.

III-B. Aplicación móvil

La herramienta propuesta se ha desarrollado para una aplicación móvil en Android. Su principal objetivo es ser un mecanismo que permita el diagnóstico, seguimiento y mejora de los pacientes afectados con el TDAH. Se tiene en cuenta la necesidad que poseen los familiares de las personas perjudicadas en saber si pueden o no padecer el trastorno.

La aplicación proporciona a los tutores un cuestionario inicial con preguntas basadas en el comportamiento, las características del paciente y las dificultades con su conducta en diferentes situaciones del día. Posteriormente, estos datos son evaluados por un miembro del personal médico, que se encargará de diagnosticar si el paciente sufre o no de la enfermedad. En caso afirmativo, indica el nivel estimado de desorden, según las respuestas obtenidas.

Después de evaluar los resultados, el personal médico envía su evaluación a los tutores del paciente y mantiene a posteriori un seguimiento de cada paciente afectado. Los tutores podrán registrar al niño, indicando los datos necesarios para su acceso a la aplicación y para llevar a cabo las actividades multimodales propuestas para ayudar a mejorar las funciones cognitivas del afectado (memoria, atención, lenguaje, percepción, resolución de problemas o planificación). Un diagnóstico precoz y un tratamiento adecuado son factores determinantes para una evolución positiva. Por lo tanto, la aplicación propuesta tiene por objeto mejorar la calidad de vida de los pacientes y familiares, y ayudar en su tratamiento con actividades y ejercicios que fortalezcan su desarrollo mental.

En el desarrollo se ha tenido en cuenta la necesidad de que los miembros de la familia resuelvan la incertidumbre acerca de si su hijo sufre el trastorno además de generar una herramienta que ayude en su tratamiento. Un objetivo

importante en el sistema es proveer un enlace entre los principales perfiles implicados y permitir el flujo de información entre ellos. Así, el sistema desarrollado unifica en una sola aplicación las principales funcionalidades presentes en este tipo de aplicaciones médicas. Por lo tanto, se propone una aplicación móvil segura que incluya diferentes funciones para cada uno de los perfiles como la detección precoz del trastorno a través de diferentes cuestionarios, el seguimiento por parte de un equipo médico de la evolución de los afectados y el tratamiento de funciones cognitivas como la atención o la memoria.

III-C. Juegos integrados en la aplicación

En la aplicación móvil propuesta se han incluido tres juegos principales que dotan de ayuda en el correcto desarrollo de las actividades cognitivas de los niños/as (Figura 2). De forma general, el juego en los niños enriquece el proceso de enseñanza y aprendizaje, por lo que es importante usarlo para un beneficio educativo. La clasificación de actividades desarrolladas en la aplicación incluyen juegos de colores, cartas y matemáticas.

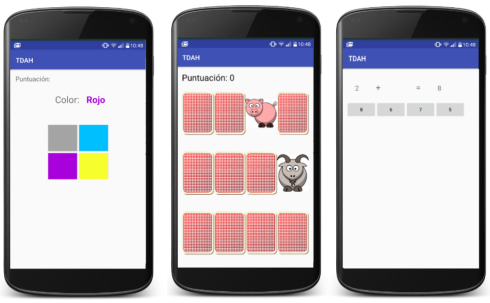


Figura 2. Juegos integrados en la aplicación

En primer lugar, el juego de colores se basa en la identificación de texto y su color. Su función es mostrar al usuario una palabra donde se lee un color determinado, este texto se muestra con un fondo de color distinto y los usuarios deben identificarlo entre las opciones propuestas. La eficacia de esta actividad ha sido estudiada en psicología, donde el efecto Stroop se aplica a la inferencia semántica en el tiempo de reacción de una tarea [15].

En segundo lugar, se propone un juego dirigido a mejorar la memoria de los niños de una manera divertida en la que también se practica la concentración y la atención. Su función principal es retener imágenes y evocar estímulos que desarrollan y fortalecen la memoria. Se muestra un conjunto de cartas boca abajo en el tablero y el usuario debe pulsar sobre cada una de ellas hasta que encuentre su pareja, eliminando pares de cartas en cada acierto.

Finalmente, también se ha incluido un juego basado en operaciones matemáticas, destinado a desarrollar el pensamiento matemático en los afectados por el TDAH. El juego consiste en varias operaciones matemáticas que van cambiando a medida que el niño acierta.

En cada uno de los juegos se han establecido diferentes niveles que permiten el progreso y el avance en su tratamiento. Además, este progreso se representa en la aplicación mediante estadísticas y gráficas que permiten observar la evolución del afectado, mostrando los resultados obtenidos en las distintas actividades.

IV. Seguridad

En el siguiente punto se realiza un estudio y un análisis de la seguridad necesaria que se debe aplicar con respecto a la información médica que se trata en esta aplicación móvil.

IV-A. Protección de datos

El m-health permite realizar, a través de dispositivos móviles, procedimientos médicos como diagnósticos, exámenes o supervisión de tratamientos. El uso de dispositivos móviles para un tratamiento médico requiere responsabilidad en el cumplimiento de la protección de información médica. Es necesario encontrar un equilibrio entre dos conceptos tan opuestos como la garantía de acceso a la información y la protección de la confidencialidad y privacidad de los usuarios. Además es importante el uso de un modelo de seguridad que asegure una protección para la información médica y que se adapte a la nueva normativa de protección de datos bajo la nueva ley GDPR (General Data Protection Regulation)[16].

En particular, los datos sanitarios se consideran en el más alto nivel de protección, se trata de información sensible y su protección es de gran importancia [17]. Esta característica es fundamental para establecer un buen cumplimiento con la nueva ley de protección de datos. Se pretende reforzar y unificar la protección de datos para todos los individuos dentro de la Unión Europea y como objetivo principal devolver el control a los ciudadanos y residentes sobre sus datos personales. Al tratarse de datos médicos y sanitarios, la aplicación móvil desarrollada cumple una serie de requisitos para que los usuarios del sistema estén informados del uso de sus datos personales. Se incluye una opción de consentimiento en la cual el usuario puede acceder a toda la información de uso acerca del acceso a su información personal.

La finalidad del tratamiento de los datos está bien definida, ya que no se recoge ningún dato que no se utilice en la aplicación y todos los datos recogidos se utilizan únicamente para las finalidades indicadas. De este modo, el usuario está informado en todo momento de su uso, se utilizan las medidas necesarias para dotar de una capa de seguridad en el tratamiento de sus datos personales.

IV-B. Comunicaciones seguras

En el desarrollo del sistema propuesto, se ha tomado la seguridad en las comunicaciones como uno de los principales temas a abordar en la seguridad del sistema. Con el objetivo de que todo el proceso de comunicación sea gestionado de forma segura, a la hora de hacer dichos usos y transacciones a través de Internet, han de tomarse una serie de medidas y buenas prácticas encaminadas a mejorar la seguridad. El protocolo HTTPS se utiliza para permitir introducir el nivel de cifrado SSL (Secure Sockets Layer) en el sistema, donde se cifran las solicitudes y respuestas que se intercambian entre el cliente y el servidor.

IV-C. Tokens de autenticación

Las técnicas de autenticación de usuario aplicadas garantizan el origen del mensaje, mientras que los esquemas de protección de la integridad utilizados garantizan que los

mensajes no se han alterado durante la transmisión, y las técnicas de cifrado implementadas protegen la confidencialidad de la información transmitida y almacenada. Los tokens de autenticación se utilizan en este sistema para garantizar la autenticidad de los datos emitidos en el intercambio de información entre la aplicación y el servidor, son contenedores de información sobre la autenticación de usuario. En general, el proceso es el siguiente: cuando un usuario inicia sesión y envía sus datos al servidor, este genera un token con el secreto para posteriormente enviar este token generado al usuario. A continuación, en cada solicitud realizada se envía este token. El servidor comprueba la firma del token y verifica su información descifrándola con el secreto. Después de esto, si todo es correcto, se envía la información solicitada.

El estándar utilizado en el sistema es el JSON Web Token (JWT) a través del algoritmo HS256 que hace uso de la función hash SHA256. El token está compuesto por tres campos generales: $header$, que contiene el algoritmo utilizado y el tipo de token, $payload$ con atributos que identifican al usuario, así como la fecha de caducidad del token y su validez y $verifySignature$ que contiene los elementos anteriores cifrados con la clave secreta para la verificación del token. A través de este procedimiento, es posible controlar la autenticidad y la información que solicita consultar cada uno de los usuarios que acceden al sistema.

IV-D. Directrices de seguridad en el sistema

La autenticidad, la integridad y la confidencialidad son las tres propiedades principales de la información que han sido protegidas en esta aplicación utilizando primitivas de seguridad en el sistema. En el desarrollo de esta aplicación se han seguido las directrices establecidas por las pautas OWASP (de sus siglas en inglés Open Web Application Security Project) para la autenticación de usuarios. Se trata de una guía de normas y directrices destinadas a mejorar la seguridad de las aplicaciones, aportando soluciones de seguridad. En el desarrollo de la aplicación presentada en este trabajo se han tenido en cuenta las normas establecidas en esta guía. Las directrices utilizadas incluyen el uso de tokens de autenticación de usuario y solicitudes HTTPS, tal como se indica en el apartado anterior. Además de esto, los mecanismos de autenticación se realizan siempre en el lado del servidor y no en el lado del cliente. Del mismo modo, para dotar de mayor seguidad en la autenticación de los usuarios se realiza una comprobación para evitar el uso de contraseñas débiles, estableciendo rutinas de cifrado para generar contraseñas suficientemente fuertes.

IV-E. Comunicación de los perfiles del sistema

En la aplicación propuesta se ha utilizado un chat en directo para establecer una comunicación en tiempo real entre los miembros del equipo médico con los familiares o con los niños/as afectados. Para ello se hace uso de un cifrado jerárquico basado en identidad HIBE (de sus siglas en inglés Hierarchical Identity Based Encryption) [18] que permite el intercambio seguro de información en las comunicaciones entre los usuarios del sistema. De este modo, se mejora la escalabilidad del esquema general de cifrado facilitando la generación de claves.

En este sistema se utiliza el email como parámetro de identidad para los usuarios, puesto que es el utilizado para el proceso de autenticación en la aplicación propuesta. Las operaciones utilizadas en este esquema son las siguientes:

- $Setup$ (k). Este algoritmo genera los parámetros de inicialización del sistema y las claves del primer nivel, k es la máxima profundidad de jerarquía. Se genera una clave secreta maestra (msk) y a partir de esta una clave pública maestra (mpk), que va a ser un conjunto de parámentros.
- $KeyGen$ ($sk_{ID_{|k}}$, mpk, $ID_{|k+1}$). Con esta función se generan las claves de cada usuario. Para ello se obtiene el nivel de profundidad del usuario. Si pertenece a un nodo padre, es decir, si es el padre del niño o el médico responsable del paciente se inicializan los parámentros igualando a la msk de la forma $sk_{ID_k} = msk$. Si el usuario es un paciente o alguna persona del equipo médico al que se le ha delegado el caso clínico se genera la clave $sk_{ID_{k+1}}$.
- $Encrypt$ ($ID_{|k}$, M, mpk). Para cifrar los mensajes se utiliza esta función que genera como salida el texto cifrado (CM) a partir de la identidad, un mensaje M y los parámetros de la clave pública mpk como entrada.
- $Decrypt$ ($sk_{ID_{|k}}$, CM, mpk). El algoritmo de descifrado es el encargado de obtener el mensaje original M utilizando la clave privada $sk_{ID_{|k}}$, el mensaje cifrado (CM) y los diferentes parámetros que se obtienen de la clave pública mpk.

A nivel general, la metodología utilizada para una comunicación entre el usuario A y B es la siguiente: en primer lugar se genera la clave privada para A mediante el algoritmo $Setup$. Mediante el algoritmo $KeyGen$, se produce la clave privada correspondiente al usuario $B(ID_i)$. El algoritmo de cifrado $Encrypt$ cifra los mensajes utilizando el conjunto de datos almacenados en la mpk y el algoritmo de descifrado $Decrypt$ descifra el texto cifrado utilizando la clave privada de B. El sistema también admite la delegación en la que los usuarios pueden recibir claves privadas restringidas que sólo permiten la delegación a una profundidad limitada (Figura 3).

V. CONCLUSIONES

En general, como conclusión del sistema propuesto se ha desarrollado una aplicación móvil que integra los principales perfiles implicados en el TDAH, incorporando un perfil para cada uno de los roles: familiares, equipo médico, profesorado y pacientes con las funciones y permisos correspondientes. La implementación de la aplicación, incluyendo la gamificación, ha sido el enfoque principal para el tratamiento de este trastorno. En este sentido, se incluyen varias actividades de lógica y memoria a las que se puede acceder a través del perfil del paciente. El equipo médico realiza un seguimiento a través de los resultados y progresos obtenidos, generando un diagnóstico provisional. Por otro lado, los tutores pueden acceder al sistema de prediagnóstico mediante la integración de un cuestionario y también realizar un seguimiento de los resultados de su hijo.

Además de las funcionalidades desarrolladas en la aplicación, se han tenido en cuenta las medidas necesarias para conseguir unas comunicaciones seguras. Dado que los datos

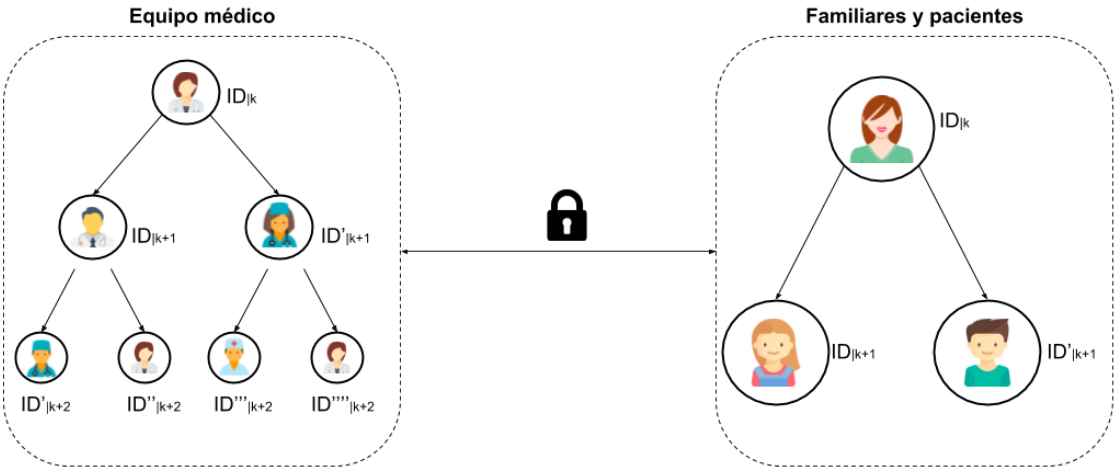


Figura 3. Esquema de comunicación entre los usuarios

de los usuarios de la aplicación propuesta son en gran medida información médica, que requieren un alto nivel de protección, es importante que no se vean comprometidos en ningún caso. Así, en el desarrollo del sistema, se han implementado diferentes medidas criptográficas para proporcionar a la aplicación de autenticidad para los usuarios, privacidad y confidencialidad de los datos.

El sistema también ha incluido un método para generar consultas a través de un protocolo seguro. En cuanto a la autenticación, se han utilizado métodos basados en tokens para garantizar la autenticidad de las consultas al servidor y para controlar el acceso a la API. El sistema ha seguido las reglas de OWASP para cumplir con la guía de buenas pautas de seguridad en las aplicaciones. Además, se ha aplicado el cifrado jerárquico basado en identidad para cifrar las comunicaciones entre los usuarios del sistema. En una futura versión de este documento, se incluirán características de seguridad adicionales.

Como trabajo futuro, se plantean nuevas mejoras en la funcionalidad de la aplicación y en el protocolo de seguridad utilizado para las comunicaciones entre los usuarios del sistema. Principalmente se plantea la aplicabilidad de permisos revocables a los usuarios en las comunicaciones mediante el protocolo R-HIBE. Del mismo modo, se profudizará en el desarrollo de una aplicación web que permitirá la visualización en detalle de los avances de los pacientes por parte del equipo médico y ayudar en su tratamiento. Además de esto, al tratarse de una colaboración con el Hospital Universitario de Canarias, como trabajo futuro también se incluye la prueba de la aplicación en un entorno real.

Agradecimientos

Investigación subvencionada por el Ministerio de Economía y Competitividad de España, el fondo Europeo FEDER, la Fundación CajaCanarias, TESIS2015010102, RTC-2014-1648-8, MTM2015-69138-REDT, DIG02-INSITU y TEC2014-54110-R.

Referencias

[1] B. E. Rodillo, "Trastorno por déficit de atención e hiperactividad (tdah) en adolescentes," *Revista Médica Clínica Las Condes*, vol. 26, no. 1, pp. 52–59, 2015.

[2] S. B. Campbell, "Attention-deficit/hyperactivity disorder," in *Handbook of developmental psychopathology*. Springer, 2000, pp. 383–401.

[3] E. Cardo and M. Servera-Barceló, "Prevalencia del trastorno de déficit de atención e hiperactividad," *Revista de neurología*, vol. 40, no. 1, pp. 11–15, 2005.

[4] M. L. East and B. C. Havard, "Mental health mobile apps: from infusion to diffusion in the mental health social system," *JMIR mental health*, vol. 2, no. 1, 2015.

[5] M. L. Lorusso, E. Biffi, M. Molteni, and G. Reni, "Exploring the learnability and usability of a near field communication-based application for semantic enrichment in children with language disorders," *Assistive Technology*, vol. 30, no. 1, pp. 39–50, 2018.

[6] C. Herodotou, "Mobile games and science learning: A comparative study of 4 and 5 years old playing the game angry birds," *British Journal of Educational Technology*, vol. 49, no. 1, pp. 6–16, 2018.

[7] A. S. Juarascio, S. M. Manasse, S. P. Goldstein, E. M. Forman, and M. L. Butryn, "Review of smartphone applications for the treatment of eating disorders," *European Eating Disorders Review*, vol. 23, no. 1, pp. 1–11, 2015.

[8] C. G. Fairburn and E. R. Rothwell, "Apps and eating disorders: A systematic clinical appraisal," *International Journal of Eating Disorders*, vol. 48, no. 7, pp. 1038–1046, 2015.

[9] J. Nicholas, M. E. Larsen, J. Proudfoot, and H. Christensen, "Mobile apps for bipolar disorder: a systematic review of features and content quality," *Journal of medical Internet research*, vol. 17, no. 8, 2015.

[10] J. R. Rachels and A. J. Rockinson-Szapkiw, "The effects of a mobile gamification app on elementary students' spanish achievement and self-efficacy," *Computer Assisted Language Learning*, vol. 31, no. 1-2, pp. 72–89, 2018.

[11] B. Martínez-Pérez, I. De La Torre-Díez, and M. López-Coronado, "Privacy and security in mobile health apps: a review and recommendations," *Journal of medical systems*, vol. 39, no. 1, p. 181, 2015.

[12] A. K. Jain and D. Shanbhag, "Addressing security and privacy risks in mobile applications," *IT Professional*, vol. 14, no. 5, pp. 28–33, 2012.

[13] L. Casati and A. Visconti, "The dangers of rooting: Data leakage detection in android applications," *Mobile Information Systems*, vol. 2018, 2018.

[14] R. N. Akram, H. Chen, J. López, D. Sauveron, and L. T. Yang, "Security, privacy and trust of user-centric solutions," *Future Generation Comp. Syst.*, vol. 80, pp. 417–420, 2018.

[15] J. López-Villalobos, I. Serrano-Pintado, J. L. Andrés-De, J. Sánchez-Mateos, S. Alberola-López, and M. Sánchez-Azón, "Usefulness of the stroop test in attention deficit hyperactivity disorder," *Revista de neurologia*, vol. 50, no. 6, pp. 333–340, 2010.

[16] GDPR, "General data protection regulation," 2018. [Online]. Available: https://www.eugdpr.org/

[17] A. Papageorgiou, M. Strigkos, E. Politou, E. Alepis, A. Solanas, and C. Patsakis, "Security and privacy analysis of mobile health applications: The alarming state of practice," *IEEE Access*, vol. 6, pp. 9390–9403, 2018.

[18] D. Boneh, X. Boyen, and E.-J. Goh, "Hierarchical identity based encryption with constant size ciphertext," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2005, pp. 440–456.

Arquitectura de gestión dinámica de riesgos basada en ontologías y reglas de comportamiento

Irene Romero
Departamento de Sistemas Telemáticos
Univ. Politécnica de Madrid
iromero@dit.upm.es

Raúl Riesco
National Cybersecurity Institute of Spain (INCIBE)
raul.riesco@incibe.es

Francisco Barea
Departamento de Sistemas Telemáticos
Univ. Politécnica de Madrid
jfbarea@dit.upm.es

Víctor A. Villagrá
Departamento de Sistemas Telemáticos
Univ. Politécnica de Madrid
villagra@dit.upm.es

***Resumen*—La elevada dependencia de las organizaciones con respecto a sus sistemas de información hace que el análisis de riesgos de la mismo sea muy valioso. Debido a la rapidez con la que evolucionan las amenazas, los enfoques de análisis de riesgo estáticos pierden validez al no actualizarse en tiempo real ante los eventos o la nueva información que pueda aparecer. En este artículo se propone un modelo de arquitectura para la gestión del riesgo dinámico que incorpora inteligencia haciendo uso de ontologías, reglas de inferencia y probabilidades condicionadas. Asimismo, para hacer frente a la naturaleza cambiante de las amenazas, es necesario el intercambio de información sobre estas en tiempo real. Para ello, en lugar de compartir indicadores de compromiso efímeros (como podrían ser indicadores basados en firmas) se propone el intercambio de la propia inteligencia en forma de reglas de comportamiento.**

***Index Terms*—gestión de riesgo dinámico, ciberseguridad, inteligencia de amenazas , ontologías, STIX.**

I. Introducción

La dependencia que tienen las organizaciones con respecto a sus sistemas de información es cada vez mayor y, al convertirse estos en activos indispensables, implica que puedan ser el objetivo de ataques e interferir con la misión de la organización. Por ello, es necesario orientar esfuerzos hacia la defensa y seguridad de estos elementos; así como la evaluación del riesgo de los mismos de cara a ganar una conciencia situacional lo más completa posible.

Un análisis de riesgos (AARR) es el proceso sistemático utilizado para estimar la magnitud de los riesgos a los que está expuesta una organización. Debe formar parte de la actividad continua de gestión de la seguridad. En general, las metodologías de análisis y gestión de riesgos forman parte de conceptos similares (que se pueden ver en la Fig. 1).

El proceso de un análisis y gestión de riesgos comprende un ejercicio de comprensión, catalogación y valoración de aspectos que adquieren gradualmente una complejidad sustancial. Un AARR ha de ser riguroso y permitir ser contrastado de manera objetiva, de otro modo se introduciría un sesgo que condicione las decisiones basadas en sus resultados.

Teniendo en cuenta que los sistemas de información cambian continuamente, la necesidad de metodologías de AARR dinámicas (y no estáticas) es innegable. La dinamicidad de dichas metodologías permite a los sistemas actualizar su nivel de riesgo en tiempo real, lidiando con eventos a medida que se producen y la naturaleza cambiante de las amenazas. Sin embargo, las metodologías estáticas no cubren estas necesidades al no ser un proceso continuo, sino la repetición del proceso en intervales discretos y regulares de tiempo.

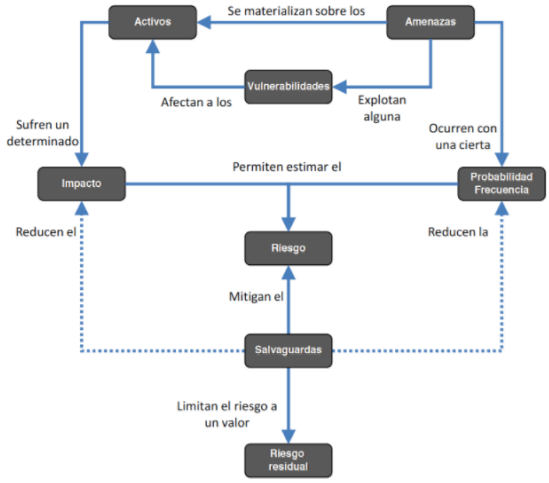


Figura 1. Conceptos del análisis y gestión de riesgos

Además, el análisis eficiente de la información de la Cyber Threat Intelligence (CTI) compartida es crucial para la evaluación de riesgos de red. Esto es debido a que enfoque estático de la seguridad tradicional basada no está a la altura de la nueva generación de amenazas caracterizada por su naturaleza dinámica, compleja y resiliente. Para hacer frente a esto es necesario conseguir y compartir información sobre las amenazas en tiempo real.

La Threat Ingelligence (TI) engloba todo el conocimiento que se posee sobre las posibles amenazas para poder tomar las decisiones pertinentes. Cuando la información compartida es técnica, es donde cobran importancia los indicadores de compromiso (IOC). Sin embargo, los esquemas basados en IOC son poco eficientes, pues se basan en firmas (que son efírmeras). Si fuera posible identificar los TTP (Tecnicas, Tacticas y Procedimientos) o las herramientas utilizadas, etc. se podría evolucionar hacia un enfoque más eficiente, caracterizando y contextualizando procesos por ejemplo para la distribución de malware, técnicas *fileless*, etc. Estándares como STIX™, CybOX™ o TAXII™ comienzan a ganar fuerza de cara al intercambio de información de amenazas, pues proporcionan una estructura para indicadores de ciberseguridad, caracterización de amenazas y diferentes opciones para el intercambio de la información.

En este artículo se presenta una propuesta teórica dentro del marco del análisis y la gestión del riesgo dinámico. Así,

se presenta un detallado estudio del arte en la sección II, identificando las carencias de las metodologías existentes. En la sección IV se presenta un modelo simplificado y a alto nivel para gestión de riesgos dotado con inteligencia, basado en ontologías y reglas de comportamiento; en el que se propone compartir información sobre amenazas y riesgo. Por último en la sección V se presentan las conclusiones y líneas futuras de esta propuesta.

II. Estado del arte

A lo largo de esta sección se repasan los mecanismos que emplea cada metodología estudiada para el cálculo del riesgo de una organización.

Antes de explorar las metodologías concretas estudiadas, cabe comentar la norma ISO 27005 [1], el estándar internacional que se ocupa de la gestión de riesgos de seguridad de informarción suministrando directrices aplicables a toda organización que tenga intención de gestionar los riesgos que puedan complicar la seguridad de la información de su organización. Dicha norma no facilita una metodología concreta de AARR, sino que describe el proceso recomendado para un análisis en una organización de cualquier tipo. También la norma ISO31000, que propone un nuevo concepto de riesgo (dejando atrás la visión clásica del riesgo como la probabilidad del evento y sus consecuencias) como la incertidumbre vinculada a los objetivos de la organización [2]. Además, la norma ISO31000 introduce como concepto el çontexto"de una organización. Este contexto tiene en cuenta variables culturales, políticales, sociales, tecnológicas, económicas, etc. de la organización; considerando qué factores pueden afectar dichas variables.

A continuación, se divide el análisis en metodologías estáticas y dinámicas; clasificándo estas últimas en grupos en función del enfoque que utilizan.

II-A. Metodologías de análisis de riesgo estáticas

Las metodologías estáticas tienen como resultado un valor del riesgo en tiempo discreto, es decir, calculan el riesgo para un instante concreto, sin tener en cuenta las posibles variaciones de las condiciones de las que depende este cálculo.

En primer lugar, y mirando hacia el campo de las tecnologías de la información en España, destaca MAGERIT. MAGERIT [4] hace uso de los conceptos básicos de los AARR, siendo imprescindible identificar los aspectos de seguridad de los activos, conocer los vulnerabilidades de cada uno de ellos y valorar el impacto de un ataque potencial sobre ellos. MAGERIT hace uso de la definición de riesgo clásica que tiene en cuenta el impacto y la probabilidad de las amenazas.

Por otro lado, la metodología más promovida internacionalmente (por el Instituto Internacional de Estándares y Tecnología, EEUU) es la NIST SP 800-30 [5], basada en un proceso de nueve pasos a realizar consecutivamente para la determinación del riesgo de una organización. De nuevo, se fundamenta en la idea del riesgo como función entre la probabilidad de ocurrencia de una amenaza que explota una vulnerabilidad y el impacto de la misma. También desde EEUU, la metodología OCTAVE [6], que se basa en que los activos de la organización a analizar son la información.

Entre otras metodologías estáticas relevantes destacan CRAMM y Mehari. CRAMM [7] está desarrollada por el gobierno británico, se desarolla en tres etapas: definición global de objetivos de seguridad, análisis de riesgos identificando amenazas y vulnerabilidades, y selección de medidas de seguridad aplicadas. Mientras que Mehari [8], propuesta por el Club de la Seguridad francés (CLUSIF), es una guía de implantación de la seguridad en una entidad a lo largo del ciclo de vida. Evalúa los riesgos en base a los criterios de disponibilidad, integridad y confidencialidad.

Todas estas metodologías estáticas abordan el problema de la cuantificación del riesgo, haciendo uso de la definición clásica del mismo (impacto multiplicado por probabilidad).

Otras metodologías estáticas son STOPE [9], que propone proceso que adopta las cinco fases cíclicas del modelo DMAIC (definir, medir, analizar, mejora y control), u OSSF [10], diseñado teniendo en mente características especiales de servicios online así como el ciberespacio. Para la determinación de la probabilidad y los niveles de impacto, OSSF propone la metodología OWASP.

II-B. Metodologías de análisis de riesgos dinámicas

Este tipo de metodologías obtienen el riesgo en tiempo real. A continuación se exponen las más relevantes en el estado del arte según distintos enfoques.

II-B1. Enfoques basados en métricas: La definición clásica de riesgo, como ya se ha comentado, es la multiplicación de la probabilidad y el impacto. Muchas de las propuestas hacen uso de esta definición combinándola con otras técnicas, como sucede en [11] donde la gestión de riesgos se basa en sistemas multi-modelo de predicción para control industrial.

Otro enfoque que fundamenta su base en el uso de métricas es BlueGen [12]. Este algoritmo define dos modelos para el análisis: modelo de entorno, que describe el nivel de amenaza, y el catálogo de referencia, un conjunto de capacidades de atacantes, capacidades de mitigación y tipos de organización; y un conjunto de relaciones entre ellos. El riesgo de cada misión es función de la exposición de la organización y la criticidad de la misión.

Otro enfoque se puede encontrar en [14], donde se define una métrica de riesgo temporal, que tiene en cuenta los nuevos eventos de seguridad que vayan sucediendo y las respuesta seleccionadas. Esto quiere decir que la exposición al riesgo general de una situación dada depende de la secuencia completa de acciones tomadas para recuperarse de ella. Así, el riesgo depende del número de activos afectados, el número de eventos, el daño producido a la organización en un instante concreto y la probabilidad de ocurrencia de un exploit por un evento de vulnerabilidad en un determinado instante.

En cuanto a algoritmos orientados a métricas, se pueden encontrar IDAM&IRS [13] y el definido por Stakhanova en [26].

IDAM&IRS tiene por objetivo establecer una estrategia de respuesta basado en el riesgo existente en el escenario de intrusión. Para ello, define la métrica *risk index* en función de la confianza en la alerta, el número de alertas (total y del mismo tipo) recibidas en un slot de tiempo, el grado de vulnerabilidad del sistema atacado y la severidad del sistema. Compara este nivel con los umbrales definidos y en función de esto, decide que acción de respuesta es la mejor.

Por otro lado, el método de selección de la respuesta propuesto por Stakhanova consta de tres fases para cada

una de las cuales se define una métrica de seguridad. La primera fase del algoritmo define la métrica de probabilidad de ocurrencia, que permite determinar en qué momento es necesario reaccionar frente a una intrusión. En segundo lugar, se define la métrica de reducción de daño, que consiste en evaluar el daño y coste de la respuesta, las respuestas acordes con esta métrica se seleccionan como candidatas para la última fase. Finalmente, la métrica de máximo beneficio al menor riesgo será la encargada de de seleccionar la respuesta óptima en la fase final del algoritmo

II-B2. Enfoques probabilísticos y de predicción: Una de las propuestas de cara a un plantear un modelo dinámico de AARR es el empleo de grafos bayesianos (BAG, *Bayesian Attack Graph*). En [15] se realiza el análisis representando en un grafo las amenazas y su impacto en la red. El juicio sobre la estimación de riesgo asociado a cada amenaza es subjetivo y depende la probabilidad de que una amenaza se active y la dificultad de la misma. A medida que se den incidentes en la red, se modifica la probabilidad usando técnicas de inferencia Bayesianas, forward and backward propagation. La propagación hacia delante modifica la probabilidad de los atributos sucesores que están directamente influenciados por las evidencias. La propagación hacia atrás, ajusta la hipótesis inicial sobre todos los atributos. Posteriormente, las probabilidades posteriores (probabilidades incondicionales actualizadas) reflejan la probabilidad de otros resultados basado en los eventos detectados.

Las redes Bayesianas también se emplean en [16], ilustrado con una prueba de concepto basada en una inyección SQL, y en [17], donde la métrica se basa en el producto de la probabilidad condicional de una amenaza, la severidad de cada amenaza y un peso fijado. En este último, la probabilidad de una determinada amenaza se calcula en base a los logs de tráfico que se podrían obtener de diferentes componentes de seguridad, como un IDS. Por último, reglas Bayesianas se aplican también en [18] para calcular la probabilidad de que una amenaza afecte a un activo en particular, tras haber realizado un estudio de los activos identificados, las amenazas potenciales y su impacto en las diferentes dimensiones.

El sistema propuesto en [19] también basa su análisis en la definición de un grafo; esta vez, a partir de los informes recibidos por los detectores de intrusiones. A continuación, el grafo se transforma en modelos de Markov que tienen en cuenta la evolución de los ataques en curso y la evolución del estado del sistema objetivo. Existen otras propuestas que hacen uso de los modelos de Markov. Es el caso, por ejemplo, de [27], que presenta un IPS distribuido con capacidad de predecir niveles de amenaza a partir de un Modelo Oculto de Markov (HMM) y estimar riesgos sobre los activos mediante lógica difusa. En concreto infiere el riesgo sobre los activos en función del estado (Normal, Intento de Intrusión, Intrusión en Progreso, Ataque Exitoso). En la misma línea se mueve el modelo desarrollado en [28].

El uso de probabilidades condicionadas también se presenta en [20], donde se modelan los ataques en dos fases, en las que el atacante primero accede a un sistema (brecha) conectado a internet para después escalar a sistemas internos explotando vulnerabilidades locales. Asumiendo que el riesgo se expresa como la probabilidad por el impacto, dicha probabilidad se expresa en función de la probabilidad de ocurrencia de un ataque y la probabilidad condicionada de comprometer el sistema, habiendo exisstido un ataque; términos que dependen principalmente de las vulnerabilidades del sistema y las vulnerabilidades que explotan los *toolkits* de atacantes.

El problema de los modelos probabilísticos presentados basados en el teorema de Bayes es que modelan fases de los ataques sobre la entidad y no sobre información recibida de entidades externas, como se propone en este artículo para solventar este problema.

Otro enfoque probabilístico es el PRA [21], *Probability Risk Assessment*, que pretende cuantificar la probabilidad de que una amenaza suceda. Para ello, se identifican en primer lugar los servicios esenciales y no esenciales para la misión de la organización. A continuación, se enumeran las amenazas a los servicios esenciales y se cuantifica el riesgo (valor medio del daño esperable) para cada escenario de intrusión identificado.

II-B3. Otros enfoques: Otro enfoque que no tiene cabida en las secciones anteriores es el propuesto en [22], donde la gestión de riesgos se basa en la construcción de una matriz (tanto desde una perspectiva empresarial como técnica). Para determinar los valores con los que completar la matriz, los autores proponen la creación de una encuesta para medir diferentes indicadores clave (dependencia tecnológica, impacto de la pérdida de información, etc.) que se enviará a las personas a cargo de la gestión de cada aplicación.

La representación gráfica también puede ser de utilidad en este campo, como sucede en [23] donde se asume un entorno cambiante, ilustrado utilizando diagramas con bucles causales de los principales causantes de riesgos. En [24] se propone la representación de las dependencias entre vulnerabilidades en un grafo acíclico, complementada con probabilidades condicionales de ataque e impacto.

La epidemiología es una disciplina que propone conceptos aplicables a este problema, en concreto la teoría de peligro de Matzinger [25]. En [14] se propone la aplicación del concepto *distress signal* como un modificador del riesgo (calculado previamente con el impacto y la probabilidad. Esta probabilidad se calcula haciendo uso de un sistema inteligente basado en redes neuronales).

III. Ontologías

Una ontología es una herramienta formal y explícita de representar conceptos, su significado y sus relaciones. Es explícita porque define conceptos, propiedades, relaciones, funciones, taxonomías, axiomas, reglas y restricciones. Es formal porque está definida por un lenguaje interpretable por máquinas.

En [29], Herzog *et al.* definen una ontología genérica para el dominio de la seguridad, modelando conceptos como activos, amenazas, vulnerabilidades, contramedidas y las relaciones entre ellos. En [30], Fenz *et al.* contribuyen con ontologías para un análisis cuantitavo de riesgos, teniendo en cuenta el daño causado por ciertas amenazas, costes y tiempo de recuperación. En [31], los autores integran el concepto de ontología para riesgo de procesos de negocio, basándose en NIST. La comunicación entre la ontología de seguridad y el motor de riesgo utiliza XML. Por último, en [32] se propone

el uso de ontologías basadas en STIX y reglas para el cálculo del impacto de una amenaza en una organización.

IV. Arquitectura propuesta

La arquitectura propuesta en este artículo forma parte de un trabajo en desarrollo actualmente, por lo que se describirá a alto nivel (sus módulos, entradas, bases de datos, etc.) pero no la implementación concreta a bajo nivel.

Podemos dividir la arquitectura en dos módulos:

- Análisis de riesgos dinámico: Inteligencia de riesgo e intercambio de información de riesgo. En este módulo se calcula el valor de riesgo, para lo que se valoran los activos, servicios o procesos de negocio en base a las dimensiones (C, I, D), se analizan las dependencias y se hereda la valoración de activos. También se realiza en esta fase el análisis de amenazas, incluyendo el concepto de probabilidad e impacto.
- Gestion y tratamiento del riesgo dinámico: Una vez calculado el riesgo, se pasa a la fase gestión y tratamiento del riesgo, donde se definirá la politica a implementar para mitigarlo.

Además del proceso de análisis de riesgos y gestión del mismo; ha de existir un conjunto de sensores que alimentan dicho modelo. Estos sensores manejarán diferentes tipos de información (sobre ciberataques, riesgos físicos, riesgos humanos, etc.). Han de comunicarse entre ellos e contarán con inteligencia para una mejor definición del contexto del sistema. Estos sensores además permitirán determinar la fiabilidad del dato y filtrar eventos que no aporten valor.

También es necesario contar con una base de conocimiento donde almacenar las evidencias que permitan realizar los cálculos de probabilidad para el riesgo, así como acciones de respuesta para tratamiento del riesgo, su efectividad, etc.

En este artículo se presenta el diseño de la parte del análisis de riesgos, dejando el diseño de la parte de gestión como trabajo futuro.

En el análisis de riesgo distinguimos entre tres procesos diferentes: el intercambio de información de riesgo, la inteligencia de riegos y el análisis de riesgo, proporcionando un valor de riesgo.

IV-A. Intercambio de información de riesgo

La arquitectura estará abierta a cualquier fuente de datos, siempre que esté estructurada y cuente con cierta lógica.

Como ya se ha comentado, basar una arquitectura de AARR únicamente en IOC no es eficiente pues son indicadores efímeros. De esta forma, se propone tratar de identificar los TTPs. Los TTPs tienen un problema, su formalización; pero se propone modelarlos usando un lenguaje de reglas semánticas -como SWRL- que nos permita identificar conceptos y las relaciones entre ellos.

La arquitectura propuesta, una vez reciba la información, realizará un proceso de parseo o transformación al formato STIX™, en caso de que no se encuentre ya en dicho formato. STIX™ es un lenguaje para la especificación, captura, caracterización y comunicación de información estandarizada sobre ciberamenazas. Lo hace de forma estructurada para facilitar los procesos de gestión de ciberamenazas y la automatización de los mismos. Este estándar proporciona un mecanismo común para abordar la información estructurada de ciberamenazas a través de esta gama completa de casos de uso que mejora la coherencia, la eficiencia, la interoperabilidad y la conciencia situacional general.

La útlima versión de STIX™, 2.0, utiliza el formato JSON; sin embargo, se propone en este caso el uso de una ontología para la representación de esta información. Esta ontología resolvería cuestiones no resueltas aún por STIX, pudiendo describir situaciones mediante reglas y relaciones (como TTP). Además, las ontologías resuelven el problema de la interoperabilidad entre las metodologías de gestión de riesgos, proporcionando un lenguaje común

Asímismo, además de los sensores que alimenten esta inteligencia de amenazas, se incluye en el proceso de AARR la información proveniente de bases de datos externas, como CVE, CPE, NVD, etc.

Se propone compartir las reglas definidas para la inteligencia de amenazas y de riesgos. La recepción de dichas reglas permitiría a la organización anticiparse a las amenazas antes de que sucedan. Así, en lugar de recibir un IOC (el enfoque clásico) podremos recibir el algoritmo que permita detectar un TTP o incluso un APT en la organización. Este enfoque requiere de una anonimización de la información y el uso del protocolo TLP (*Traffic Light Protocol*).

IV-B. Inteligencia de riesgos

Este módulo es el responsable de la inteligencia semántica relacionada con el riesgo. También es encargado del cálculo del contexto de la organización, incluyendo activos y dependencias entre ellos; partiendo de la información toda la información interna o externa que se ha compartido. Se basa en un proceso de razonamiento semántico que descubre nuevas amenazas, vectores de ataque o patrones anómalos.

IV-C. Análisis del riesgo

Este módulo es el encargado de proporcionar un nivel cuantitativo de riesgo en base al que se tomarán las contramedidas necesarias a decidir por el módulo de gestión y tratamiento del riesgo.

Este riesgo dependerá de las variables de probabilidad e impacto. Se propone de nuevo el uso de ontologías y probabilidades condicionadas, como en [33] donde las probabilidades condicionadas se basan en una base de datos con el histórico de eventos sucedidos y se calculan usando SWRL. El cálculo de la probabilidad es crucial pues las decisiones sobre contramedidas a implementar estarán basadas en este valor.

V. Conclusiones

La gran mayoría de metodologías actuales de análisis y gestión de riesgos adoptan un enfoque en el que miden de forma iterativa el estado de los activos de la organización así como los procesos de negocios. En base a este estado, se calcula el riesgo de la organización, basado principalmente en el juicio subjetivo de expertos. Además, en base a este riesgo se toman las decisiones al respecto de las contramedidas a desplegar en la organización.

Este enfoque pierde validez y efectividad debido a la rapidez con la que evolucionan las amenazas y su sofisticación, por lo que el concepto de gestión de riesgo dinámico es clave

para abordar el nuevo panorama de ciberseguridad ante el que nos encontramos.

Por otra parte, el intercambio de información relativa al riesgo entre organizaciones se puede explotar haciendo uso de una inteligencia de riesgos, complementada con otro tipo de datos como puede ser el contexto de la organización, e inferir un valor de riesgo más completo y fiable.

Tras un detallado estudio del estado del arte en el campo de la gestión de riesgos, se identifican áreas con debilidades. Este artículo propone de forma teórica una arquitectura dinámica novedosa que pretende solucionar algunas de las problemáticas identificadas en el horizonte actual. Esta nueva arquitectura se basa en el uso de inteligencia (implementada a través de ontologías) para relacionar los conceptos involucrados en un análisis de riesgos: amenazas, activos, impacto y probabilidad; calculando esta última a partir de probabilidades condicionadas.

Esta arquitectura soluciona también la ineficiencia que resulta de compartir indicadores de compromiso basados en firmas, siendo estas efímeras y perdiendo validez rápidamente. Lo que se propone compartir en este caso es la inteligencia en sí (y no la información sensible), permitiendo a organizaciones identificar comportamientos maliciosos que se hayan dado en otras organizaciones. Esta información ha de ser anonimizada y gestionada con el protocolo TLP.

Como línea futura principal se encuentra el diseño del módulo de tratamiento y gestión de riesgos, donde también tiene cabida el uso de ontologías y razonadores semánticos que identifiquen y seleccionen la respuesta óptima ante diferentes niveles de riesgo, como se propone en [34]. Este módulo además habría de evaluar la efectividad de la respuesta y realimentarlo al sistema, para que pueda ser de utilidad a la hora de la selección de la respuesta.

REFERENCIAS

[1] ISO/IEC 27005:2008, Information technology - Security techniques – Information security risk management, 2008.

[2] ISO 31000:2018, Risk Management - Guidelines, 2018.

[3] ITIL, «Fundamentos de la Gestión TI». Disponible en el sitio web: http://itil.osiatis.es/Curso_ITIL/Gestion_Servicios_TI/ fundamentos_de_la_gestion_TI/que_es_ITIL/que_es_ITIL.php

[4] MAGERIT versión 3, Metodología de Análisis y Gestión de Riesgos de los Sistemas de Información. Ministerio de Hacienda y Administraciones públicas, 2012.

[5] G. Stoneburner, A. Goguen, y A. Feringa, «Risk Management Guide for Information Technology Systems, Recommendations of the National Institute of Standards and Technology», jul. 2002.

[6] R. Caralli, J. Stevens, L. Young, y W. Wilson, «Introducing OCTAVE Allegro: Improving the Information Security Risk Assessment Process»,2007.

[7] Siemens - Insight Consulting, The Logic behind CRAMM's Assessment of Measures of Risk and Determination of Appropriate Countermeasures, 2005.

[8] A. Huerta, «Introducción al Análisis de Riesgos – Metodologías», «Deconstruyendo PILAR», «Análisis de Riesgos con PILAR», conjunto de publicaciones online disponible en: http://www.securityartwork.es/author/ahuerta/

[9] M. S. Saleh y A. Alfantookh: "A new comprehensive framework for enterprise information security risk management" en *Applied Computing and Informatics*, vol. 9, n. 2, pp. 107-118, 2011.

[10] J. Meszaros y A. Buchalcevova. Introducing OSSF: A Framework for Online Service Cybersecurity Risk Management en *Computers & Security*. vol.65, pp. 300-313, 2017.

[11] Q. Zhang, C. Zhou, N. Xiong, Y. Qin, X. Li, S. Huang. "Multimodel-Based Incident Prediction and Risk Assessment in Dynamic Cybersecurity Protection for Industrial Control Systems" en *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. vol. 46, pp. 1-16, 2015.

[12] T. Llanso "BluGen: An Analytic Framework for Mission-Cyber Risk Assessment and Mitigation Recommendation" en *Proceedings of the 50th Hawaii International Conference on System Sciences*. 2017.

[13] C.P. Mu, X.J. Li, H.K. Huang, y S.F. Tian, "Online Risk Assessment of Intrusion Scenarios Using D-S Evidence Theory", en *European Symposium on Research in Computer Security (ESORICS)*, pp.35-48, 2008.

[14] L. Beaudoin, N. Japkowicz, y S. Matwin "Autonomic computer network defence using risk state and reinforcement learning". *Cryptology and Information Security Series*, vol. 3, pp. 238-248, 2009.

[15] N. Poolsappasit , R. Dewri y I. Ray "Dynamic security risk management using bayesian attack graphs" en *IEEE Trans Dependable Secure Comput* vol.9, n. 1, pp. 61–74, 2012.

[16] D. Henshel "Modeling cybersecurity risks: Proof of concept of a holistic approach for integrated risk quantification" en *2016 IEEE Symposium on Technologies for Homeland Security (HST)*, pp. 1-5, 2016.

[17] M. S. K. Awan, P. Burnap, O. Rana y A. Javed, "Continuous Monitoring and Assessment of Cybersecurity Risks in Large Computing Infrastructures" en *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems* pp. 1442-1447.

[18] A. U. Khan, M. Oriol, M. Kiran, M. Jiang y K. Djemame, "Security risks and their management in cloud computing," en *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*, pp. 121-128, 2012.

[19] W. Kanoun, N. Cuppens-Boulahia, F. Cuppens, S. Dubus y A. Martin "Success likelihood of ongoing attacks for intrusion detection and response systems" en *Computational Science and Engineering, 2009. CSE'09. International Conference* vol. 3, pp. 83-91, 2009.

[20] L. Allodi y F. Massacci "Security Events and Vulnerability Data for Cybersecurity Risk Estimation" en *Risk Analysis*, vol. 37, pp. 1606–1627, 2017.

[21] C. Taylor, A. Krings and J. Alves-Foss. "Risk analysis and probabilistic survivability assessment (RAPSA): An assessment approach for power substation hardening" en *Proc. ACM Workshop on Scientific Aspects of Cyber Terrorism,(SACT)*, vol. 64, 2002.

[22] G. Ollard, E. Disson, G. Talens, S. Ducroquet "Proposition of a method to aid Security Classification in Cybersecurity context" pp. 88-95, 2016.

[23] S. Naumov y I. Kabanov, "Dynamic framework for assessing cyber security risks in a changing environment" en *International Conference on Information Science and Communications Technologies (ICISCT)*, pp. 1-4, 2016.

[24] W. Wu, R. Kang y Z. Li, "Risk assessment method for cybersecurity of cyber-physical systems based on inter-dependency of vulnerabilities" en *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 1618-1622, 2015.

[25] P. Matzinger "Essay 1: the Danger model in its historical context" en *Scandinavian journal of immunology*, vol. 54, n. 1-2, pp. 4-9, 2001.

[26] N. Stakhanova, S. Basu and J. Wong "A Cost-Sensitive Model for Preemptive Intrusion Response Systems" en *AINA* vol. 7, pp. 428-435, 2007.

[27] H. Kjetil, A. Ajith, y J.K. Svein, "DIPS: A Framework for Distributed Intrusion Prediction and Prevention Using Hidden Markov Models and Online Fuzzy Risk Assessment", en *Third International Information Assurance and Security Symposium (IAS)*, pp.183-190, 2007

[28] A. Årnes, K. Sallhammar, K. Haslum, T. Brekne, M.E. Gaup Moe, y S.J. Knapskog, "Real-time Risk Assessment with Network Sensors and Intrusion Detection Systems", en *International Conference on Computational Intelligence and Security (CIS-05)*, Xian, China, publicado en Springer LNCS vol.3801/3802, 2005.

[29] A. Herzog, N. Shahmehri, C. Duma, .An Ontology for Information Security.en *International Journal of Information Security and Privacy*, 2007.

[30] S. Fenz, A. Ekelhart, M. Klemen, E. Weippl, "Security Ontologies: Improving Quantitave Risk Analysis.en *International Conference on System Sciences (HICSS'07)*, 2007.

[31] S. Fenz, A. Ekelhart, M. Klemen, E. Weippl, .Ontology-based Generation os IT-Security Metrics.en *International Conference on System Sciences (HICSS'08)*, 2008.

[32] S. Qamar, Z. Anwar, M. Ashiqur Rahman, E. Al-Shaer, Bei-Tseng Chu, "Data-driven analytics for cyber-threat intelligence and information sharing", en *Computers & Security*, vol.67, 2017.

[33] H. Bulu, A. Alpkocak, P. Balci, Üncertainty modeling for ontology-based mammography annotation with intelligent BI-RADS Scoring.en *Computers in Biology and Medicine*, vol. 43, pp.301-311, 2013.

[34] V. Mateos, V. A. Villagrá, F. Romero, .Ontologies-Based Automated

Intrusion Response System.[en] *Computational Intelligence in Security for Information Systems 2010*, pp. 99-106, 2010.

Herramienta para la identificación de requisitos de seguridad en un Modelo de Desarrollo Seguro

José Carlos Sancho Núñez
Universidad de Extremadura
Escuela Politécnica
jcsanchon@unex.es

[1]Andrés Caro Lindo
[2]Lucio Fondón Terrón
Universidad de Extremadura
[1]andresc@unex.es
[2]lfondont@alumnos.unex.es

José Andrés Félix de Sande
Viewnext S.A.
CENIT Cáceres (PCTEx)
jafelix@viewnext.com

Abstract- **La seguridad en el desarrollo del software debe considerarse en todas las fases del ciclo de vida. Desarrollar software seguro nace por integrar actividades de seguridad desde fases tempranas. Sin embargo, la escasa importancia que, tradicionalmente, tiene la seguridad en el desarrollo del software comienza a originar graves problemas en las empresas de desarrollo. Afectando incluso a su reputación. Tener en cuenta la seguridad de forma preventiva reduce la posibilidad de ataques maliciosos. Esta contribución plasma el proceso de trazabilidad para identificar y generar requisitos o historias de seguridad que permitan la codificación segura. Se presenta el diseño de una herramienta, aún en desarrollo, que facilita y simplifica este proceso de obtención de requisitos. La herramienta se apoya en una batería de preguntas para detectar los aspectos de seguridad fundamentales. Se describen sus características y los puntos fuertes que la diferencian y mejoran con respecto a otras técnicas conocidas.**

Index Terms- **software seguro, seguridad en aplicaciones, riesgos de seguridad, requisitos de seguridad, ingeniería del software, desarrollo seguro.**

I. Introducción

La multitud de ciberataques que se producen en aplicaciones con éxito son debido a la falta de seguridad en su desarrollo. La explotación de vulnerabilidades produce un gran impacto en las empresas [1], afectando gravemente a sus procesos de negocio, y haciendo que la seguridad en el desarrollo del software se haya convertido en un factor determinante. Los avances actuales que se producen en el ámbito de la detección y corrección de vulnerabilidades, evolucionan, por lo general, hacia enfoques de seguridad reactivos [2][3].

Sin embargo, entre las causas de estos ataques tienen que ver, en su mayoría, con la falta de seguridad en el proceso de desarrollo del software. En este sentido, Apvrille y otros [4] ponen de referencia la necesidad de considerar la seguridad en todas las tareas de las fases del ciclo de vida del software. Así mismo, Solinas y otros [5] indican que la seguridad para el software deja de ser opcional y se hace obligatoria.

En contrapunto y pese a lo que indican estos estudios, las empresas de software utilizan la fase de pruebas para determinar la calidad del software e identificar y solventar vulnerabilidades. Incluso conociendo que esperar hasta la etapa de pruebas para detectar y solventar los fallos de seguridad aumenta exponencialmente el coste de su resolución [6]. Esto ocurre porque se realiza la identificación de vulnerabilidades desconociendo su origen, lo que ocasiona un rediseño y excesivos cambios en la implementación del software. Por este motivo, Sodiya y otros discuten en [7], los problemas de conseguir productos de software seguros.

La contribución que se presenta se centra en mantener una trazabilidad que permita mejorar la seguridad del software durante todas las fases de ciclo de vida del software. Los autores de esta investigación analizan en trabajos anteriores [8], los principales modelos de desarrollo de software seguro con el objetivo de detectar cuales son las actividades de seguridad esenciales que deben llevarse a cabo durante el proceso de construcción del software. Tras categorizar en [9] las actividades de seguridad detectadas se consigue la base que permite la particularización de un modelo de desarrollo seguro propio. En Fig.1 se detalla el modelo conformado e implantado en un contexto real y basado en una metodología sistemática y planificada. Compuesto por catorce actividades de seguridad que se encuentran organizadas en cuatro áreas de desarrollo.

Fig. 1. Modelo de Desarrollo de Software Seguro Viewnext-UEx

El objetivo del modelo es aplicar la seguridad en el software desde las fases más tempranas, evitando que la implementación y diseño de software se oriente únicamente a la funcionalidad y que, como ocurría hasta ahora, no se identifiquen y solventen vulnerabilidades hasta la fase de pruebas. Cuando el desarrollo del software está finalizado.

Al igual que Mellado y otros [10], se centran en el proceso de ingeniería de requisitos de seguridad para minimizar fallos o posibles ataques maliciosos, esta contribución detalla el proceso de trazabilidad seguido para identificar riesgos, obtener requisitos o historias de seguridad y definir buenas prácticas de codificación segura.

II. Estado del Arte

La complejidad en la obtención de requisitos o historias de seguridad ha puesto de manifiesto la creación de distintas técnicas. Enfocadas en su mayoría hacia la definición de historias de usuario para las metodologías ágiles, frente la obtención de requisitos de seguridad de los procesos tradicionales. En este sentido, podemos hablar de [11]

OWASP Cornucopia, un mecanismo en forma de un juego de cartas ideado para ayudar a los equipos de desarrollo de software a identificar los requisitos de seguridad en procesos de desarrollo ágiles, convencionales y formales. Parecido se define SAFECode [12], una guía de seguridad de software para profesionales del mundo ágil que permite al Product Owner, de darle voz a las necesidades de negocio, en este caso llenar en forma de épicas de seguridad el backlog.

Sin embargo, más relacionado con la propuesta realizada por estos autores es lo indicado en [13] por Siiskonen y otros, que presentan como obtener historias de usuarios de seguridad de forma genérica.

III. Trazabilidad

La incorporación sistemática de prácticas de seguridad correctamente planificadas en el ciclo de vida de desarrollo necesita un proceso de trazabilidad. En Fig.1 se observa que este proceso se centra en el área de desarrollo –Metodología SDL- orientada al proceso de construcción de software seguro por defecto. Así, Yu y otros, proponen en [14] una técnica de trazabilidad basada en la refactorización e integrada con otras actividades de mantenimiento del software seguro.

El proceso de trazabilidad diseñado debe identificar los requisitos o historias de seguridad mediante un cuestionario de preguntas relacionadas con los componentes, la arquitectura, el diseño y la funcionalidad de la aplicación a desarrollar. Se extraen los aspectos de seguridad intrínsecos en las respuestas funcionales proporcionadas. Obtener el listado de requisitos o historias de seguridad facilita el proceso de modelado de amenazas.

La trazabilidad busca mantener un hilo conector durante todo el desarrollo del software que compruebe el estado de los aspectos de seguridad que afectan al software. Así, se evita que el diseño e implementación del software se oriente únicamente a la funcionalidad de la aplicación, teniendo en cuenta la seguridad desde fases tempranas.

IV. Diseño de la herramienta

Los autores han diseñado una herramienta, actualmente en desarrollo, que debe simplificar el proceso de obtención de requisitos o historias de seguridad. En Fig.2 se observa la que esta herramienta utiliza como base un cuestionario de preguntas sobre la aplicación a desarrollar y aporta como resultado final, un listado de requisitos o historias de seguridad.

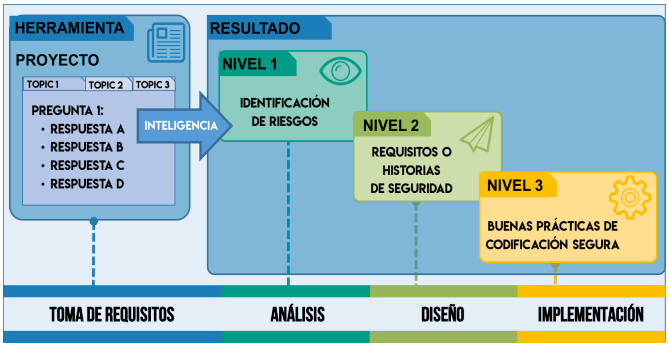


Fig. 2. Proceso descriptivo de la metodología que sigue la herramienta.

Para permitir un mayor entendimiento, se describen los conceptos utilizados en la definición de la metodología que desarrolla la herramienta.

A. *Topics de seguridad*

Se han estudiado en [15] los tipos de vulnerabilidades más comunes en la actualidad con la finalidad de crear una serie de *topics* o categorías mapeados que permitan cubrir ampliamente diversos aspectos de la seguridad a la hora de construir el software. Fig. 3 muestra que estos *topics* son los encargados de recoger de forma categórica las preguntas sobre la aplicación. Ejemplos de *topics* pueden considerarse:

- Arquitectura: Indica aspectos sobre el tipo de la aplicación, la estructura, los componentes y la conexión entre las distintas partes del software.
- Autenticación: Proceso de verificación para el acceso de los usuarios a la aplicación.
- Autorización: Métodos o procesos que verifican el permiso a los recursos de la aplicación.
- Validación de entradas: Método de validación sobre los datos que la aplicación permite introducir.
- Manipulación y registro de errores: Relativo al proceso de comprobación sobre la información que muestra la aplicación cuando se producen fallos en ella.
- Marco regulatorio: Indica los aspectos legales relativos al sector de la empresa donde se utiliza el software.
- Administración de usuarios: Relativo a los datos que se almacenan y al ciclo de vida de las cuentas de los usuarios.
- etc.

B. *Batería de preguntas y respuestas*

Las preguntas respondidas por el analista de negocio recogen aspectos funcionales de la aplicación. Ejemplos de preguntas son:

- ¿Qué tipo de aplicativo es?
- ¿El sitio requiere autenticación?
- ¿A través de qué método se autentica el usuario?
- ¿Cuándo cambia el identificador de sesión?

Tras responder las preguntas planteadas en los diversos *topis* de seguridad se obtienen de forma automatizada los siguientes elementos.

C. *Listado de riesgos*

Un riesgo es el proceso vulnerable que compromete el activo de un sistema. Un ejemplo de riesgo son los ataques de fuerza bruta automatizados para la autenticación de usuarios.

D. *Requisitos o historias de seguridad*

Derivado de la identificación de riesgos anterior podemos definir como requisito o historia de seguridad la implementación de medidas y consideraciones de protección de autenticación, ante ataques automatizados.

E. *Buenas prácticas de codificación segura*

Una vez obtenido el requisito se generan las buenas prácticas asociadas a este. En ejemplo del anterior es solicitar al usuario que en la autenticación introduzca sólo una parte de la contraseña *"dígitos primero, tercero y sexto"*, siendo esta regla aleatoria cada vez.

F. *Conexión con otros sistemas*

Como funcionalidad adicional este software se integra dentro del proceso de integración continua del desarrollo del software permitiendo la conexión con otros sistemas dedicados a la gestión de tareas de planificación, seguimiento de los proyectos como Jira o Redmine.

Fig. 3. Listado de topics de seguridad y ejemplo de pregunta realizada en la aplicación.

V. Aspectos Diferenciadores

Las técnicas o herramientas no comerciales [11] [12] que permiten la identificación de requisitos o historias de seguridad son procedimientos manuales. Esto impide el seguimiento automatizado de los requisitos de seguridad desde la integración, la implementación, las pruebas y hasta la liberación del código.

La herramienta propuesta busca que tras responder una serie de preguntas relacionadas con la aplicación a desarrollar se obtenga el listado de requisitos o historias de seguridad automatizado y transferirlas a herramientas de integración continua como Jira o Redmine. Así se observa en *Fig 4*. De esta forma se mantiene la trazabilidad de los requisitos sin perder en ningún momento el estado de los aspectos de seguridad de un proyecto software, pudiendo comprobar su implementación al llegar a fases finales del proyecto.

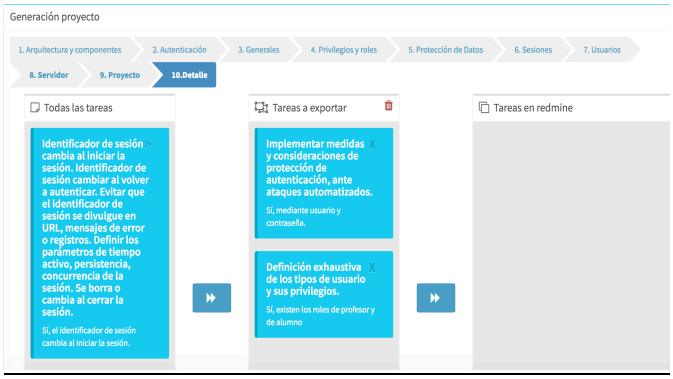


Fig. 4. Listado de requisitos de seguridad a seleccionar para su exportación a Redmine.

El principal punto fuerte de la herramienta es la flexibilidad en la configuración del cuestionario para cada proyecto. Permite añadir o quitar cualquier elemento que posteriormente derive en la generación de los requisitos de seguridad.

Otro aspecto ventajoso que ofrece la herramienta es la creación de plantillas que permitan su reutilización. Una vez terminada la configuración total de un cuestionario se posibilita su transformación a plantilla y se ofrece para su posterior uso, ahorrando tiempo a los analistas encargados de la toma de requisitos.

VI. Arquitectura

La arquitectura de la aplicación está compuesta por el Stack MEAN. Este framework está compuesto por el siguiente conjunto de subsistemas: MongoDB, ExpressJS, AngularJS y NodeJS. En Fig.4 se puede observar gráficamente la arquitectura e interacción entre los distintos componentes.
Se utiliza MongoDB como sistema gestor de bases de datos empleado para almacenar la información generada por la aplicación. Un frontal implementado con AngularJS utiliza los servicios del backend con NodeJS. Recalcar la existencia de otro servicio NodeJS, conectado con el anterior y encargado de generar y exportar los requisitos enviados a los diferentes gestores de tareas.

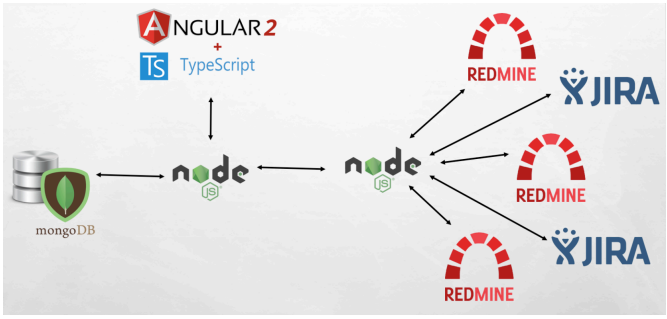


Fig. 5. Arquitectura y conexión entre los componentes de la aplicación.

VII. Desarrollo Modular

Este apartado define los módulos que componen la aplicación y sus respectivas funcionalidades:

- Dashboard: Muestra una pantalla de bienvenida con gráficos informativos del uso de la aplicación, estadísticas, contador de usuarios, número de proyectos, etc.
- User: Incluye funcionalidades relacionadas con la seguridad de acceso a la aplicación y a los recursos. Tales como el control de acceso, registro de usuarios, edición del perfil, gestión de roles, etc.
- Project: Engloba lo relacionado con la configuración para lanzar el proyecto y garantizar su funcionamiento. Así como de la gestión de las dependencias.
- Configurator: Encargado de la creación de las preguntas, las respuestas y la asignación de los requisitos asociados.
- Generator: Permite elegir o desechar los requisitos que se

exportarán a Jira y Redmine. Como se observa en *Fig.6*, este módulo solicita los datos de acceso: dirección IP, puerto de conexión, nombre del proyecto, usuario y contraseña del gestor de tareas que corresponda.

- Template: Ofrece la posibilidad de guardar un proyecto como plantilla para compartir entre los usuarios e importarla como plantilla base cuando se genere un nuevo proyecto.
- Mediatory: Se encarga de comunicar la aplicación web con las aplicaciones a las que se desea exportar los recursos.

Fig. 6. Ejemplo de conexión para la exportación de requisitos.

VIII. Trabajos futuros

Los trabajos futuros se centran en ampliar la batería de preguntas relacionadas con *topics* genéricos y poder disponer de mayor número de requisitos de seguridad contemplados. Los autores consideran que es fundamental aplicar diversas pruebas de conceptos en proyectos de desarrollo de software reales que permitan retroalimentar y mejorar la usabilidad y funcionalidad de la aplicación.

Ampliar el abanico de herramientas para la exportación de requisitos es un factor diferenciador, por lo que se convierte en otra de las tareas pendientes. Jenkins y Trello son dos herramientas a valorar para su inclusión.

IX. Conclusiones

El proceso de obtención de requisitos o historias de seguridad se considera complejo. En parte, debido a que comúnmente el proceso de toma de requisitos se centra puramente en cuestiones de funcionalidad, dejando a un lado la seguridad.

Los autores de esta contribución proponen y diseñan una herramienta que ayuda y simplifica el proceso de obtención de requisitos o historias de seguridad. Basada en la resolución de una batería de preguntas genera de manera automatizada un listado de requisitos o historias de seguridad que pueden ser exportadas a otras herramientas de seguimiento o integración continua del desarrollo de software.

Agradecimientos

Los autores agradecen la financiación recibida para esta investigación por parte de la Junta de Extremadura (Consejería de Economía, Comercio e Innovación), FEDER (Fondo Europeo de Desarrollo Regional) y ViewNext, empresa de Servicios de Tecnologías de la Información.

Referencias

[1] I. Cisco Systems, «Informe anual sobre ciberseguridad de 2017», 2017. [En línea]. Disponible en: http://www.cisco.com/c/dam/m/digital/1226019/Cisco_2017_ACR_es-xl.pdf. [Accedido: 20-jun-2017].

[2] H. Tran, E. Campos-Nanez, P. Fomin, y J. Wasek, «Cyber resilience recovery model to combat zero-day malware attacks», *Comput. Secur.*, vol. 61, pp. 19-31, 2016.

[3] S. S. Murtaza, W. Khreich, A. Hamou-Lhadj, y A. B. Bener, «Mining trends and patterns of software vulnerabilities», *J. Syst. Softw.*, vol. 117, pp. 218-228, 2016.

[4] A. Apvrille y M. Pourzandi, «Secure software development by example», *IEEE Secur. Priv.*, vol. 3, n.º 4, pp. 10-17, 2005.

[5] M. Solinas, L. Antonelli, y E. Fernandez, «Software secure building aspects in computer engineering», *IEEE Lat. Am. Trans.*, vol. 11, n.º 1, pp. 353-358, 2013.

[6] National Institute of Standards and Technology, «The Economic Impacts of Inadequate Infrastructure for Software Testing», *Quality*, 2002. [En línea]. Disponible en: https://www.nist.gov/sites/default/files/documents/director/planning/report02-3.pdf. [Accedido: 18-jun-2017].

[7] A. S. S. Sodiya, S. A. A. Onashoga, y O. B. B. Ajayi, «Towards building secure software systems», *Inf. Universe J. Issues Informing Sci. & Inf. Technol.*, vol. 3, pp. 635-646, 2006.

[8] J. C. Sancho Núñez, A. Caro Lindo, y P. García Rodríguez, «Análisis de metodologías de Desarrollo de Software Seguro», en *Jornadas Nacionales de investigación en Ciberseguridad (JNIC)*, 2016, pp. 42-47.

[9] J. C. Sancho Núñez, A. Caro Lindo, P. García Rodríguez, y Á. Quesada, «Categorización de Actividades de Seguridad en el Desarrollo de Software», en *Jornadas de Ingeniería del Software y Bases de Datos*, 2016, pp. 565-568.

[10] D. Mellado, E. Fernandez-Medina, y M. Piattini, «A Security Requirements Engineering Process in Practice», *Lat. Am. Trans. IEEE (Revista IEEE Am. Lat.*, 2007.

[11] OWASP Project, «OWASP Cornucopia». [En línea]. Disponible en: https://www.owasp.org/index.php/OWASP_Cornucopia. [Accedido: 01-mar-2018].

[12] V. Asthana, I. Tarandach, N. O'Donoghue, B. Sullivan, y M. Saario, «Practical Security Stories and Security Tasks for Agile Development Environments», *SAFECode*, p. 34, 2012.

[13] V. Ylimannela y M. Helenius, *Handbook of The Secure Agile Software Development Life Cycle*. 2014.

[14] Y. Yu, J. J??rjens, y J. Mylopoulos, «Traceability for the maintenance of secure software», en *IEEE International Conference on Software Maintenance, ICSM*, 2008.

[15] OWASP, «OWASP Top 10 - The Ten Most Critical Web Application Security Risks», *Owasp*, p. 22, 2017.

.

Ciberdefensa empresarial: Un marco conceptual y práctico en un entorno digitalmente inestable

Jeimy J. Cano M.
Universidad del Rosario
Bogotá, Colombia
jeimy.cano@urosario.edu.co

Abstract- **La empresas del siglo XXI avanzan cada vez más hacia un entorno digitalmente modificado. Las ventajas de contar con capacidades especializadas asistidas por terceros de confianza y despliegues ágiles, contrastan con las evidentes manifestaciones de poder y control de los Estados, reveladas en ciberconflictos globales, frente a los nuevos activos digitales que definen las promesas de valor de las empresas. En este contexto, las organizaciones deben desarrollar nuevas conversaciones desde la geopolítica e infopolítica para dar cuenta con los retos que implica mantener una posición estratégica en el escenario actual. En consecuencia este documento presenta un marco conceptual y práctico de ciberdefensa empresarial, como una manera de enriquecer las reflexiones sobre los ciber riesgos y comprender como la organización participa y puede ser afectada en un escenario volátil, incierto, complejo y ambiguo con el actual.**

***Index Terms*- ciberseguridad, ciberdefensa, ciberriesgo**

I. INTRODUCCIÓN

De acuerdo con recientes estudios la inestabilidad global se puede interpretar desde tres elementos fundamentales: el vacío de poder, la cuarta revolución industrial y la vulnerabilidad digital [1]. En este contexto, las personas, las organizaciones y las naciones asisten a un escenario donde se experimentan tensiones por el control de activos estratégicos en el ciberespacio, que repercuten en el ciberdominio, como aquella expresión social y de dinámica sociopolítica que surge por la interacción de las personas en el ciberespacio [2].

Esta realidad de superávit de información, de recursos tecnológicos disponibles en los nuevos ecosistemas digitales y la hiperconectividad de los individuos, establece un nuevo momento en la evolución de la sociedad, donde las fronteras de la soberanía se hacen cada vez más difusas y los alcances de las acciones realizadas en el ciberespacio se hacen más sistémicos y con consecuencias inesperadas.

Lo que en la guerra cinética o regular, se entendía como alcances y efectos del conflicto, los cuales podían ser definidos y controlados, en el contexto digital y de relaciones propias de un ciberdominio que reclama una ciudadanía digital, es una apuesta incierta, como quiera que la afectación de un objeto específico puede tener efectos no previstos sobre el objetivo particular de la acción. El concepto de efecto colateral, que bien se puede conocer en el mundo físico, en el escenario digital, resulta menos evidente. En este último, puede ocasionar desde una pérdida de información sensible hasta causar la muerte de una persona con un dispositivo médico implantado [3].

En consecuencia, las organizaciones actuales deben acelerar su "despertar" a nuevas realidades como los "ciberconflictos" (o ciberguerra) y "la guerra en la red – netwar", las cuales ya no son ciencia ficción, sino evidentes manifestaciones de las luchas de poder y control que se han planteado los estados, y donde las empresas terminan siendo, por lo general, botines de la guerra digital, generando efectos no deseados sobre la dinámica empresarial y las promesas de valor para sus clientes [4].

Comprender la dinámica de los ciberconflictos desde la vista empresarial, implica actualizar la lectura ejecutiva de los miembros de junta, incorporar nuevas variables de análisis en sus reflexiones y sobremanera, provocar una ruptura de las decisiones empresariales, donde los informes de gestión tradicionales, ahora se cambian por análisis de perspectivas de entorno y la dinámica digital internacional, para comprender el grado de interdependencia que tiene la organización y cómo moverse en este escenario incierto de una realidad digitalmente modificada.

Así las cosas, las empresas deben crear un nuevo marco de discusión directiva y de gobierno de la corporación, ahora desde el entendimiento de las tensiones internacionales, articulando su actuar desde las prácticas de seguridad y cumplimiento, siguiendo protocolos de acción desde la estrategia gubernamental de protección digital de la gobernabilidad de un país [5].

En consecuencia, este documento presenta una revisión de la evolución del concepto de la guerra en el entorno digital, para establecer un marco conceptual y práctico que permita a las organizaciones identificar y asegurar los activos digitales estratégicos que dispone en la actualidad, cómo alinearse con las prácticas y directrices estatales sobre seguridad cibernética y desarrollar nuevas capacidades de defensa complementarias a las inicialmente articuladas y desplegadas por los estados

II. EVOLUCIÓN DEL CONCEPTO DE LA GUERRA: DE LO CINÉTICO A LO DIGITAL

El concepto de la guerra ha venido evolucionando a lo largo del tiempo, con relativos pocos cambios en sus motivaciones, estrategias y resultados. Dichos cambios en el contexto digital, que se comentan a continuación, ahora se concentran en situaciones que son novedosas frente a la dinámica de los conflictos regulares:

- Las partes involucradas desconocen las capacidades de acción o afectación que tiene una frente a la otra.
- La población civil se encuentra en medio del territorio de combate: ciberespacio.
- Los actores no necesariamente son estados. Los actores no estatales se convierten en elementos desestabilizadores de la dinámica digital global.

Estas característica propias de los ciberconflictos en el

escenario digital, establecen nuevos retos, no sólo para los estados sino para las organizaciones, las cuales hacen parte de la dinámica propia que se genera en el ciberdominio, quienes en terminarán siendo afectadas de forma directa (si son parte de la infraestructura crítica) a través de ciberarmas con implicaciones en su operación y deterioro de servicios a nivel de la nación, y de forma indirecta, para aquellas que hacen uso de la infraestructura crítica, no sólo por la afectación de sus negocios, sino por posibles afectaciones en sus infraestructuras tecnológicas o sistemas de información, donde se pueden generar fugas y/o pérdidas de información [6].

Hablar de una guerra en el entorno digital, no sólo es una reflexión que activa las alarmas en los entorno gubernamentales, sino que requiere especial atención de las organizaciones como quiera que éstas debe jugar al menos tres (3) roles claves para sobrevivir a las tensiones y retos que situaciones o acciones hostiles de actores no identificados, puedan afectar su adecuado funcionamiento y proteger sus activos digitales estratégicos, muchas veces fundados en capacidades claves asistidas por terceros de confianza, como pueden ser los proveedores de servicios en la nube.

Un primer papel de las organizaciones es *ser promotor de la apropiación de la ciberseguridad y la ciberdefensa* como nuevas distinciones de negocio de nivel estratégico, las cuales ahora son prácticas y capacidades que protegen directamente la promesa de valor para con sus clientes y su posición estratégica en un sector particular. Esto es, ilustrar y detallar cómo las posibles acciones hostiles que se inician en el ciberdominio, pueden terminar en actividades adversas planeadas y deliberadas en el ciberespacio, ocasionando efectos contrarios en su imagen, procesos de negocio y planes estratégicos.

En segundo lugar, las organizaciones deben *ser habilitadores para compartir información de inteligencia y riesgos emergentes*, que puedan afectar a otros actores del sector o fuera de él. Este es un ejercicio de construcción de confianza entre pares, que demanda una declaración de protección conjunta, que procura alinear y focalizar las capacidades de actuación que le indique al posible agresor (estatal o no) que existe una vista de defensa conjunta que deberá surtir, tanto en modalidad pasiva como activa.

En tercer lugar, las empresas deben *ser parte activa de las estrategias de ciberseguridad nacional, alineando sus prácticas y capacidades vigentes con dichos lineamientos*, con el fin de enriquecer los análisis previstos por la autoridad de la nación y participar de los ejercicios y simulación de seguridad cibernética del país, no sólo para saber cómo deben actuar ante una amenaza de alcance nacional, sino para reconocer sus interdependencia con los diferentes actores del ahora ecosistema digital del cual hace parte.

Mientras en el mundo de la guerra cinética, las empresas jugaban el juego de la política y la inteligencia en medio de las tensiones del poder internacional, para lograr evadir bien un cerco económico o un ataque dirigido; en el contexto digital, las empresas deben estar analizando el territorio de las tensiones geopolíticas y las posibles repercusiones que pueden generarse frente a sus activos digitales que están a la fecha sobre el ecosistema digital donde ahora opera, crea y captura valor para sus clientes [7].

III. IDENTIFICANDO LOS ACTIVOS DIGITALES ESTRATÉGICOS DE LAS ORGANIZACIONES

En este nuevo entorno de negocios digitales, las empresas crean nuevos activos con características del mundo digital, los cuales se configuran de otras formas particulares y demandan un ejercicio de protección diferente al tradicional que se conoce para la información clave en las organizaciones, como quiera que en su definición y constitución, existen terceros que hacen parte de la promesa de valor que se quiere transferir a la experiencia del cliente [8].

Los activos digitales actuales corresponden con las capacidades que las organizaciones requieren para desarrollar sus negocios con mayor agilidad, alienadas con las expectativas cambiantes de sus clientes, los cuales demandan experiencias cada vez más novedosas. Para lograrlo, las empresas cuentan con terceros de confianza, que le proporcionan bases de datos agregadas, plataformas de despliegue rápido de aplicaciones, capacidades analíticas de datos y aplicaciones inteligentes que monitorizan, controlan, optimizan y funcionan de manera autónoma sobre los productos que desarrollan las empresas [9].

En este contexto, ya no solamente se tienen activos de información, los cuales generalmente son valorados según su nivel de riesgo e impacto frente a eventos no deseados que se puedan presentar, sino activos digitales, que tienen una construcción de carácter relacional y sistémica, como quiera que, son las conexiones que se habilitan desde los terceros sobre los productos y/o servicios, los que capitalizan las experiencias de los usuarios y la protección de la información que se deriva de dicha interacción, que es en últimas, la que se compila para crear un mejor conocimiento del individuo y así mejorar en cada interacción posterior [10].

Los activos digitales estratégicos se convierten en elementos claves asociados con la promesa de valor de la empresa. Esto es, las capacidades provistas por los terceros que configuran un producto y/o servicio que termina convirtiéndose en un activo intangible, que leído desde la vista de la contabilidad, tiene un valor, es posible tener posesión de él, se puede configurar un derecho de uso, es posible transferirlo y declarar una propiedad particular sobre el mismo[30].

Ejemplos de activos digitales se puede identificar en empresas globales como Amazon o Facebook. Para Amazon, su portal de acceso a la tienda, es un activo digital que esta articulado por capacidades creadas por un desarrollo interno de la unidad de negocio *Amazon Web Services* (AWS). Este define el medio por el cual, la empresa captura y recrea las experiencias de sus clientes y articula la interacción desde la selección del producto, la venta, el servicio post venta y sus preferencias. Un evento no deseado sobre este portal, puede deteriorar la imagen de la empresa, el cual sin una adecuada atención, es capaz de debilitar el posicionamiento de marca y la confianza que se ha logrado en años de operaciones.

De igual forma, Facebook es una plataforma de servicios de interacción entre personas. Para lograr esta configuración, la empresa incorpora capacidades de empresas como Microsoft Azure, en su plataforma de aplicaciones para desarrollar interacciones entre sus participantes, Amazon como base de la infraestructura tecnológica disponible para el manejo de sus contenidos y la empresa OpenID como su proveedor de servicios de autenticación para asegurar la

conectividad de los usuarios a su plataforma. Eventos adversos o de mal uso de la información sobre Facebook puede socavar la confianza de sus clientes, habida cuenta que la protección de la información personal se convierte en una práctica fundamental, dadas las características de la dinámica de funcionamiento de los participantes de esta red social digital [11].

Por tanto, estos nuevos activos digitales, que resultan de una interacción de diferentes componentes tecnológicos orientados por la promesa de valor de la compañía, requieren no solo proteger y asegurar la información como activo clave de la empresa, sino ahora prepararse para desarrollar capacidades de Ciberdefensa, que le permitan tanto a la empresa como a sus aliados estratégica, defender y anticipar los riesgos y amenazas emergentes con el fin de alcanzar mayor resiliencia digital ante situaciones o eventos inesperados en un contexto cada vez más asimétrico e inestable.

IV. Capacidades de defensa: el elemento ciber en la lectura empresarial

Las empresas digitalmente modificadas y aquellas que aspiran a transformarse digitalmente para desarrollar nuevas habilidades y capacidades para competir en un escenario incierto como el actual, están expuestas a una dinámica empresarial y global novedosa, donde no solamente los competidores están listos para crear disrupciones tecnológicas y de servicios inesperadas, sino que actores no documentados, inciertos o mercenarios digitales a nombre de naciones, estarán atentos para crear situaciones que puedan comprometer no sólo su operación o información estratégica, sino sus activos digitales para degradar su postura estratégica en el ciberespacio.

Comprender las capacidades de defensa en el entorno empresarial, implica entender a la organización como un actor más de la geopolítica e infopolítica global, que interactúa con diferentes actores de su ecosistema digital. En este sentido, no solamente es necesario distinguir lo que la organización debe proteger al interior de sus fronteras, sino detallar las interacciones e impactos de sus relaciones con los diferentes participantes de su ecosistema para comprender los alcances y acciones que debe desarrollar para ser parte, bien de la malla de protección conjunta, o de la falla generalizada que afecta a todos en el entorno [28].

En el ejercicio de modelar el ecosistema de la empresa en el contexto de sus relaciones digitales con el entorno, se deben tener en cuenta al menos un conjunto actores y configurar un perfil de ciber riesgos [27] básico, que le permita establecer y priorizar las acciones que debe adelantar para diseñar o fortalecer las capacidades de defensa, que no son sólo las prácticas de seguridad y control, que ya deben estar formalizadas y en operación en la empresa, sino un conjunto de nuevas actividades (generalmente asistidas con terceros de confianza especializados) que le permiten aprender, analizar, responder y anticipar frente al entorno incierto y ambiguo donde actualmente opera.

De acuerdo con Kaplan, Bailey, O'Halloran, Marcus, & Rezek [12] los siguientes son el conjunto mínimo de actores que se deben considerar para comprender cómo la organización participa de su ecosistema digital: autoridades de regulación, clientes, proveedores (de negocio, tecnología y servicios), aseguradoras, gremios económicos, sociedad civil, entidades del gobierno, defensa y los atacantes. En cada uno de ellos, es posible encontrar relaciones relevantes que deben ser incluidas en los análisis que se realicen frente a la comprensión de la empresa y sus retos en el entorno digital.

Si bien a la fecha no existe una definición para el tema de ciber riesgo [3], si es posible establecer un conjunto de características claves que se deben atender para dar cuenta de los retos que este tipo de riesgo implica. En el ejercicio de conceptualizar un ciber riesgo, siguiendo las reflexiones de Maitra [3], se pueden establecer al menos cinco elementos esenciales para comprender su dinámica:

- la actividad no autorizada o agresión - acciones realizadas de manera intencional o no en el contexto de la organización.
- el agresor – persona u organización que planea y efectúa la actividad no autorizada o agresión. Ejemplo: Actores estatales y no estatales, crimen organizado, empleados internos, mercenarios digitales
- la vulnerabilidad - determinadas por las prácticas y estándares que la organización tiene sobre la gestión de la tecnología, sus procesos y las personas. Ejemplo: Compatir las contraseñas, contraseñas por defecto.
- el ataque – el aprovechamiento de las vulnerabilidades conocidas o desconocidas para concretar acciones que interrumpan, deterioren, alteren, revelen o destruyan activos y/o servicios claves de la empresa. Ejemplo: Malware, DDos
- la consecuencia - los efectos que se generan basados en las intencionalidades de los agresores. Ejemplo: Revelar información, espionaje, extorsión, robo de información, sabotaje, fraude.

De forma complementaria, Hermans & Diemont [14] establecen los elementos básicos que debe contener un perfil de ciber riesgo que dan forma a los elementos previamente planteados. Entorno de negocio, regulación, vulnerabilidades/resiliencia, atacantes y objetivos claves o estratégicos, fundan el marco general en el cual la organización debe profundizar para definir el nivel de exposición al riesgo en el escenario digital y hacer una lectura empresarial de sus resultados con miras, bien a fortalecer las práctica de seguridad y control vigentes, o a desarrollar las capacidades que son necesarias para defender y anticipar las amenazas digitales a las cuales puede estar expuesta la organización.

Así las cosas, las capacidades de defensa en una organización digitalmente modificada son cinco básicas, las cuales son semejantes en sus nombres a las que pudiese haber en el contexto militar, pero limitadas en su alcance y forma, dada su implementación en el ámbito empresarial. Las capacidades definidas son: defensivas, ofensivas, disuasivas, diplomáticas y de inteligencia, las cuales se detallan en la siguiente sección.

V. Ciberdefensa empresarial. De las reflexiones militares a las capacidades empresariales

Hablar de ciberdefensa en el contexto empresarial puede sonar extraño y contrario al espíritu de lo empresarial. Sin embargo, considerando las amenazas emergentes y los ciberconflictos vigentes en el escenario internacional, las organizaciones deben estar preparadas para comprender y movilizarse en un entorno digital lleno de incertidumbres y

enemigos invisibles, que no van solamente a aprovecharse de vulnerabilidades para mancillar su imagen, sino a desarrollar todo un ejercicio de operaciones en el ciberespacio para concretar fines superiores, usando a las organizaciones bien como puentes o botines de esta nueva guerra asimétrica, en la que la población y las empresas hacen parte del campo de batalla.

Basado en lo anterior, cinco (5) son las capacidades que las corporaciones deben desarrollar como parte natural de su programa de defensa digital (ver figura 1), el cual debe estar alineado con las prácticas de seguridad de la información y la dinámica que implica la ciberseguridad empresarial para alcanzar cada vez más niveles de resiliencia digital [15]. Para efectos de este marco conceptual se entenderá una capacidad, adaptando las reflexiones de Su & Linderman [16], como un patrón de aprendizaje y desaprendizaje de construcción colectiva a través del cual una organización y sus colaboradores generan, modifican y actualizan sistémica y sistemáticamente sus reflexiones ejecutivas y rutinas operativas para alcanzar una mayor efectividad en sus procesos y un mejor posicionamiento estratégico.

Figura 1. Marco general de ciberdefensa. Elaboración propia

La primera es la capacidad *defensiva*. En el entorno militar implica proteger una posición estratégica en el territorio de las operaciones, lo que supone establecer el nivel de barreras, monitoreo y control de un área específica con el fin de tener una forma de actuar y atender cualquier eventualidad que se pudiese presentar [17]. Esta capacidad leída en clave empresarial implica en definir y desarrollar un perímetro de seguridad proactivo, es decir, con capacidades analíticas y cognitivas que permitan anticipar nuevos vectores de ataque, así como contener y demorar a los posibles agresores mientras se toman las acciones de tratamiento del evento en curso.

La siguiente tiene que ver con los temas *ofensivos*. Para la milicia la capacidad ofensiva establece los desarrollos armamentísticos disponibles para comprometer, vulnerar y conquistar a la contraparte, con el fin de ganar territorio del enemigo y así dominarlo y someterlo a las condiciones de la tropa que muestra la superioridad táctica y operativa. Esta misma capacidad en el escenario corporativo tiene algunas restricciones dada las posturas encontradas a la fecha sobre el derecho a responder que tienen las empresas frente a ciberataques [18].

En el escenario ofensivo, las organizaciones pueden desplegar actividades propias de un monitoreo pasivo y activo representado en redes señuelo (o honeypots), configuración de "agujeros negros" para consumir tráfico mal formado o malicioso hacia sus infraestructuras, que permitan estudiar el comportamiento del agresor y pensar en actividades de contra ataque como desactivación de las redes de los atacantes o desactivar sus servidores o infraestructura, acciones que deberán ser analizadas y validadas desde la vista ejecutiva empresarial y las directrices nacionales de la seguridad digital.

Hablar de *disuasión* en el campo militar implica desanimar a la contraparte para efectuar acciones hostiles en contra de su territorio, ubicando en el imaginario del agresor que se cuentan con armas o estrategias que pueden afectar gravemente la integridad o posición del atacante. En la empresa, la disuasión implica motivar y proyectar un compromiso institucional con el reto de los ciber riesgos, para lo cual se adelantan con frecuencia ejercicios basados en juegos de guerra, cooperación gubernamental e internacional y se habilitan espacios para compartir información estratégica que permite una defensa empresarial colectiva asistida por capacidades conjuntas entre los sectores productivos y el estado [19].

Las capacidades *diplomáticas*, si bien no son parte inherente del escenario militar, si es claro que requiere una visión geopolítica e infopolítica de los ejecutivos de la empresa, para motivar conversaciones que trasciendan la esfera empresarial y de sector, habida cuenta que los miembros de las juntas directivas, a nivel internacional, no solo actúan como representantes de la empresa, sino como embajadores del país donde tiene sede principal la organización. Este ejercicio debe estar coordinado con las autoridades nacionales que manejan las relaciones internacionales y los organismos multilaterales que procuran una estabilidad global digital [20].

Finalmente y no menos importante, las capacidades de *inteligencia*. En el ambiente militar la inteligencia compila información desde diferentes fuentes, establece nuevos escenarios de actuación y procura movilizar y afectar su entorno, para materializar aquellos que le son más favorables, manteniendo activas técnicas de contraseñuelos o engaños, para asegurar que sus acciones son aquellas se confirman en la realidad [21].

En el entorno empresarial, la inteligencia demanda un proceso evolutivo desde lo operacional a lo cognitivo. Esto es, pasar de un centro de seguridad de operaciones (en inglés SOC) que observa en tiempo real eventos, identifica desviaciones sobre patrones conocidos para generar alertas basadas en reglas estáticas, a una identificación de patrones emergentes, analítica de eventos inusuales en tiempo real basado en heurísticas, que haga tránsito para alcanzar la distinción de un centro cognitivo de seguridad, donde la organización, asistida por algoritmos de inteligencia artificial, ejecuta simulaciones de ataques inusuales, aprende y pronostica sobre las variaciones y asimetría de su entorno de amenazas digitales [22].

En pocas palabras, podemos resumir que una ciberdefensa empresarial efectiva, parafraseando a Donaldson et al [23], es aquella que interrumpe los ataques para disuadir a los agresores con menos determinación; detecta aquellos que están en curso, retrasa a los atacantes antes que tenga éxito, afecta a su contraparte y en el mejor de los casos, supera y conquista a su adversario, esto último, si y sólo si, está autorizada la organización a contraatacar, para lo cual debe existir una validación desde las instancias ejecutivas a nivel gubernamental.

Lo anterior sugiere que la ciberdefensa empresarial es un ventana de aprendizaje/desaprendizaje permanente para una organización que busca mantenerse operativa y proyectada en

el futuro, a pesar de las inestabilidades y amenazas digitales emergentes del entorno. Para ello, se requiere no solo un desarrollo clave de capacidades analíticas y cognitivas que la preparen frente a escenarios inciertos, sino un compromiso del nivel ejecutivo, que asumiendo su "deber de liderazgo", sea capaz de adelantar las conversaciones infopolíticas entre sus miembros del directorio para dar cuenta de la nueva realidad organizacional en un escenario digitalmente modificado.

VI. Un ejemplo de aplicación del modelo general de ciberdefensa

Recientemente se ha evidenciado un ciberataque al sistema de pagos de electrónicos interbancarios (SPEI) del Banco Central de México (Banxico), donde al menos cinco instituciones financieras estuvieron involucradas, en las cuales hubo una extracción de alrededor de 300 millones de pesos mexicanos (aprox. US$ 15 Millones). El ataque, según el informe del Banxico buscaba "vulnerar las conexiones de las instituciones con el SPEI, inyectando instrucciones de pago fraudulentas a partir de cuentas inexistentes" [29].

Siguiendo el modelo de ciberdefensa, el Banxico aplica su capacidad defensiva, con dos niveles de control y validación. Un sistema donde se encuentran los participantes (bancos, casas de bolsa, etc), el cual termina afectados por las acciones de los adversarios y otro aparte, que no es comprometido, donde están las cuentas de los clientes finales con verificaciones individuales por operación.

Sus capacidades ofensivas, estaban orientadas a una estrategia de defensa pasiva, considerando su monitoreo perimetral y mecanismos de registros de auditoría en los sistemas previamente comentados. Basado en su revisión y análisis se establece el modo de operación y acción de los atacantes, estableciendo las lecciones aprendidas del evento y ajuste de sus normativas en temas de seguridad y control. Acciones de defensa activa, si estuviesen previstas, deberían estar ajustadas con el marco de ciberseguridad nacional de México [31].

El Banxico, como supervisor del sistema financiero mexicano, comparte información y establece un marco de cumplimiento normativo, que actúa como capacidad disuasiva, para los diferentes actores del entorno, dadas las exigencias de procedimientos y controles que se encuentran previstos, particularmente en la circular 14/2017, denominada Reglas de Sistema de Pagos Electrónicos Interbancarios. Sin embargo, aún no se concretan acciones o ejercicios colectivos para este tipo de amenazas.

El Banco Central de México, como autoridad monetaria, establece los canales diplomáticos y comunicaciones pertinentes con las autoridades nacionales e internacionales, para efectos de tratar casos donde se pueda ver comprometida la información y afectados los intereses tanto de los participantes del sistema financiero como sus clientes. En este caso, no se reporta en los informes disponibles actividades concretas en este sentido.

Es probable, de acuerdo con la información disponible de en el sitio web del Banxico, que las capacidades de inteligencia, basadas en el uso de centros de operaciones de seguridad, no se hubiesen visto involucradas dado que el SPEI es un sistema aparte y cerrado para uso exclusivo de los participantes del sistema financiero. En este contexto, habría elementos de mejoramiento que deben aplicarse como una forma de fortalecer la estrategia defensiva del Banco Central frente a ciberataques.

En consecuencia, es claro que la estrategia de defensa del Banxico debe ser actualizada y ajustada frente a los nuevos retos y exigencia de un entorno más abierto y con actores desconocidos que desean generar inestabilidad en el contexto financiero del país y crear una escenario de incertidumbre y desconfianza nocivo para la salud económica del país y la reputación del Banco Central.

VII. Conclusiones

En un contexto de ciberconflictos, donde las organizaciones y los ciudadanos están en medio de las operaciones digitales no autorizadas, se hace necesario actualizar la lectura de la protección de la información y la resiliencia digital en las empresas, para comprender y ampliar el entendimiento de la dinámica digital de su entorno, para concretar nuevas capacidades de acción y respuesta que le permitan proteger sus activos digitales de actividades encubiertas o subrepticias que puedan comprometer las operaciones de sus procesos, impactar sus grupos de interés o destruir su posición estratégica en un mercado.

Si bien, los diferentes países vienes adoptando marcos de defensa cibernética y ciberseguridad nacional, donde los diferentes participantes revelan aquellos puntos neurálgicos que pueden tener impacto en el orden nacional, las empresas deben desarrollar capacidades complementarias que permitan coordinar y desplegar las acciones que los comandos de seguridad cibernética de las naciones tienen previstas. En este sentido, no es un contraataque lo que las organizaciones deben adelantar, sino motivar estrategias de defensa pasiva y activa que permitan al menos interrumpir, detectar y retrasar las acciones de los terceros contra sus diferentes procesos, infraestructura tecnológica o posicionamiento empresarial [2, 26].

En consecuencia, mientras los comandos cibernéticos, están habilitados para desarrollar capacidades ofensivas para vulnerar y superar a su contraparte, las empresas deben complementar sus estrategias de seguridad digital y resiliencia digital, teniendo en cuenta la configuración de sistemas de monitoreo y seguimiento en línea, inteligencia de amenazas, y sistemas de señuelos en red, que habiliten a la compañía para confinar el mayor tiempo posible la amenaza emergente sobre dicha organización, creando confusión en el atacante y tiempo para acciones conjuntas con aliados estratégicos.

Por tanto, deconstruir el concepto de ciber riesgos en la ciberdefensa empresarial, significa desarticular el paradigma de defensa conocido desde la esfera eminentemente militar y bélica basada en una guerra cinética, para conectarla con una realidad geográficamente distribuida, políticamente afectada y socialmente enriquecida, donde lo digital es parte inherente de la forma como las diferentes partes interesadas interactúan.

Es decir, desarrollar un movimiento teórico significativo alrededor de la razonabilidad de las medidas de defensa aplicadas frente al riesgo evidenciado, que procure entender que "*toda evaluación de riesgo es y se mantiene como algo sujeto al contexto. No existe ni psicológicamente ni bajo las condiciones sociales dominantes una preferencia positiva o negativa del riesgo*" [24, p.77].

Lo anterior, demanda de la ciberdefensa empresarial y de los cuerpos de gobierno corporativo una revisión conceptual de su postura frente a la realidad de los ciberconflictos, como fundamento de la construcción de una valoración del ciber riesgo enriquecida, que puede anticipar afectaciones fruto de sus análisis y conectar con el entorno geopolítico de la empresa y el país, "como prisma explicativo de los acontecimientos mundiales actuales" [25, p.61].

En resumen, el modelo conceptual propuesto, leído desde el marco general de ciberdefensa, establece una manera diferente de comprender la dinámica de la organización en su nuevo ecosistema digital. Por una parte, reconoce y asegura las prácticas de seguridad y control que debe tener desarrolladas e implementadas como base de su operación interna, y por otro, habilita una reflexión de alto nivel que permite entender a la empresa en sus relaciones externas con los diferentes actores de su entorno geopolítico e infopolítico, para comprender su posición estratégica derivada de sus acciones en este entorno.

AGRADECIMIENTOS

El autor agradece las reflexiones, experiencia y comentarios sobre los temas de ciberdefensa de la Mayor Milena Realpe del Ejército Nacional de Colombia, el Coronel Fredy Bautista García de la Policía Nacional de Colombia y el Contraalmirante John Fabio Giraldo Gallo de la Armanda Nacional de Colombia, los cuales fueron insumos fundamentales para el desarrollo de este artículo.

REFERENCIAS

[1] CIDOB (2017) El mundo en 2018: diez temas que marcarán la agenda internacional. *Reporte*. Recuperado de: http://bit.ly/2CEow1p

[2] Kello, L. (2017) *The virtual weapon and international order*. New Heaven, CT: Yale University Press.

[3] Romanosky, S. & Goldman, Z. (2017) Understanding cyber collateral damage. *Journal of National Security Law and Policy*. 9(2). 233-257.

[4] Brose, R. (2015) Cyberwar, Netwar, and the Future of Cyberdefense. En Maybaum, M., Osula, M. & Lindström, L. (Eds) (2015) *7th International Conference on Cyber Conflict: Architectures in Cyberspace*. Tallinn, Estonia: NATO CCD COE Publications. 25-38.

[5] Almeida, V., Doneda, D., & De Souza Abreu, J. (2017) Cyberwarfare and digital governance. *IEEE Internet Computing*. March/April. 68-71

[6] Rid, T. & McBurney, P. (2012) Cyber-weapons. *RUSI Journal*. February/March. 157(1). 6-13

[7] Tirenin, W. & Faatz, D. (1999) A concept for strategic cyber defense. *Conference Proceedings MILCOM 1999*. IEEE Military Communications. Atlantic City, NJ. USA. Vol. 1. pp. doi: 10.1109/MILCOM.1999.822725.

[8] Peppard, J. & Ward, J. (2016) *The Strategic Management of Information Systems: Building a Digital Strategy*. 4th Edition. Chichester, West Sussex, UK: John Wiley & Sons.

[9] Porter, M. & Heppelmann, J. (2014) How Smart connected products are transforming competition. *Harvard Business Review*. November. 1-23.

[10] Porter, M. & Heppelmann, J. (2015) How Smart, connected products are transforming companies. *Harvard Business Review*. Octubre.

[11] Iyer, B. & Henderson, J. (2010) Preparing for the Future: Understanding the Seven Capabilities of Cloud Computing. *MIS Quarterly Executive*. 9(2). June. 117-131

[12] Kaplan, J., Bailey, T., O'Halloran, D., Marcus, A. y Rezek, C. (2015) *Beyond cybersecurity. Protecting your digital business*. Hoboken, New Jersey. USA: Wiley.

[13] Maitra, A. (2015) Offensive cyber-weapons: technical, legal, and strategic aspects. *Environmental Systems & Decisions*. 35. 169–182. DOI: 10.1007/s10669-014-9520-7

[14] Hermans, J. & Diemont, T. (2017) Treating cyber risk. En Antonucci, D. (2017) *The cyber risk Handbook. Creating and measuring effective cybersecurity capabilities*. Hoboken, New Jersey. USA: John Wiley & Sons. P.111

[15] Choi, J., Lung, H., & Kaplan, J. (2017) A framework for improving cybersecurity discussions within organizations. *Mckinsey Quarterly*. November. Recuperado de: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/a-framework-for-improving-cybersecurity-discussions-within-organizations

[16] Su, H. y Linderman, K. (2016) An Empirical Investigation in Sustaining High-Quality Performance. *Decision Sciences*. 47(5). 787-819.

[17] Denning, D. (2014) Framework and principles for active cyber defense. *Computers & Security*. 40. 108-113. Doi: https://doi.org/10.1016/j.cose.2013.11.004

[18] Hoffman, W. & Levite, A. (2017) Sector Cyberdefense. Can active measures help stabilize cyberspace? *Carnegie Endowment for International Peace*. Recuperado de: http://ceip.org/2swaiib

[19] Cano, J. (2017) Juegos de guerra. Un ejercicio de construcción conjunta en ciberseguridad y seguridad de la información. *Revista Sistemas*. No. 143. Abril- Junio. 60-66.

[20] Clark, D., Berson, T., & Lin, H. (Editors) (2014) *At the Nexus of Cybersecurity and Public Policy: Some Basic Concepts and Issues*. National Research Council. Washington, D.C. USA: National Academies Press. Recuperado de: https://docs.house.gov/meetings/IF/IF02/20150303/103079/HHRG-114-IF02-20150303-SD006.pdf

[21] Heckman, K. & Stech, F. (2015) Cyber Counterdeception: How to Detect Denial & Deception (D&D). En Jajodia, S., Shakirian, P., Subrahmanian, V., Swarup, V., & Wang, C. (Editors) (2015) *CyberWarfare. Building the Scientific Foundation*. Advances in Information Security. Vol. 56. Switzerland: Springer Verlag. Chapter 6.

[22] Chaudhary, R. & Mcknight, D. (2017) The security intelligence center. Next steps: Beyond response to Anticipation. *Internal Audit Foundation – Crowe Horwath*.

[23] Donaldson, S., Siegel, S., Williams, C. y Aslam, A. (2015) *Enterprise cybersecurity. How to build a successful cyberdefense program against advanced threats*. New York, USA: Apress

[24] Luhmann, N. (2006) *Sociología del riesgo*. México, México: Universidad Iberoamericana - Instituto Tecnológico y de Estudios Superiores de Occidente, A.C (ITESO).

[25] Cuéllar, R. (2012) Geopolítica. Origen del concepto y su evolución. *Revista de Relaciones Internacionales de la UNAM*. No.113, mayo-agosto. 59-80.

[26] Green, J. (editor) (2015) *Cyber warfare. A multidisciplinary analysis*. New York, NJ. USA: Routledge.

[27] Eling, M. & Schnell, W. (2016) What do we know about cyber risk and cyber risk insurance? *Journal of Risk Finance*. 17(5). pp. 474-491. Doi: https://doi.org/10.1108/JRF-09-2016-0122.

[28] Choucri, N. (2012) *Cyberpolitics in international relations*. Boston, MA. USA: MIT Press.

[29] BCM – Banco Central de México (2018) Puntos Importantes sobre la Situación Actual del SPEI. *Reporte Oficial*. Recuperado de: http://www.banxico.org.mx/inicio/banner/informacion-importante-sobre-la-situacion-del-spei/%7B022CD9D7-11A9-68E6-D1A5-965F57A23F60%7D.pdf

[30] Laney, D. (2018) *Infonomics. Infonomics: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage*. New York, NJ. USA: Bibliomotion, Inc.

[31] Gobierno de México (2017) Estrategia Nacional de Ciberseguridad. Recuperado de: https://www.gob.mx/gobmx/documentos/estrategia-nacional-de-ciberseguridad

Anonimización de datos no estructurados a través del reconocimiento de entidades nominadas

Fadi Hassan, Josep Domingo-Ferrer, Jordi Soria-Comas
Universitat Rovira i Virgili
Department d'Enginyeria Informàtica i Matemàtiques
Cátedra UNESCO de privacidad de datos
CYBERCAT-Centre de Recerca en Ciberseguretat de Catalunya
Av. Països Catalans 26, 43007 Tarragona
E-mail {fadi.hassan, josep.domingo, jordi.soria}@urv.cat

***Resumen*—La anonimización de datos estructurados ha sido ampliamente estudiada en el pasado reciente. Sin embargo, anonimizar datos no estructurados (típicamente documentos de texto) sigue siendo una tarea manual que necesita más atención por parte de los investigadores. La principal dificultad a la hora de tratar datos no estructurados es que no existe un esquema de base de datos que pueda usarse para medir riesgos en la privacidad. De hecho, datos confidenciales y valores cuasi-identificadores pueden estar repartidos a lo largo de los documentos a anonimizar. En este trabajo proponemos usar un marcador de reconocimiento de entidades nominadas basado en aprendizaje automático. La finalidad es crear un sistema capaz de detectar todos los atributos que tienen implicaciones en la privacidad (identificadores, cuasi-identificadores y atributos confidenciales). Concretamente, presentamos una prueba de concepto centrada en la detección de atributos confidenciales. Consideramos un caso de estudio en el cual los valores confidenciales a detectar sean nombres de enfermedades en diagnósticos médicos. Una vez detectados estos valores de atributos confidenciales, se pueden usar técnicas estándar de control de revelación estadística para datos estructurados con el objetivo de controlar el riesgo de revelación.**

***Index Terms*—Anonimización, datos no estructurados, reconocimiento de entidades nominadas, campos condicionados aleatorios.**

I. Introducción

Hoy en día, se están recogiendo grandes cantidades de datos de fuentes muy diversas, a menudo sin que los individuos afectados tengan constancia de ello. Una recogida de datos tan sistemática, acoplada a nuevas técnicas de análisis de datos, ha dado paso a lo que se conoce como *big data* o *megadatos*. Aunque a veces se califiquen como moda, los megadatos implican un cambio significativo en la forma de manipular los datos. En este trabajo, nos preocupamos de las implicaciones de privacidad de los megadatos, concretamente de los megadatos no estructurados.

En el entorno tradicional, los datos eran recogidos principalmente a través de encuestas u otras fuentes de datos administrativos. Como resultado, solían tener una naturaleza estructuradas (una tabla). La amplia variedad de fuentes de datos en el contexto actual de megadatos (por ejemplo, correos electrónicos enviados y recibidos, participación en redes sociales, etc.) nos fuerza a considerar otros tipos de datos, como datos semiestructurados o no estructurados (texto libre). Ya en 2005, se comentaba en [9] que un 80 % de los datos comerciales y médicos se guardaban de forma no estructurada. En el contexto sanitario, el uso adecuado de estos datos es crítico para la investigación y la creación de políticas, además de ser útil para industrias relacionadas, como seguros de salud.

La nueva Regulación General de Protección de Datos europea (en inglés, *General Data Protection Regulation* o GDPR, [8]) afirma que es necesario el consentimiento explícito de los individuos afectados para usar información identificable personalmente (en inglés, *personally identifiable information* o PII) para fines secundarios (diferentes a los fines primarios que motivaron su recogida, como la sanidad o la facturación de un servicio). Idealmente, el recolector de datos debería esforzarse en adquirir dicho consentimiento. Sin embargo, esto puede no ser posible en la práctica. Puede ser difícil contactar con los individuos para obtener su consentimiento. Además, los individuos con condiciones anómalas suelen preocuparse más por su privacidad, lo que los hace menos propensos a dar consentimiento para usar sus datos. Debido a estos inconvenientes, es probable que la PII obtenida mediante consentimiento sea algo sesgada.

Para evitar la necesidad del consentimiento, los datos usados para fines secundarios no deberían identificar al individuo. La anonimización, también conocida como control de revelación estadística (en inglés, *statistical disclosure control* o SDC), proporciona una manera de transformar la PII en información que no puede asociarse a un individuo concreto identificado y, por lo tanto, no está sujeta a regulaciones de privacidad.

Existe una gran cantidad de literatura sobre SDC para el caso de datos estructurados [10], [5], [4]. Los datos estructurados son aquellos que pueden describirse como un conjunto de registros que corresponden a un individuo y contienen los valores de un conjunto fijo de atributos para ese individuo. La anonimización de datos estructurados consiste normalmente en eliminar los atributos que son identificadores y enmascarar los atributos cuasi-identificadores. Estos últimos son atributos que no son identificadores por sí solos, pero que conjuntamente pueden permitir enlazar el registro con alguna fuente de datos externa que contenga identificadores y, por lo tanto, podrían reidentificar el individuo al que pertenece el registro. Alternativamente, en vez de (o además de) enmascarar los cuasi-identificadores, se pueden enmascarar los atributos confidenciales para introducir incertidumbre sobre sus valores.

Una vez decididos qué atributos son cuasi-identificadores y cuáles son confidenciales, la anonimización de datos estruc-

turados puede automatizarse (cierto es que, en algunos casos, la decisión puede no ser del todo clara, ya que depende de la información de contexto que se supone disponible para el intruso). Sin embargo, la automatización de la anonimización de datos no estructurados es mucho más compleja, ya que no hay un esquema de base de datos con el que clasificar los datos en atributos identificadores, cuasi-identificadores y confidenciales. Como resultado, anonimizar datos no estructurados sigue siendo a día de hoy una tarea en gran parte manual.

De hecho, se puede argumentar que los datos no estructurados en forma de texto son los más complejos a que debe enfrentarse la anonimización. Otros tipos de datos que pueden parecer más complejos a primera vista o bien pueden reducirse a texto no estructurado usando herramientas automatizadas de extracción semántica (como ocurre con vídeos o audios) o bien es mejor no anonimizarlos debido a la comprensión todavía incompleta de su semántica (como es el caso de los datos genéticos).

Contribución y estructura de este artículo

El propósito de este trabajo es automatizar la extracción de atributos cuasi-identificadores y/o confidenciales de datos no estructurados en forma de texto. Es decir, queremos ser capaces de identificar, de forma automática, atributos como número de pasaporte, nombre, localización, edad, fecha de nacimiento, etc. Para ser más concisos, en este trabajo nos enfocaremos en documentos de diagnósticos médicos. Una vez se complete este proceso de identificación automática de atributos, podremos aplicar algunos métodos diseñados para anonimizar datos estructurados. Para identificar atributos, usaremos un marcador de reconocimiento de entidades nominadas (en inglés, *named-entity recognition* o NER) [7].

En la sección II, presentamos brevemente algunos conceptos importantes para entender este trabajo. En la sección III, se recuerdan trabajos previos relacionados con la anonimización de documentos. En la sección IV, describimos nuestra propuesta. Los experimentos se presentan en la sección V y las conclusiones e ideas para trabajos futuros se recogen en la sección VI.

II. Antecedentes

II-A. Reconocimiento de entidades nominadas

El reconocimiento de entidades nominadas es la tarea de localizar y categorizar términos importantes de un texto [18]. Este reconocimiento es una fuente de información para diferentes aplicaciones de procesamiento de lenguaje natural. El NER ha sido utilizado para mejorar el rendimiento de varias aplicaciones, como la respuesta de preguntas [12], la traducción automática de texto [1], la recuperación de información [24], y el análisis de los sentimientos de los tweets [11].

El NER también es útil para la anonimización de datos no estructurados (p.e. documentos de texto libre). En particular, puede detectar los términos que pueden usarse para reidentificar a un individuo y aquellos términos que contienen información sensible. Una vez localizados estos términos, constituyen información estructurada que puede ser anonimizada usando métodos de SDC (tales como generalización, supresión, etc.) para mantener bajo control el riesgo de revelación.

Hay muchos esquemas de marcado para NER. En este trabajo, usamos el esquema de marcado IOB2 [22]. En IOB2, cada palabra en el texto se marca usando una de las tres posibles etiquetas: I, O o B, que indican si la palabra está dentro (inside), fuera (outside) o al principio (beginning) de una entidad nominada.

II-B. Campos condicionados aleatorios

En procesamiento de lenguaje natural, hay dos modelos comunes para resolver tareas de NER: modelos de Markov ocultos (en inglés, *hidden Markov models* o HMMs), usados en trabajos como [17], [28], y campos condicionados aleatorios (en inglés, *Conditional Random Fields* o CRFs), usados en trabajos como [3], [6], [11]. Los NER basados en CRFs son ampliamente usados y aplicados, y normalmente dan mejores resultados en varios campos [16]; por ello, en este trabajo diseñamos nuestro modelo usando CRFs.

Los CRFs [15] son modelos de grafos no dirigidos entrenados de forma condicionada que a menudo se aplican en el reconocimiento de patrones. Estos modelos se usan para calcular la probabilidad condicionada de valores en determinados nodos de salida dados los valores asignados a otros nodos de entrada.

III. Trabajo relacionado

Se han propuesto varias técnicas de anonimización para datos no estructurados en forma de texto. Muchas de ellas pueden clasificarse en una de las siguientes categorías: técnicas basadas en diccionarios y técnicas de aprendizaje automático [20].

En el pasado, la anonimización de documentos se hacía de forma manual, buscando y reemplazando las entidades nominadas. Sweeney [25] propuso el método de fregado que se basa en la definición de plantillas para las entidades nominadas, como la localización, el nombre y el país. Una vez encontradas estas entidades, se enmascara el valor relacionado.

Neamatullah et al. [19] propusieron un software para anonimizar documentos que usaba tablas de búsqueda lexicográfica, expresiones regulares y heurísticas simples, realizando comprobaciones en el contexto para localizar entidades nominadas. Tras la búsqueda, se cambiaban estas entidades por valores de categorías no indexadas (por ejemplo, reemplazar "Nueva York" por "[**Localización**]").

Vico y Calegari [27] proponen una arquitectura de software para anonimizar documentos. La idea principal es reconocer las entidades nominadas con una arquitectura de múltiples herramientas de procesamiento de lenguaje natural. Después del reconocimiento, reemplazan las entidades sensibles por un valor genérico de una categoría indexada (p.e. reemplazar "Fiebre" por "termino_generico_1").

En 2016, el United Kingdom Data Archive (UKDA) lanzó una herramienta de ayuda para la anonimización de texto [23]. Esta herramienta identifica los números y las palabras que empiezan con letra mayúscula y los reemplaza por "XXX".

Kleinberg et al. [13] diseñaron Netanos, una herramienta que permite a los investigadores anonimizar textos largos. Usan aprendizaje automático para reconocer entidades nominadas (por ejemplo, personas, localizaciones, tiempos y fechas) y las sustituyen por valores de categoría indexadas

que preservan la privacidad (tales como “Localización_1”, “Persona_1”).

IV. Metodología

El objetivo de este trabajo es localizar términos en un texto no estructurado que puedan tener implicaciones en la privacidad, ya sea porque pueden ser usados para reindentificar a un individuo o porque contienen información confidencial.

IV-A. Enfoque general

Formalmente, dada una colección de documentos de texto $D_1, \ldots, D_n$, queremos localizar los atributos que tienen relevancia para la privacidad. Específicamente, queremos obtener los atributos identificadores $\mathcal{ID} = \{ID_1, \ldots, ID_p\}$, los atributos cuasi-identificadores $\mathcal{QID} = \{QID_1, \ldots, QID_q\}$, y los atributos confidenciales $\mathcal{C} = \{C_1, \ldots, C_r\}$. El conjunto $\mathcal{ID}$ debería contener los atributos identificadores que aparecen en al menos uno de los documentos; por ejemplo, $\mathcal{ID}$ contendrá “Num Pasaporte” si al menos uno de los documentos contiene el número de pasaporte (incluso si los otros documentos no lo contienen). De forma similar, el conjunto $\mathcal{QID}$ debería contener los atributos cuasi-identificadores que aparecen en al menos un documento, y el conjunto $\mathcal{C}$, los atributos confidenciales que aparecen en al menos un documento.

Una vez determinados estos conjuntos, la colección de documentos puede considerarse como un fichero *estructurado* con registros $D_1, \ldots D_n$ y atributos que son elementos de $\mathcal{ID} \cup \mathcal{QID} \cup \mathcal{C}$. Obviamente, este fichero estructurado probablemente sea escaso, ya que no todos los atributos toman valores en todos los documentos. Para anonimizar este fichero, procedemos como en el caso de ficheros estructurados. Los valores de los atributos en $\mathcal{ID}$ deberían suprimirse de todos los registros/documentos y se deberían enmascarar los atributos en $\mathcal{QID}$ y/o $\mathcal{C}$. Dependiendo del tipo de enmascaramiento usado, puede ser necesario tratar primero los atributos que falten en algunos documentos; una posibilidad es imputarlos usando síntesis parcial [5], [10].

De esta manera, el problema de anonimizar datos no estructurados se reduce a localizar las apariciones de los atributos relevantes para la privacidad en la colección de documentos y anonimizar el conjunto de datos estructurado resultante. Podemos abordar la tarea de localizar las particiones de atributos creando varios modelos de aprendizaje automático, cada uno destinado a reconocer un tipo diferente de entidad nominada. Por ejemplo, un primer modelo para reconocer atributos identificadores (tales como número de pasaporte, número de la seguridad social, etc.), otro para reconocer atributos cuasi-identificadores (tales como localización, fecha de nacimiento, edad, código postal, etc.), y otro más para reconocer atributos confidenciales (como nombres de enfermedades).

IV-B. Prueba de concepto

Como prueba de concepto, nos centramos en localizar datos confidenciales en diagnósticos médicos. Proponemos un modelo basado en CRFs para extraer los nombres de enfermedades de una historia clínica. Dado un texto, este modelo predice una secuencia de etiquetas IOB2.

Una vez predicha la secuencia de etiquetas IOB2 para cada parte de la historia clínica, podemos interpretar esta secuencia de etiquetas y extraer la entidad “enfermedad”. Por ejemplo, si tenemos la frase “La retinopatía fue evaluada por oftalmoscopia” y la secuencia de etiquetas IOB2 {B-DIS, O, O, O}, nos movemos por la secuencia y cada palabra correspondiente a una etiqueta B-DIS se considera el principio de una entidad enfermedad y cada palabra correspondiente a una etiqueta I-DIS se considera dentro de una entidad enfermedad. Por lo tanto, una palabra B-DIS junto con las siguientes palabras I-DIS forman una entidad enfermedad. De hecho, las etiquetas B-DIS e I-DIS tienen la misma finalidad, pero B-DIS se encarga de distinguir dos entidades enfermedad consecutivas.

Tabla I
Extracción de características.

Característica	Descripción
Raíz de la palabra	Por ejemplo, la raíz de “enfermedad” es “enfermo”. Extraemos las raíces usando “SnowballStemmer”, de la librería nltk [2].
Longitud de palabra	Cantidad de letras de la palabra.
Forma de la palabra	Forma de la palabra, la cual puede ser ’minúsculas’, ’mayúsculas’, ’primera mayúscula’, ’combinada’.
Tipo de palabra	Tipo al que pertenece la palabra (nombre, adverbio, adjetivo, etc.). Usamos el marcador “Stanford POS” para extraer esta característica [26].

La Figura 1 muestra la estructura del modelo propuesto para reconocer nombres de enfermedades. Consiste en tres pasos:

- El primer paso es el análisis, que divide una frase en fichas (*tokens*).
- El segundo paso es el extractor de características; en este paso, usamos una ventana de tres palabras (palabra actual, previa y siguiente) y extraemos las características de esas palabras. La Tabla I describe las características que consideramos.
- El tercer paso utiliza un modelo CRF, que toma las características del segundo paso y produce una secuencia de etiquetas para la frase completa.

V. Resultados Experimentales

En esta sección describimos los resultados experimentales de la prueba de concepto mencionada anteriormente. Hemos programado los experimentos en Python y hemos usado “sklearn-crfsuite” para CRF [14] y “SnowballStemmer” para la extracción de raíces de palabras [2].

V-A. Datos

En nuestros experimentos, hemos utilizado textos médicos etiquetados para estudiar la relación entre enfermedades y tratamientos. Estos archivos fueron obtenidos de MEDLINE 2001 usando los 100 primeros títulos y los 40 primeros resúmenes de 59 archivos medline01n*.xml, disponibles en [21].

Estos datos contienen 3645 frases etiquetadas. Las etiquetas son: “DISONLY”, “TREATONLY”, “TREAT PREV”, “DIS PREV”, “TREAT SIDE EFF”, “DIS SIDE EFF”, “DIS VAG”, “TREAT VAG”, “TREAT NO” y “DIS NO”. Como únicamente estábamos interesados en las enfermedades, mantuvimos sólo las 629 frases con la etiqueta “DISONLY”.

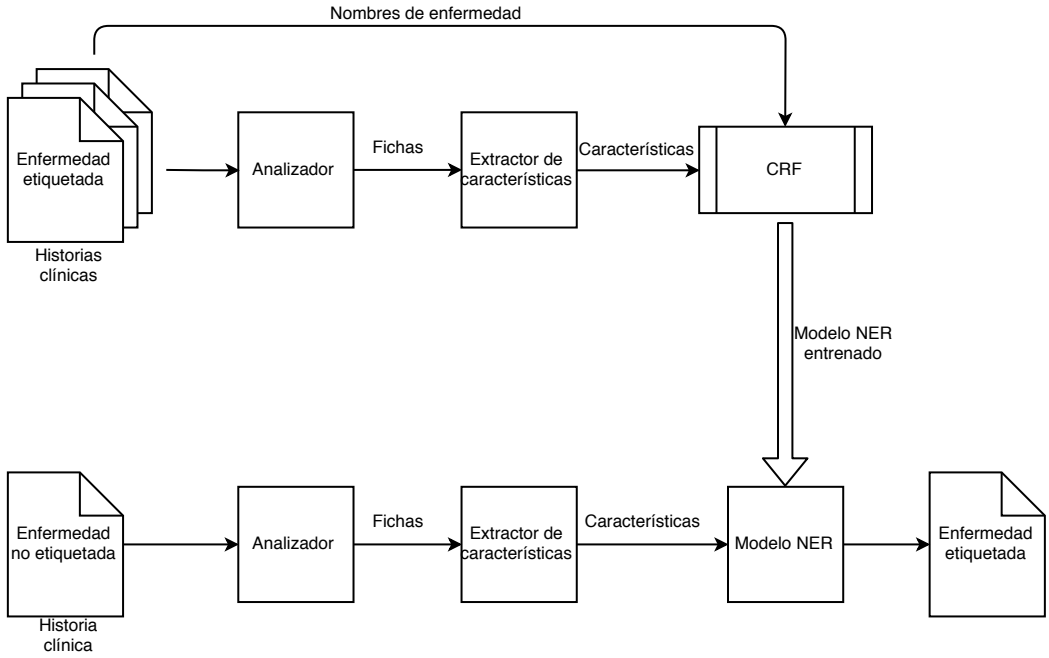


Figura 1. Arquitectura del marcador de reconocimiento de entidades nominadas

V-B. Métricas de evaluación

Usamos tres métricas para evaluar el rendimiento del modelo propuesto para el reconocimiento de enfermedades:

- *Precisión*. Número de enfermedades identificadas correctamente por el clasificador dividido por el número total de enfermedades identificadas:

$$Precision = \frac{|S \cap T|}{|S|},$$

donde S es el conjunto de enfermedades identificadas por el clasificador y T es el conjunto de enfermedades correctas de acuerdo con el fichero original.

- *Recuerdo*. Número de enfermedades identificadas correctamente por el clasificador dividido por el número de enfermedades correctas en el fichero original:

$$Recuerdo = \frac{|S \cap T|}{|T|}.$$

- *F1*. Media armónica de precisión y recuerdo:

$$F_1 = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recuerdo}}{\mathrm{Precision} + \mathrm{Recuerdo}}.$$

V-C. Resultados y discusión

Realizamos la evaluación experimental en dos fases: entrenamiento del modelo y prueba del modelo. De las 629 muestras de frases etiquetadas, 503 fueron usadas para el entrenamiento del modelo (80 % de las muestras), y 126 fueron usadas para la prueba (20 % de las muestras).

La fase de entrenamiento se llevó a cabo usando 10 veces una validación cruzada, explicada a continuación. Partimos el fichero destinado al entrenamiento en 10 submuestras del mismo tamaño. De las 10 submuestras, una la mantuvimos como datos de validación para probar el modelo durante esta fase, y las nueve restantes se usaron para entrenar el modelo.

Tabla II
EVALUACIÓN DEL MODELO EN EL FICHERO DE PRUEBA A NIVEL DE PALABRA

	Precisión	Recuerdo	Puntuación F1
B-DIS	0.766	0.677	0.719
I-DIS	0.789	0.709	0.747
promedio / total	0.778	0.693	0.733

Aunque muchas de las palabras del fichero fueron etiquetadas como O (fuera de enfermedad), estábamos interesados en palabras etiquetadas como B-DIS (principio de enfermedad) e I-DIS (dentro de enfermedad). Por lo tanto, calculamos la precisión, el recuerdo y la puntuación F1 sólo para B-DIS e I-DIS. Por ejemplo, si tenemos la frase "Evaluación diagnóstica del paciente con presión sanguínea elevada", las partes de la frase son {"Evaluación", "diagnóstica", "del", "paciente", "con", "presión", "sanguínea", "elevada"} y sus correspondientes etiquetas {O, O, O, O, O, B-DIS, I-DIS, I-DIS}. La entidad nominada contiene tres palabras "presión sanguínea elevada". La Tabla II muestra la evaluación de la predicción de etiquetas comparada con las etiquetas correctas a nivel de palabra (de forma separada para cada palabra). En cambio, la Tabla III muestra las mismas métricas de evaluación para entidades enteras. Es decir, en el ejemplo anterior, la Tabla **??** se refiere por separado a las tres palabras "presión", "sanguínea" y "elevada", mientras que la Tabla III se refiere a la entidad "presión sanguínea elevada"; en este último caso, a menos que *las tres palabras por completo* de la entidad fuesen correctamente etiquetadas, la entidad entera se consideraría desclasificada.

De acuerdo con la Tabla III, nuestro modelo funcionó significativamente mejor en precisión que en recuerdo. Es muy posible que el valor de recuerdo pueda incrementarse

Tabla III
EVALUACIÓN DEL MODELO EN EL FICHERO DE PRUEBA A NIVEL DE ENTIDAD

	Precisión	Recuerdo	Puntuación F1
Entidad Enfermedad	0.742	0.660	0.698

usando más muestras de entrenamiento. Consideramos que los resultados anteriores son prometedores, ya que se acercan a los del etiquetado manual de los datos usados.

VI. CONCLUSIONES Y TRABAJO FUTURO

En este trabajo, hemos tratado la anonimización de datos no estructurados en forma de texto. Como prueba de concepto, nos hemos centrado en localizar nombres de enfermedades (es decir, atributos confidenciales) en historias clínicas. Una vez localizados, estos atributos confidenciales pueden protegerse usando técnicas estándar de SDC para datos estructurados.

La principal contribución de este trabajo se basa en la arquitectura del reconocedor de entidades nominadas. El modelo propuesto está basado en aprendizaje automático y es mejor que los enfoques de NER basados en diccionarios. Específicamente, evita el problema que surge cuando las entidades que se quiere localizar no aparecen en el diccionario usado.

Como trabajo futuro, planeamos extender la prueba de concepto presentada para la detección de identificadores y cuasi-identificadores. Esto conllevará un esfuerzo considerable para generar ficheros que incluyan notas para atributos como nombres, localizaciones, edades, etc. Estos ficheros serán usados posteriormente para entrenar los modelos de detección de identificadores y cuasi-identificadores.

AGRADECIMIENTOS

Agradecemos los fondos recibidos de las siguientes organizaciones: Comisión Europea (proyectos H2020-644024 “CLARUS” y H2020-700540 “CANVAS”), Generalitat de Cataluña (Premio ICREA-Acadèmia para J. Domingo-Ferrer) y Gobierno de España (proyectos TIN2014-57364-C2-1-R “SmartGlacis” y TIN2015-70054-REDC). Las opiniones de este documento son de los autores y no reflejan necesariamente las opiniones de UNESCO o cualquiera de los financiadores.

REFERENCIAS

[1] B. Babych y A. Hartley. Improving machine translation quality with automatic named entity recognition. En *Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT Through Other Language Technology Tools: Resources and Tools for Building MT (EAMT ’03)*, pp. 1–8. Association for Computational Linguistics, 2003.

[2] S. Bird, E. Klein y E. Loper. *Natural Language Processing with Python — Analyzing Text with the Natural Language Toolkit*. O’Reilly, 2009. El Natural Language Tooking software (nltk) está disponible en: https://www.nltk.org

[3] A. Culotta, R. Bekkerman y A. McCallum. *Extracting Social Networks and Contact Information from Email and the Web.* Computer Science Department Faculty Publication Series, no. 33. University of Massachusetts-Amherst, 2004.

[4] J. Domingo-Ferrer, D. Sánchez y J. Soria-Comas. *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Morgan & Claypool, 2016.

[5] J. Drechsler. *Synthetic Datasets for Statistical Disclosure Control*. LNS 201. Springer, 2011.

[6] A. Ekbal, R. Haque y S. Bandyopadhyay. Bengali part of speech tagging using conditional random field. En *Proceedings of the Seventh International Symposium on Natural Language Processing (SNLP-2007)*, 2007.

[7] J. R. Finkel, T. Grenager y C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. En *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL ’05)*, pp. 363–370. Association for Computational Linguistics, 2005.

[8] EU General Data Protection Regulation, 2016/679. https://gdpr-info.eu

[9] S. Grimes. Structure, models and meaning. *Intelligent Enterprise*, Mar. 2005.

[10] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte-Nordholt, K. Spicer y P.P. de Wolf. *Statistical Disclosure Control*. Wiley, 2012.

[11] M. Jabreel, F. Hassan y A. Moreno. Target-dependent sentiment analysis of tweets using bidirectional gated recurrent neural networks. In *Advances in Hybridization of Intelligent Methods*, pp. 39–55. Springer, 2018.

[12] M. A. Khalid, V. Jijkoun y M. De Rijke. The impact of named entity normalization on information retrieval for question answering. En *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval (ECIR’08)*, pp. 705–710. LNCS 4956, Springer, 2008.

[13] B. Kleinberg, M. Mozes, Y. van der Toolen y B. Verschuere. *Netanos - Named Entity-based Text Anonymization for Open Science*. Open Science Framework, Jan. 31, 2018. https://osf.io/w9nhb

[14] M. Korobov. sklearn-crfsuite, 2015. https://sklearn-crfsuite.readthedocs.io/en/latest/

[15] J. Lafferty, A. McCallum y F. C. N. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. En *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, pp. 282-289. ACM, 2001.

[16] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang y H. Xu. A comprehensive study of named entity recognition in Chinese clinical text. *Journal of the American Medical Informatics Association*, 21(5):808–814, 2013.

[17] S. Morwal, N. Jahan y D. Chopra. Named entity recognition using hidden Markov model (HMM). *International Journal on Natural Language Computing*, 1(4):15–23, 2012.

[18] D. Nadeau y S. Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, 2007.

[19] I. Neamatullah, M. M. Douglass, L. H. Lehman, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark y G. D. Clifford. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):32, 2008.

[20] R. Pérez-Laínez, A. Iglesias y C. de Pablo-Sánchez. Anonimytext: anonymization of unstructured documents. Universidad Carlos III de Madrid, 2009. https://e-archivo.uc3m.es/handle/10016/19829

[21] B. Rosario y M. A. Hearst. Classifying semantic relations in bioscience texts. En *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*. Association for Computational Linguistics 2004. Datos disponibles en: http://biotext.berkeley.edu/dis_treat_data.html

[22] E. F. Sang y J. Veenstra. Representing text chunks. En *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 173–179. Association for Computational Linguistics, 1999.

[23] United Kingdom Data Service. *Text Anonymization Helper Tool*. Último acceso: Mar. 24, 2018. https://bitbucket.org/ukda/ukds.tools.textanonhelper/wiki/Home

[24] B. M. Sundheim. Overview of results of the MUC-6 evaluation. En *Proceedings of the TIPSTER Text Program: Phase II*, pp. 423–442. Association for Computational Linguistics, 1996.

[25] L. Sweeney. Replacing personally-identifying information in medical records, the Scrub system. En *Proceedings of the AMIA Annual Fall Symposium*, p. 333. American Medical Informatics Association, 1996.

[26] K. Toutanova, D. Klein, C. Manning y Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. En *Proceedings of HLT-NAACL 2003*, pages 252-259. 2003.

[27] H. Vico y D. Calegari. Software architecture for document anonymization. *Electron. Notes Theor. Comput. Sci.*, 314(C):83–100, 2015.

[28] G. Zhou y J. Su. Named entity recognition using an HMM-based chunk tagger. En *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL’02)*, pp. 473–480. Association for Computational Linguistics, 2002.

Privacy in Microblogging Online Social Networks: Issues and Metrics

Samia Oukemeni
Internet Interdisciplinary Institute (IN3)
Universitat Oberta de Catalunya
CYBERCAT - Center for Cybersecurity Research of Catalonia
Barcelona, Spain
soukemeni@uoc.edu

Helena Rifà-Pous
Internet Interdisciplinary Institute (IN3)
Universitat Oberta de Catalunya
CYBERCAT - Center for Cybersecurity Research of Catalonia
Barcelona, Spain
hrifa@uoc.edu

Joan Manuel Marquès Puig
Internet Interdisciplinary Institute (IN3)
Universitat Oberta de Catalunya
Barcelona, Spain
jmarquesp@uoc.edu

***Resumen*—In the era of digital life, privacy is threatened by the rapid expansion of online social media, especially Microblogging Online Social Networks (MOSNs). Whether users are connected in real life or not, microblogging systems are a popular form of Online Social Networks (OSNs) that allow users to post short messages, share interests, and communicate with each other. Despite this popularity, they have a poor reputation in terms of protecting the privacy of users. The present paper examines different types of privacy threats and concerns in MOSNs. The threats can be either classic and common to other online platforms or specific to the environment of OSNs. In addition, this article describes various models that assess and quantify privacy in OSNs and finally proposes a new generic framework to evaluate the privacy in MOSNs. The proposed framework indicates the level of privacy protection provided by the system and helps to compare different MOSNs.**

***Index Terms*—Microblogging online social networks, privacy, security, issues, measurements, metrics, framework;**

I. Introduction

Privacy is recognized internationally as a fundamental human right. However, in this data-driven world, protecting privacy becomes one of the dominant issues as it is threaten by the terabytes of personal information revealed everyday. Privacy is an intuitive term that can be interpreted differently from abstract and wide to technical and specific. It is a common concept yet not easy to define. According to Oxford dictionary [1], privacy is defined as:

> ” A state in which one is not observed or disturbed by other people”.

The protection of privacy is usually associated with cryptography, authorization, and anonymization. Westin in 1968 defined privacy as ” the right to select what personal information about me is known to what people” [2]. Bünnig et al. [3] described privacy as protecting personal information from malicious and unauthorized entities. Privacy can also be defined as a set of policies that enforces the protection of private information [4]. Another definition of privacy is to hide some details from others [5]. All these different explanations and yet no single definition of privacy encompasses all aspects of the term.

Due to the exponential development of information technologies, protecting the privacy becomes extremely important, especially in the field of social media and microblogging services. A microblogging network is a popular form of Online Social Networks. It is a weblog where users are allowed to send snippets of a small number of characters (between 140 and 310 characters) [6]. The number of registered users for microblogging services is increasing each month. It reaches nearly 1 billion monthly active users between Twitter, Tumblr, etc. [7], [8]. MOSNs have become a source of news coverage and means of propagating all sorts of information [9], [10] as it happens in the coverage of the 2016 US presidential election [11].

As different as they can be, all microblogging systems offer to their users the ability to create customized profiles, follow and be followed, share interests and keep updated on the trending topics and news [12]. Users can include multimedia content in their posts, like pictures, links, or video links, and they can keep track of activities from other users, trends, companies, brands or celebrities [13].

However, such information attracts malicious users and attackers to gather, aggregate and exploit the data generated by users to commit cyber crimes against the users such identity theft, phishing, social scams, and social engineering. For example, the website Please Rob Me [14] raises the awareness of the danger of oversharing in social media, especially the information about the geographical location on Twitter. The website scans the feeds from Twitter and shows when the users tweet out locations other than their home. Furthermore, most of the known microblogging companies handle the users' data and generate their revenue by gathering and selling the data to third-party channels for advertisement or statistical purposes [15]. Despite the fact that MOSNs might provide privacy settings to fine-tune the visibility of profiles and posts, they often remain insufficient. Furthermore, privacy policies offered by the systems are too vague and expressed in a language that makes them difficult for users to understand how their information is handled and shared [18]. In fact, on the 17th of March 2018, a former Cambridge Analytica contractor admitted that the firm harvested more than 50 million Facebook profiles without permission to build an algorithm that targeted US voters with personalized political advertisements based on their psychological profile [16]. The data was collected through an app called ”thisisyourdigitallife”, built for academic purposes [17].

MOSNs have brought new challenges to privacy-oriented companies and academic community. Researchers have dis-

cussed new privacy-preserving controls and techniques and they have proposed different microblogging online social systems to protect and enhance the users' privacy like Diaspora [19] and Galaxy 2 [20]. As the definition of privacy is ambiguous and elusive, there is no standard means of how to build an efficient privacy-protecting system. Thus, the need for techniques to evaluate and measure the privacy of MOSNs. Some research analyzed the privacy settings provided in the systems to evaluate the privacy level [21]. Some other works proposed new models based on mathematical formulas to assess the privacy in the MOSNs. Our findings show that in spite of the number of proposed methods for measuring the privacy, there is no comprehensive framework that takes in all the aspects of privacy goals, nor a general proposal for measuring any MOSN and so, a global evaluation and comparison between different systems is not possible.
In this paper, we summarize privacy issues and challenges in MOSNs as described in section II. Section III provides an extensive catalog of privacy-specific measurement models in OSNs in general. This paper also contributes to the former subject by providing a comprehensive privacy measuring framework. An overview of the framework is described in section IV. Section V presents the conclusion of the present paper, including future work.

II. Privacy in Social Networks

Privacy has been a concern, even before the rise of new information technologies. But with the booming of online applications and services that require creating profiles with personal information, several data breaches targeting private information have intensified. The best-known example of data breach happened in 2014 when Sony Corp suffered the biggest data breach known in modern times, it was even labeled "the Hack of the Century". The cyber-invasion cost Sony Corp billions US dollars not only in terms of financial losses but also in terms of reputation. The attackers stole confidential internal documents such as movie scripts and highly classified and personal information of Sony employees (internal emails, salaries, and more than 47,000 social security numbers) [22]. These repeated incidents have highlighted the need to protect sensitive and private data, beyond deploying basic access control policies. As one of the answers to this need, the EU Parliament approved the General Data Protection Regulation (GDPR) on 14 April 2016. The new regulation was designed to harmonize data privacy laws across Europe and to protect the privacy of the users. It gives the users the right to access, download and erase their data. Also, the concept of privacy by design becomes a legal requirement in building new online platforms. GDPR is enforced by 25 May 2018 [23]

II-A. Understanding Privacy in the Context of MOSNs

MOSNs are now a trendy way for users to express themselves and connect with their entourage. Compared to regular OSNs, microblogging networks meet the need of a faster mode of communication to reach a larger group of people. By using shorter posts, the time to write and post an update is drastically shorten which allows a microblogger to post frequently several messages every day. Also, the popularity of MOSNs is due to the openness and the flexibility provided to the users. They can connect and communicate with their favorite celebrity, brand, politician, athlete or even other regular users without the obligation of a pre-existing social relationship.
To stay active in an MOSN, the users voluntarily share personal information about themselves without prior knowledge of who can access their private data or how it is handled by the service providers. A study of Twitter users sharing behavior shows that between 40 % to 50 % of tweets include personal information about the author, including personally identifiable information, contact data, health information and location data [24]. This fact of providing sensitive information willingly makes the protection of privacy more complex.
Furthermore, nearly all online companies like Twitter and Tumbler generate their profits by gathering, storing and processing users' data in order to sell them for advertisement or statistical purposes. These companies claim that before selling the data, they anonymize them, meaning that they remove any explicit information from the dataset that can directly identify the users (name, Social Security Numbers (SSNs)...) [25]. However, recent research [26] indicates that from the anonymized dataset of 1.5 million people, a person can be identified with 95 % accuracy in only four spatiotemporal points. Moreover, Montjoye et al. [27] studied an anonymized dataset of credit card transactions of 1.1 million people and were able to re-identify 90 % of individuals knowing again only four spatiotemporal points. In other words, it is simple and easy to identify a person based on non-identifying attributes (sex, birthdates..). Therefore, posting news and interest in social media can put the privacy of data in jeopardy when it is placed in the wrong hands.

II-B. Privacy Issues and Attacks in MOSNs

With the increasing popularity, microblogging systems and online social networks in general have become a hub for cybercriminal activities. Attackers and malicious users are drawn to these platforms and specifically to the critical data revealed, intentionally or unintentionally, by the users. With simple fake accounts, the adversary can engage the victims and spread malicious contents. Most of the attacks are driven by the purpose of harassment, identity theft and stealing information related to bank accounts or social security numbers. [28].

Some privacy risks are more amplified in MOSNs compared to traditional service systems. These include:

- Malicious insiders that can connect with the victims and act as legitimate users.
- Unintentional disclosure of personal information from users like geographic location, interests, etc.
- The joint utilization of different online social networks that can bring in a new type of attacks based upon the fusion of multiple profiles of the same user across multiple OSNs.
- Third-party applications that use the API provided by the MOSNs and they can access the users' profiles. Also, these applications may have vulnerabilities that attackers can exploit to get to the users' accounts. For example, a vulnerability in Twitter Counter, a popular tool for analyzing Twitter followers, was exploited in 2017, which has led to taking control of hundreds of high-profile Twitter accounts like the European Parliament, UNICEF, and Amnesty International [29].

In addition to classic attacks on any online platform, like Denial-of-Service (DoS) Attack, SQL injection, Cross-Site Scripting XSS, social spamming, flooding, phishing and malware attacks [28], [30], [31], [32], there are other specific social media attacks. For example:

- Identity clone and theft attacks aim to create a fake profile in the system or clone and duplicate user's online presence to fool other users and commit fraud or espionage [33].
- Social profiling refers to the process of collecting information and constructing a user's profile. This occurs through aggregating information that is publicly and voluntarily published in MOSNs [34], [35].
- Social link prediction and disclosure occurs when an adversary predicts or discloses hidden links and masks relationship between two users, a relationship that the users would like to remain hidden from the public [36].
- Conversation and communication tracking is a type of profiling attack where the adversary tracks the communication feed of users and collects information about the users and their interests. This allows the adversary to create a more detailed user profile [37].
- Sybil attack or fake profiles attack [30], [38] is an attack where a single entity masquerades as multiple identities with the objective to make as many friends as possible with legitimate accounts. The influence of the adversary increase in the network and thereby engages in malevolent activities like spamming and identity fraud. This type of attacks is used also to increase the visibility of the content and manipulate the view counts and the network decisions.
- Clickjacking is used to trick users to redirect them to malicious sites by clicking on attractive buttons or links in a post [28], [39].
- De-anonymizing attacks are used to re-identify a particular user in an anonymized dataset [30].
- Crawling is the collection and aggregation of available information across the profiles of multiple users in the MOSN. In this attack, the adversary doesn't target one particular user. The information gathered can be used for users' activities analysis or in marketing advertising [37], [40].

III. Measuring and Evaluating Privacy in MOSNs

With the large number of threats and attacks on MOSNs, privacy oriented service providers and researchers have introduced new systems that offer microblogging functionalities and at the same time advocate for privacy protection. Some systems add a privacy layer while others are built using privacy by design methodologies.

With a multitude of privacy controls and techniques implemented in these new MOSNs, a necessity of services evaluating models appeared. These models are used to evaluate the effectiveness of these private networks. However, the challenge is how to measure and evaluate the privacy in MOSNs since defining privacy itself is a challenging issue. Privacy is subjective because it is related to what people consider sensitive, i.e. which information each person wants to keep secret. This can change based on the context and with the course of the time.

III-A. Metrics and Measurements

When we talk about privacy assessment and evaluation, we talk about metrics. A metric is defined as a system of measurement, i.e. the techniques and procedures, that evaluate and quantify an issue [41]. Metric and measurements are similar enough that the two terms are commonly used interchangeably. However, there is a difference between the two terms: measurements provide single-point-in-time views of specific factors while metrics provide standardized procedures and calculation methods to generate relevant numbers of the measured system.

III-B. Privacy Evaluation and Scoring in MOSNs

Privacy metrics can be used for decision making and in assessing, monitoring and predicting potential privacy threats in the system. Privacy evaluation empowers the academic community with a strong understanding of privacy and a better protection of information in the MOSNs.

In online social networks, some attempts to evaluate and quantify the privacy are found in the literature. In 2009, Maximilien et al.[43] proposed a framework to calculate the privacy score based on the sensitivity and the visibility of attributes of a social network. They conducted a survey where the questions were designed to determine the privacy degree that users were willing to disclose each information in their profiles. The authors didn't offer any dataset to measure the effectiveness of their model. Liu and Terzi extended the approach in [44]. They developed a mathematical model to measure the privacy score of the users, based on the sensitivity and the visibility of attributes, using concepts from Item Response Theory (IRT) [45]. To evaluate the effectiveness of the score, the authors used both synthetic and real-world datasets. However, this model is not generalized to any kind of OSN since it assumes that the users are independent, the attributes are independent and it doesn't take into consideration the inferred data. Srivastava and Geethakumari extended Liu et al.'s model and included the hidden data in [46]. They also introduced privacy leakage that quantifies the privacy exposure for some user from a message. Both models [44] and [46] assumed that the sensitivity and visibility are the same across all users. Petkos et al. [47] enhanced the previous models and proposed a PScore framework. PScore considers the user's personal preferences in scoring the attributes, it includes the hidden and inferred information and it is structured based on different types of information. Pensa and Di Blasi introduced a new privacy assessment framework in [48]. This work was inspired from the model proposed by Liu and Terzi [44], it takes into consideration the circle (friends) of the users where the willingness ratio of a user to disclose information is proportional to the number of her or his friends. The framework measures the privacy leakage and set a model of privacy preferences for each user. If the score exceeds a given threshold, the framework notifies the user about the privacy risk. The privacy score is based on both the sensibility and the visibility of user profile attributes.

Some researchers took another approach to evaluate privacy beyond the sensitivity and visibility properties. Becker et al. [49] introduced PrivAware, a tool that quantifies the privacy risk from the amount of information inferred in social networks. PrivAware maps the privacy risk to a grading score and

sets recommended actions for users. Ngoc et al. [50] presented a privacy metric calculated based on the probability and entropy theory. This metric quantifies the information leaked in the users' posts. The authors built the metric based on the idea of how much an attacker can reveal of hidden sensitive information of a user from the sentences in the posts. Talukder et al. [51] proposed Privometer to measure the leakage of sensitive information based on the profiles of users and their social graphs. Privometer takes into consideration also the leakage of sensitive information from applications installed in the user's friend profiles. Privometer ranks the relationships of users based on the amount of information leakage and suggests self-sanitization recommendations to control the leakage. Akcora et al. [52] suggested measuring the risk score based on the feedback from users about others users and the sensitive information disclosure. The framework computes the risk level in terms of the friends' attitude and the similarities with the users. The authors used Facebook and real datasets to evaluate the effectiveness of the model. Vidyalakshmi et al. [53] developed a privacy scoring framework of friends to assist users in assessing their information sharing behavior and in taking a decision of who can see what information. The scores are based on the output of a friend and its position in the sorted list of friends and the total number of friends of the user.

Nepali and Wang [54] introduced a new model to monitor privacy exposure in real time. The model uses real data from social networks, instead of data from surveys. The privacy risk indicator (PIDX) is calculated based on the sensitivity and the visibility of attributes. The obtained value is used for privacy monitoring and risk control. SONET is based on 2 components, attribute to attribute (actor model) and user to user relationships (community model). SONET included hidden information that is not firsthand available, but it infers from direct data. The model is used to monitor the level of privacy in the system and to protect users from sensitive information disclosure. The authors extended the actor model of SONET in [55] and the community model in [56]. They included 3 metrics: known attribute list (direct, hidden and virtual), attributes sensitivity and attributes visibilities. The authors introduced the OSNPIDX tool in [56] as an implementation of SONET model. OSNPIDX defines an actor with 20 static-assigned privacy impact factor attributes.

The existing work in the field of privacy scoring models reveals that it is rather limited and is still relatively unexplored. Additionally, all the models evaluate privacy in OSNs from the user's perspective (the profile, the visibility and sensitivity of the items published, the information leakage from social graph...). There is no holistic view on evaluating the privacy in social networks and clearly, it still lacks a generic framework that evaluates the privacy and not specific to only one aspect or to one MOSN. *Table I* summarizes the reviewed privacy scoring approaches.

IV. Privacy Measurement Framework for Microblogging Online Social Networks

As explained in the above (see section III), all the reviewed frameworks and scores focus on one aspect to assess and evaluate the privacy in OSNs, either they compute privacy scores based on the sensitivity and visibility of some attributes in the systems or they measure the information leakage from social graphs. Some of these frameworks are system-specific tools that can be used to evaluate only one specific OSN. They don't include other aspects of privacy protection goals like how the system protects the confidentiality of the data and users, or how the storage type can affect the privacy of the data. Also, all the frameworks focus on the impact of the visibility of data to other users but not to the system provider. In response to the limitations of these privacy evaluation frameworks, we propose an enhancing generic privacy scoring framework to quantify, assess and evaluate privacy in MOSNs. The framework is generic and universal, and allows comparing the privacy protection between different systems. Our proposed framework is designed with these considerations in mind:

Figura 1. PDSA life cycle

- It must be generic and applicable to different MOSNs and also it can be fine-tuned to meet the needs of specific situations and be modified to fit the needs of specific system.
- It must take into consideration all the MOSNs stakeholders namely: the users (profile, relationships, groups...), the data generated by the users, the system itself and third-party components.
- It must include security metric as well because of the existing synergy between security and privacy and also, because security is necessary to achieve privacy.
- It must take into consideration all types of data as proposed in [57].

The proposed framework follows a four-steps methodology inspired by the Plan-Do-Study-Act cycle (PDSA) [58] (See *Fig.1*):

- **Step 1** sets the boundaries, it determines the objectives and the scope of of the framework.
- **Step 2** follows the Goal-Question-Metric (GQM) paradigm [59] to design the engine of the framework. The engine monitors and gathers data from different sources, including user feedback, risk assessment reports, research surveys, event loggers..., and answers the questions defined in each goal. We define 4 metrics that provide quantitative information to answer the following goals:
  - Metric 1: How the system protects itself from the privacy and security point of view.
  - Metric 2: How the privacy and the security of data are handled in the system.
  - Metric 3: How the system protects the users and the data.
  - Metric 4: How various assumptions and functions in the system might affect the privacy and the security.

Tabla I
OVERVIEW OF THE REVIEWED PRIVACY SCORING APPROACHES

Privacy score	Description	Features
Privacy Scores: [43], [44], [46], [47], [48]	A score generated based on the sensitivity and visibility of the items posted by an OSN user. Some scoring frameworks take into consideration the hidden and inferred information and the circle (friends) of the users.	Profile items, sensitivity per item, visibility per item. For some articles, leakage per item is also included.
PrivAware: [49]	The score is calculated based on the total of visible attributes divided by the total of attributes in a profile.	The amount of information inferred in social networks
Privometer: [51]	A score generated based on the sensitivity of the profiles of users and their social graphs.	Sensitive attribute inference from the information available in immediate friends' profiles.
Privacy Index: [54], [55], [56]	A score calculated based on the sensitivity and the visibility of public attributes.	Sensitivity score per item, visibility level per item.
Privacy score from social graphs: [52], [53]	A score calculated based on the risk level in terms of the friends attitude and the similarities with the users.	How much an attacker can reveal of relationships

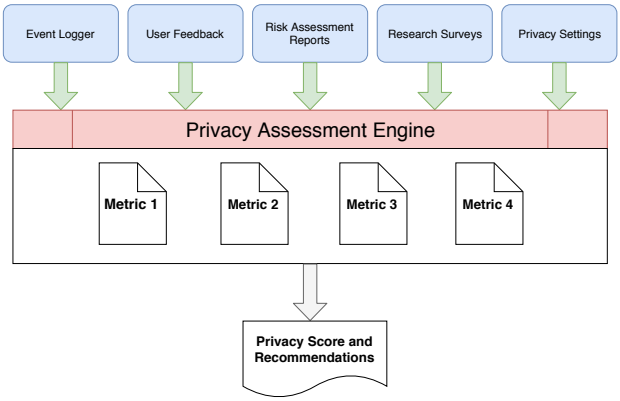


Figura 2. Privacy scoring process flow

The proposed framework, then, computes an overall privacy score based on the assessment of the impact of information security objectives (confidentiality, integrity, and availability) in addition to the privacy objectives as defined by the National Institute of Standards and Technology (NIST): predictability, manageability and disassociability [60]. Based on the obtained score, the framework offers suggestions and recommendations for effectively controlling the privacy and security of the system. The process of building the engine is summarized in *Fig.2*.

- **Step 3** determines whether the results of the proposed framework are accurate, clear, understandable and fully explained in case of uncertainty.
- **Step 4** is based on the results of step 3, it derives what can be changed and improved in the framework.

It is envisioned that, due to its generality, the framework will enable the analysis of privacy in MOSNs in more details. At the same time, it will empower the ability to measure the performance, the efficiency, and the effectiveness of the controls and settings provided in terms of privacy and security and to gauge how well the system under investigation is meeting the privacy and security objectives, thereby reducing the number of threats and attacks on MOSNs.

V. CONCLUSION

In this paper, we discussed two important and trending topics in microblogging online social networks: privacy challenges and issues and privacy measurements and evaluation. While users enjoy sharing interests and connecting with friends via online social media, a large amount of personal information becomes accessible to everyone. In addition, advanced data retrieval and analytic techniques have made it easier for attackers to collect and aggregate unlimited data and to target different types of attacks. This paper discussed also some privacy metrics to measure and quantify the privacy in social networks with the goal of evaluating the privacy choices in a system. Our study shows that all the proposed metrics are system-specific and there is a lack of a generic model. To answer this lack, we presented a high-level overview of a comprehensive framework to assess and evaluate the privacy of social media and compare different OSNs. As future work, we plan to implement and extend the proposed framework on different types of microblogging social networks.

ACKNOWLEDGMENTS

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness (TRA2013-48180-C3-P, TRA2015-71883-REDT, TIN2014-57364-C2-2-R and TIN2015-70054-REDC), FEDER, and the Erasmus+ Program (2016-1-ES01-KA108‐02346).

REFERENCIAS

[1] "Privacy — Definition of privacy in English by Oxford Dictionaries", in *Oxford Dictionaries — English*, [Online]. Available: https://en.oxforddictionaries.com/definition/privacy.

[2] A.F. Westin: "Privacy And Freedom", in *25 Wash. & Lee L.*, Rev. 166, 1968.

[3] C. Bunnig, and C.H. Cap: .Ad Hoc Privacy Management in Ubiquitous Computing Environments", in *Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services*, 2009.

[4] Q. Ni, E. Bertino, J. Labo, C. Brodie, C.M. Karat, J. Karat, and A. Trombeta: "Privacy-Aware Role-Based Access Control", in *ACM Transactions on Information and System Security (TISSEC)*, vol. 13, n.3, pp.24-31, 2010.

[5] S. Taheri, S. Hartung and D. Hogrefe: .Achieving Receiver Location Privacy in Mobile Ad Hoc Networks,in *2010 IEEE Second International Conference on Social Computing, Minneapolis, MN*, pp. 800-807, 2010.

[6] A. Java and X. Song and T. Finin, and B. Tseng: "Why We Twitter: Understanding Microblogging Usage And Communities", in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis - WebKDD/SNA-KDD 07*, pp. 56–65, 2007.

[7] Statista: "Twitter MAU worldwide 2017 — Statistic", in *Statista*, [Online]. Available: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

[8] Statista: "Tumblr: total number of blogs 2018 — Statistic", in *Statista*, [Online]. Available: https://www.statista.com/statistics/256235/total-cumulative-number-of-tumblr-blogs/.

[9] M. Broersma and T. Graham: "Twitter As News Source", in *Journalism Practice* , vol. 7, pp. 446-464, 2013.

[10] C. Chamberlain: "How Has Twitter Changed News Coverage?", in *News Bureau — Illinois, 22-Oct-2015*, [Online]. Available: https://news.illinois.edu/view/6367/267046.

[11] M. Das Sarma: “Tweeting 2016: How Social Media is Shaping the Presidential Election”, in *Inquiries Journal*, 2016. [Online]. Available: http://www.inquiriesjournal.com/a?id=1454.

[12] H. Kwak and C. Lee and H. Park, and S. Moon: “What Is Twitter, a Social Network or a News Media?”, in *Proceedings of the 19th international conference on World wide web - WWW 10*, pp. 591–600, 2010.

[13] T. Aichner and J. Frank: ”Measuring the Degree of Corporate Social Media Use”, in *International Journal of Market Research 57*, n. 2, pp. 257-276, 2015.

[14] “Raising Awareness About Over-Sharing”, in *Please Rob Me*. [Online]. Available: http://pleaserobme.com/.

[15] P. Gadkari: “How Does Twitter Make Money?”, in *BBC News*, 2013. [Online]. Available: http://www.bbc.com/news/business-24397472.w

[16] P. Greenfield: “The Cambridge Analytica Files: The Story so Far,” in *The Guardian*, 2018. [Online]. Available: https://www.theguardian.com/news/2018/mar/26/the-cambridge-analytica-files-the-story-so-far.

[17] E. Graham-Harrison and C. Cadwalladr: “Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach,” in *The Guardian*, 2018. [Online]. Available: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.

[18] “Gpen Privacy Sweep 2017 Finds Ambiguity in Privacy Policies”, in *Office of Information and Data Protection Commissioner*, 2017. [Online]. Available: http://www.idp.al/2017/10/25/gpen-privacy-sweep-2017-finds-ambiguity-in-privacy-policies/?lang=en

[19] “Welcome to JoinDiaspora*”, in *diaspora* social network*. [Online]. Available: http://www.joindiaspora.com/.

[20] “Galaxy2”, in *Omeka RSS*. [Online]. Available: https://socialmediaalternatives.org/archive/collections/show/10.

[21] A. Srivastava and G. Geethakumari: “A Framework to Customize Privacy Settings of Online Social Network Users”, in *2013 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pp. 187–192, 2013.

[22] P. Elkind: “Sony Pictures: Inside the Hack of the Century”, in *Fortune.com*, pp. 65–88, 2015.

[23] “Home Page of EU GDPR,” in *EU GDPR Portal*, 2018. [Online]. Available: https://www.eugdpr.org/eugdpr.org.html.

[24] C. Honeycutt and S.C. Herring: ”Beyond Microblogging: Conversation and Collaboration via Twitter”, in *Proceedings of the 42nd Hawaii International Conference on System Sciences (HICSS-42)*, pp. 1–10, 2009.

[25] B.C. Fung and K. Wang and and P.S. Yu: “Anonymizing Classification Data for Privacy Preservation”, in *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, n. 5, pp. 711–725, 2007.

[26] Y.A.D. Montjoye and C.A. Hidalgo and M. Verleysen, and V.D. Blondel: “Unique in the Crowd: The Privacy Bounds of Human Mobility”, in *Scientific Reports*, vol. 3, n. 1, 2013.

[27] Y.A.D. Montjoye and L. Radaelli and V.K. Singh and A. Pentland: “Unique in The Shopping Mall: On The Reidentifiability of Credit Card Metadata”, in *Science*, vol. 347, n. 6221, pp. 536–539, 2015.

[28] G. Nalinipriya and M. Asswini: “A Survey On Vulnerable Attacks in Online Social Networks”, in *International Confernce on Innovation Information in Computing Technologies*, pp. 1–6, 2015.

[29] J. Russell: “Prominent Twitter Accounts Compromised After Third-Party App Twitter Counter Hacked”, in *TechCrunch*, 2017. [Online]. Available: https://techcrunch.com/2017/03/15/twitter-counter-hacked/.

[30] L. Bilge and T. Strufe and D. Balzarotti, and E. Kirda: “All Your Contacts Are Belong to Us: Automated Identity Theft Attacks On Social Networks”, in *Proceedings of the 18th international conference on World wide web - WWW 09*, pp. 551–560, 2009.

[31] G. Yan and G. Chen and S. Eidenbenz and N. Li: “Malware Propagation in Online Social Networks”, in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security - ASIACCS 11*, pp. 196–206, 2011.

[32] I. Kayes and A. Iamnitchi: “Privacy and Security in Online Social Networks: A Survey”, in *Online Social Networks and Media*, vol. 3-4, pp. 1–21, 2017

[33] O. Goga and G. Venkatadri and K. Gummadi: “Exposing Impersonation Attacks in Online Social Networks”, in *ACM CONFERENCE ON ONLINE SOCIAL NETWORKS (COSN’14)*, 2014.

[34] D. Chandrasekaran and D. Costello and P. Stubbs: ”Social Media Profiling”, in *US Patent App. 13/465,335, issued November 7, 2013*, 2013.

[35] Electronic Privacy Information Center: “EPIC - Privacy and Consumer Profiling”, in *Electronic Privacy Information Center*. [Online]. Available: https://www.epic.org/privacy/profiling/.

[36] E. Zheleva and E. Terzi and L. Getoor: ”Privacy in Social Networks”, in *S.l.: Morgan & Claypool*, 2012.

[37] L.A. Cutillo and M. Manulis and T. Strufe: “Security and Privacy in Online Social Networks”, in *Handbook of Social Network Technologies and Applications*, pp. 497–522, 2010.

[38] C. Zhang and J. Sun and X. Zhu and Y. Fang: ”Privacy And Security for Online Social Networks: Challenges and Opportunities”, in *IEEE Network*, vol. 24, n. 4, pp. 13-18, 2010.

[39] U.U. Rehman and W.A. Khan and N.A. Saqib, and M. Kaleem: “On Detection and Prevention of Clickjacking Attack for OSNs” ,in *2013 11th International Conference on Frontiers of Information Technology*, pp. 160–165, 2013.

[40] C. Canali and M. Colajanni and R. Lancellotti: ”Data Acquisition in Social Networks: Issues And Proposals”, in *Proceedings of the International Workshop on Services and Open Sources (SOS’11)*, 2011.

[41] E. Chew and M. Swanson and K. Stine and N. Bartol and A. Brown and W. Robinson: “Performance Measurement Guide for Information Security”, in *National Institute of Standards and Technology (NIST)*, 2008. [Online].

[42] S.C. Payne: “SANS Institute – A Guide to Security Metrics - Research,” 2006.

[43] E.M. Maximilien and T. Grandison and T. Sun and D. Richardson and S. Guo, and K. Liu: ”Privacy-As-A-Service: Models, Algorithms, And Results On The Facebook Platform”, In *Proceedings of Web*, vol. 2, 2009.

[44] K. Liu and E. Terzi: “A Framework for Computing the Privacy Scores of Users in Online Social Networks”, in *2009 Ninth IEEE International Conference on Data Mining*, pp. 1–30, 2009.

[45] F.B. Baker and S.H Kim: İtem Response Theory: Parameter Estimation Techniques”, in *CRC Press*, 2004.

[46] A. Srivastava and G. Geethakumari: “Measuring Privacy Leaks In Online Social Networks”, in *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 2095–2100, 2013.

[47] G. Petkos and S. Papadopoulos and and Y. Kompatsiaris: “PScore: A Framework for Enhancing Privacy Awareness in Online Social Networks”, in *2015 10th International Conference on Availability, Reliability and Security*, pp. 592–600, 2015.

[48] R.G. Pensa and G.D. Blasi: “A Privacy Self-Assessment Framework for Online Social Networks”, in *Expert Systems with Applications*, vol. 86, pp. 18–31, 2017.

[49] J. Becker and H. Chen: ”Measuring Privacy Risk in Online Social Networks”, in *University of California, Davis*, 2009.

[50] T.H. Ngoc and I. Echizen and K. Komei and H. Yoshiura: “New Approach to Quantification of Privacy on Social Network Sites”, in *2010 24th IEEE International Conference on Advanced Information Networking and Applications*, pp. 556–564, 2010.

[51] N. Talukder and M. Ouzzani and A.K. Elmagarmid and H. Elmeleegy and M. Yakout: “Privometer: Privacy Protection in Social Networks”, in *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, pp. 266–269, 2010.

[52] C. Akcora and B. Carminati and and E. Ferrari: “Privacy in Social Networks: How Risky Is Your Social Graph?”, in *2012 IEEE 28th International Conference on Data Engineering*, pp. 9–19, 2012.

[53] B.S. Vidyalakshmi and R.K. Wong and C.H. Chi: ”Privacy Scoring of Social Network Users as a Service”, in *2015 IEEE International Conference on Services Computing, New York, NY*, pp. 218-225. 2015.

[54] R.K. Nepali and Y. Wang: ”SONET: A SOcial NETwork Model for Privacy Monitoring and Ranking”, in *2013 IEEE 33rd International Conference on Distributed Computing Systems Workshops, Philadelphia, PA*, pp. 162-166, 2013.

[55] Y. Wang and R.K. Nepali: “Privacy Measurement for Social Network Actor Model”, in *2013 International Conference on Social Computing*, pp. 659–664, 2013.

[56] Y. Wang and R.K. Nepali and J. Nikolai: “Social Network Privacy Measurement and Simulation”, in *2014 International Conference on Computing, Networking and Communications (ICNC)*, pp. 802–806, 2014.

[57] B. Schneier: “A Taxonomy of Social Networking Data”, in *IEEE Security & Privacy*, vol. 8, n. 4, pp. 88-88, 2010.

[58] Wikipedia: ”PDCA”, in *https://en.wikipedia.org/wiki/PDCA*, 2018

[59] V.R. Basili: ”Software Modeling and Measurement: The Goal/Question/Metric Paradigm”, in *University of Maryland at College Park*, 1992.

[60] M. Garcia and N. Lefkovitz and S. Lightman: ”Privacy Risk Management for Federal Information Systems”, in *NIST Internal Report (NISTIR) 8062*, 2017.

Análisis del protocolo para la anonimización de trayectorias desde el lado del cliente

Cristina Romero-Tris* David Megías*
* Internet Interdisciplinary Institute (IN3)
Universitat Oberta de Catalunya (UOC)
CYBERCAT-Center for Cybersecurity Research of Catalonia
Email: {cromerotr,dmegias}@uoc.edu

***Resumen*—Cada vez es más común el uso de dispositivos capaces de registrar la ubicación geográfica del usuario, como por ejemplo los smartphones, o los navegadores GPS. Por otro lado, la posibilidad de generar datos históricos de trayectorias abre un gran abanico de oportunidades en diferentes campos, como la optimización de rutas de transporte público, o la personalización de ofertas comerciales según la ubicación. Sin embargo, puesto que las trayectorias contienen datos de movilidad personal, su uso incorrecto o publicación pueden suponer una amenaza para la privacidad de los usuarios. A partir de la trayectoria de un usuario se pueden llegar a deducir detalles sensibles como hábitos de comportamiento, creencias religiosas y preferencias sexuales. Por lo tanto, muchos usuarios pueden no estar dispuestos a compartir sus trayectorias sin un proceso previo de anonimización. Actualmente, existen en la literatura varias propuestas que abordan este problema. Estas soluciones se centran en anonimizar los datos antes de su publicación, es decir, cuando ya están almacenados en la base de datos del servidor. Sin embargo, este enfoque no da al usuario control alguno sobre la información que comparte. Por esta razón, proponemos anonimizar los datos en los dispositivos móviles de los usuarios, antes de enviarlos a la base de datos de una tercera entidad. Este trabajo amplia nuestro trabajo anterior, incluyendo la implementación de un protocolo para la anonimización de datos en el dispositivo del usuario, que diseñamos y publicamos previamente. En concreto, presentamos y analizamos los resultados de haber ejecutado este protocolo usando datos de trayectorias reales con varias configuraciones de parámetros diferentes.**

***Palabras clave*—criptografía (*cryptography*), privacidad (*privacy*), trayectorias (*trajectories*).**

I. Introducción

Las tecnologías basadas capaces de geolocalizar un dispositivo, como el sistema de posicionamiento global (GPS), la identificación por radiofrecuencia (RFID) o los servicios basados en la localización (LBS) se han extendido significativamente en los últimos años. La expansión de las redes móviles y la capacidad de los dispositivos portátiles han hecho que sea posible generar gran cantidad de datos de localización exacta. Estos datos se pueden ir añadiendo en orden cronológico para crear lo que llamamos una trayectoria.

La extracción y el análisis de estos conjuntos de datos de trayectoria pueden proporcionar un servicio valioso al usuario. Hay muchos ejemplos de campos en los que esta información puede ser útil: monitoreo del tráfico (e.g., detectando atascos), publicidad basada en la ubicación (e.g., envío de descuentos a clientes potenciales cercanos), seguridad nacional (e.g., detección de anomalías migratorias) o rastrear infecciones. Un ejemplo de la facilidad de la obtención de trayectorias es el Google Maps Timeline [1] (Figura 1). Con este servicio, es posible visualizar (e incluso exportar los datos para análisis) las localizaciones visitadas en cualquier momento por los usuarios. La aplicación poporciona datos avanzados como la latitud, la longitud y el nombre del sitio visitado (e.g., el nombre de un restaurante o de un hotel).

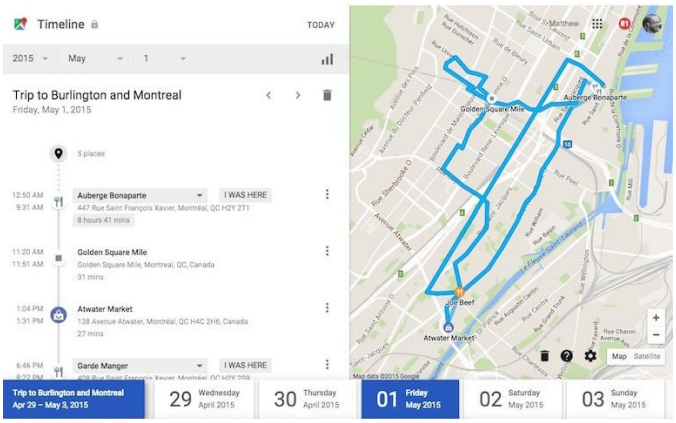


Figura 1. Trayectoria mostrada por Google Maps Timeline

Sin embargo, el problema es que a menudo las trayectorias contienen información que algunos usuarios podrían no querer compartir. Siguiendo con el mismo ejemplo, Google Maps Timeline permite generar un *ranking* de las localizaciones más visitadas. Lógicamente, en los primeros lugares de esta lista suelen aparecer las dirección de la vivienda y el lugar de trabajo del usuario, o en cualquier caso puede resultar evidente su clasificación como indican Golle y Partridge en [3]. A partir de ahí es fácil saber qué otros lugares ha visitado y deducir otros datos sobre comportamiento, modo de vida, creencias religiosas, preferencias sexuales, ideología, etc., que pueden amenazar la privacidad de los usuarios.

Una de las soluciones más extendidas en la literatura actual es anonimizar las trayectorias en la base de datos antes de

proceder a su análisis, a su publicación, o a ser enviadas a otra entidad [4], [5], [6]. Esto supone que los usuarios van a querer compartir su información real de trayectorias sin anonimizar. Sin embargo, esta suposición no siempre es correcta, y en muchas ocasiones los usuarios pueden preferir que los datos hayan sido anonimizados antes de ser enviados al servidor. Nuestra propuesta se basa en que el usuario participe activamente en el proceso de anonimización.

Por esta razón, diseñamos y presentamos el protocolo de anonimización de trayectorias publicado en [9].Este protocolo se ejecuta dentro del dispositivo del usuario y le permite saber qué partes o segmentos de su trayectoria son poco comunes, y eliminarlas. En concreto, las amenazas que intenta evitar se agrupan en tres categorías:

1. *La precisión de cada punto*. La ubicación en los dispositivos se hace por coordenadas (latitud y longitud), lo que hace que se identifique la localización unequívocamente. Para muchas aplicaciones, como la publicidad personalizada o el diseño de rutas de transporte, no es necesario revelar con tanta exactitud la ubicación, sino que es suficiente con conocer la zona en la que se encuentra el usuario. Sustituir la latitud y longitud por una zona más extensa es una forma de incrementar la privacidad del usuario. Siguiendo con la política de dar el máximo control al usuario sobre sus datos, nuestro sistema permite que sea el usuario el que decida el nivel de ofuscación de cada punto de su trayectoria.

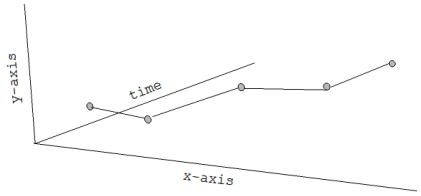


Figura 2. Ejemplo de trayectoria sin anonimizar

Para este fin, cada punto de la trayectoria se anonimiza transformándolo en un cilindro. Suponiendo un eje de tres dimensiones como el que se muestra en la Figura 2, el usuario elige un radio (e.g., 1 km) para la base del cilindro, formando un círculo que contiene la latitud y la longitud del punto original. Además, el usuario escoge un intervalo de tiempo, por ejemplo una hora, que formará la altura del cilindro y que marcará el intervalo de tiempo en el que el usuario se encontraba en ese punto. Este sistema permite al usuario decidir el grado de privacidad que desea para cada uno de los puntos de la trayectoria: escogiendo valores superiores para el radio y la altura del cilindro conseguirá mayor incertidumbre sobre su localización exacta. El resultado de un ejemplo de anonimización de la Figura 2 se muestra en la Figura 3.

2. *La singularidad de una trayectoria*. Un atacante con información previa sobre un usuario, puede volver a identificarlo en una base de datos si encuentra un registro único. Es decir, si detecta un único registro en la base de datos que corresponde con un desplazamiento que conoce de su víctima, puede identificarla unequívocamente.
Para solucionar este problema, nuestro sistema aplica un modelo de anonimización dentro de la base de datos. Para evitar que un atacante reidentifique a un usuario por la singularidad de una trayectoria en la base de datos, usamos el modelo k-anonimidad. La idea es que si k usuarios comparten trayectoria o segmentos de trayectoria, la probabilidad de vincular correctamente a la víctima con su registro es de $1/k$, cumpliendo entonces la definición de k-anonimidad. Siguiendo este modelo, un usuario no enviará una trayectoria o un subconjunto de puntos en su trayectoria, a no ser que aparezca al menos k veces en la base de datos.
3. *La confianza en el servidor*. Para saber si una trayectoria aparece k veces en la base de datos, un usuario necesitará cierto grado de interacción con el servidor. Sin embargo, revelar demasiada información al servidor puede ser una amenaza para la privacidad de los usuarios. Por esta razón, nuestra propuesta incluye un protocolo criptográfico distribuido: un usuario no preguntará directamente al servidor si una trayectoria o subtrayectoria aparece k veces en la base de datos. En su lugar, un grupo de usuarios colaborará para enviar consultas al servidor en nombre de otros miembros del grupo. La motivación para que un usuario haga esto es que otro miembro del grupo enviará sus consultas en su lugar. Para garantizar la privacidad, todas las consultas del grupo estarán cifradas. Al final del protocolo, ni el servidor ni ningún miembro del grupo pueden asociar una trayectoria a su usuario original, pero cada usuario obtiene las respuestas del servidor para sus consultas.

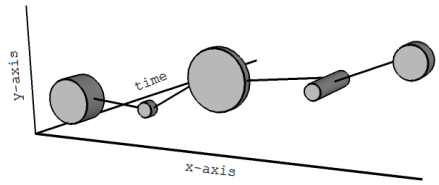


Figura 3. Ejemplo de trayectoria anonimizada

El resto del artículo se organiza de la siguiente manera. En la Sección II se detalla el estado del arte relacionado con el presente artículo. La Sección III describe de manera sencilla el protocolo de anonimización de trayectorias. La Sección IV contiene los resultados obtenidos para diferentes configuraciones de parámetros. Finalmente, la Sección V concluye el artículo y detalla las posibles líneas de investigación a seguir.

II. Estado del arte

En los últimos tiempos, el estudio de trayectorias ha atraído atención en la literatura de seguridad y privacidad en los sistemas de información. A continuación, describimos algunos de estos trabajos y la forma que tienen de anonimizar trayectorias.

Abul et al. [4] proponen un modelo llamado de (k, δ)-anonimidad, consistente en agregar k trayectorias dentro de un cilindro de radio δ, de manera que cada trayectoria queda identificada solo por ese cilindro y es indistinguible de las otras $k-1$ trayectorias.

Terrovitis et al. [5] proponen un algoritmo que elimina algunos puntos de la trayectoria. El aspecto más complicado en este caso es encontrar el conjunto óptimo de puntos a eliminar que minimice la pérdida de información relevante. De manera similar, Pensa et al. [7] proponen eliminar patrones que aparezcan en los datos con una cierta frecuencia. Esto se realiza añadiendo, borrando o sustituyendo algunos puntos de las trayectorias.

Yarovoy et al. [6] usan curvas Hilbert [8] para proyectar las localizaciones (de un espacio multidimensional) a una sola dimensión. El objetivo es encontrar "vecinos" en cada punto de la trayectoria y usarlos para crear grupos de anonimización y generalizar los datos de cada usuario.

La diferencia de todas estas propuestas con respecto a la nuestra, es que trabajan en el lado del servidor, y por lo tanto el usuario propietario de los datos no tiene ningún control sobre si esta anonimización se realizará, o si se hará de forma correcta. Nuestra propuesta es la primera de la literatura que permite que el proceso de anonimización se ejecute en el dispositivo del usuario.

III. Protocolo de anonimización de trayectorias

En esta sección, se proporciona una visión general del protocolo publicado en [9]. Como ya se ha mencionado, el objetivo del presente artículo es mostrar los resultados obtenidos al ejecutar este protocolo usando datos reales. Más información sobre cada uno de las fases, formulación y detalles criptográficos pueden encontrarse en artículo original [9].

El protocolo consta de varias fases:

1. *Creación de un grupo de anonimización* . En esta fase del protocolo, se crean grupos de manera aleatoria, formados por usuarios que quieren compartir sus datos de trayectorias manteniendo la privacidad.
2. *Anonimización de trayectoria*. Cada usuario de manera individual distorsiona los puntos de su trayectoria convirtiéndolos en cilindros similares a los de la Figura 3. Para cada punto que quiere anonimizar, elige un radio y un período de tiempo, creando así una zona espacial y temporal difusa en la que se encuentra el punto de la trayectoria original.
3. *Extracción de subtrayectorias*. En esta fase, se divide la trayectoria del usuario en varios segmentos más pequeños a los que llamamos subtrayectorias. El número de subtrayectorias a generar viene dado por un parámetro configurable del sistema.
4. *Generación de subtrayectorias falsas*. Para evitar que todas las subtrayectorias sean reales y aumentar el nivel de incertidumbre, el protocolo genera automáticamente una serie de subtrayectorias falsas.
5. *Distribución de subtrayectorias*. Las subtrayectorias reales, extraídas en el paso 3, y las subtrayectorias falsas, generadas en el paso 4, se distribuyen entre el grupo de usuarios creado en el paso 1. Aquí se ejecuta un protocolo criptográfico distribuido en el que las subtrayectorias se cifran, se mezclan aleatoriamente y se vuelven a descifrar. Al final de este paso, cada usuario posee un conjunto de subtrayectorias generadas por otros usuarios (pero desconoce a quién pertenece cada una).
6. *Envío de subtrayectorias y reenvío de respuestas*. Para cada subtrayectoria que el usuario posee, la envía en forma de consulta al servidor. El servidor responde de acuerdo con su base de datos, enviando el número de usuarios que ya tienen registrada esa subtrayectoria. Esta respuesta se envía a todos los usuarios, que tendrán en cuenta solo aquellas respuestas correspondientes a trayectorias que generaron ellos mismos, y descartarán las de los demás.
7. *Limpieza de la trayectoria anonimizada*. Las respuestas relacionadas con subtrayectorias falsas se ignoran automáticamente. A continuación, de acuerdo con la definición de k-anonimidad, cada usuario solo comparte una subtrayectoria si ésta aparece en la base de datos al menos k veces. En consecuencia, el usuario eliminará de su trayectoria todos aquellos segmentos de subtrayectorias que no cumplan con el criterio de k-anonimidad. El usuario envía el resultado al servidor para que lo almacene en su base de datos.

IV. Simulación y análisis de resultados

En esta sección presentamos los datos obtenidos después de haber implementado y ejecutado el protocolo que diseñamos. Para este fin, hemos utilizado datos reales extraídos de las trayectorias GPS de GeoLife [2]. Este conjunto de datos contiene secuencias de puntos espaciotemporales recogidos de 182 usuarios, empleando dispositivos GPS y smartphones, en un período de más de cinco años (de abril de 2007 a agosto de 2012). El conjunto de datos contiene 17.621 trayectorias con una distancia total de aproximadamente 1,2 millones de kilómetros y una duración total de más de 48,000 horas.

El protocolo propuesto requiere dos componentes: el servidor y la aplicación del usuario. Estos dos componentes se han implementado utilizando el lenguaje de programación Java para permitir la portabilidad de la aplicación. El servidor es un proceso que escucha peticiones en un puerto TCP fijo. A medida que va recibiendo solicitudes, va creando nuevos grupos y envía un mensaje a cada miembro con las direcciones IP y los puertos que se utilizarán.

IV-A. Configuración de los parámetros

El protocolo propuesto tiene varios parámetros que deben configurarse para las simulaciones:

- **El número de simulaciones**. Para evitar mostrar resultados más fiables, todas las simulaciones se han realizado tres veces para cada configuración de parámetros. Esto significa que todos los resultados presentados son la media aritmética de tres simulaciones.
- **El radio y el espacio de tiempo para anonimizar cada punto de trayectoria**. En un entorno real, los usuarios eligen estos parámetros para cada punto de sus trayectorias. Sin embargo, dado que nuestro sistema aún no se ha hecho público, para las simulaciones, nos vemos obligados a generarlos al azar. Más concretamente, para cada punto, generamos un valor aleatorio para el radio entre 0 y 1 km, y un valor aleatorio para el intervalo de tiempo entre 0 y 60 minutos.
- **El número de subtrayectorias para extraer y el número máximo de puntos en cada subtrayectoria**. Para ser completamente exhaustivos, hemos decidido extraer, para cada trayectoria, todas las subtrayectorias mínimas existentes. Esto significa que consideramos todas las combinaciones existentes ordenadas de dos puntos de puntos espaciotemporales. Por ejemplo, para una trayectoria $A \rightarrow B \rightarrow C \rightarrow D$, extraemos las subtrayectorias: $A \rightarrow B$, $A \rightarrow C$, $A \rightarrow D$, $B \rightarrow C$, $B \rightarrow D$, $C \rightarrow D$. Como resultado, de las 17,621 trayectorias en el conjunto de datos GeoLife, obtenemos 242,581 subtrayectorias asociadas a 182 usuarios diferentes.
- **El valor de k para aplicar k-anonimidad**. Con una muestra de solo 182 usuarios, hemos decidido realizar las simulaciones para valores de k pequeños, $k = 2$, $k = 3$, $k = 4$.

IV-B. Resultados

La Tabla I muestra los datos obtenidos de las simulaciones. Cada fila corresponde a un valor de k, es decir, cuántos usuarios deben la misma trayectoria en la base de datos para que el usuario actual decida compartirla. La primera columna indica el porcentaje de subtrayectorias compartidas y almacenadas en la base de datos respecto a todas las que se habían generado. La segunda columna muestra cuántos usuarios han enviado al menos una trayectoria. Por último, la tercera columna indica, para los casos en los que el usuario ha compartido al menos una trayectoria, qué porcentaje de sus trayectorias ha acabado compartiendo.

Tabla I
RESULTADOS DE k-ANONIMIDAD PARA $k = 2$, $k = 3$ Y $k = 4$

	% trayectorias	Al menos una	% trayectorias de al menos una
$k = 2$	7.12 %	148 de 182	10.84 %
$k = 3$	5.85 %	138 de 182	8.81 %
$k = 4$	3.94 %	136 de 182	6.23 %

Los resultados muestran una baja proporción de subtrayectorias anónimas enviadas. Esto se debe, en parte, a que la cantidad de datos disponibles es reducida, solo 182 usuarios diferentes. Podemos observar en la tabla que aunque hay una mayoría de usuarios que comparten al menos una trayectoria, el porcentaje de sus propias trayectorias que al final comparten sigue siendo bajo. Analizando estos datos, podemos deducir que los usuarios del conjunto de datos utilizados no tienen muchas ubicaciones en común.

Además, también se puede observar que cuanto mayor sea el valor de k, menos trayectorias se comparten. Reducir el valor de este parámetro disminuye el porcentaje de envíos de 7.12 %, a 5.85 % ($k = 3$) y a 3.94 % ($k = 4$). La cantidad de usuarios que envían al menos una subtrayectoria también se reduce, al igual que el porcentaje de subtrayectorias enviadas por éstos.

V. CONCLUSIÓN Y TRABAJO FUTURO

Los datos de trayectorias pueden ser útiles en diferentes aplicaciones, pero para poder usarlos sin amenazar la privacidad de los usuarios, es necesario anonimizarlos previamente. Las soluciones existentes en la literatura se centran en el desarrollo de técnicas que se ejecutan en el lado del servidor, una vez que todos los datos han sido recogidos. Sin embargo, en este trabajo permitimos que sean los usuarios los que controlen el nivel de anonimización de sus datos antes de compartirlos con el servidor. En este artículo, hemos presentado los resultados de implementar el protocolo presentado en [9]. Concretamente, hemos implementado el sistema y realizado simulaciones con datos reales de 182 usuarios usando la base de datos GeoLife [2]. Los resultados muestran la cantidad de datos compartidos para diferentes configuraciones de parámetros.

Los resultados indican que el número de trayectorias es reducido. Nuestro objetivo en un futuro es simular nuestro protocolo usando datos obtenidos directamente de redes sociales. La idea es usar las publicaciones que están geolocalizadas (e.g., tweets de Twitter [10]), para construir trayectorias y anonimizarlas con nuestro modelo. Además, tenemos la intención de implementar nuestro sistema en un entorno controlado con usuarios reales. La obtención de datos de usuarios reales permitirá calcular la utilidad y compararla con otras propuestas existentes.

AGRADECIMIENTOS

Este trabajo está financiado parcialmente por el Ministerio de Economía y Competitividad a través del proyecto TIN2014-57364-C2-2-R “SMARTGLACIS”.

REFERENCIAS

[1] Google Maps TimeLine, 2016. https://www.google.es/maps/timeline
[2] Y. Zheng, L, Zhang , X. Xie and Ma, Wei-Ying. “Mining interesting locations and travel sequences from GPS trajectories” *Proceedings of the 18th international conference on World wide web*, pp. 791–800, 2009
[3] P. Golle, and K. Partridge. “On the anonymity of home/work location pairs.” *7th International Conference on Pervasive Computing*, pp. 390-397, Nara, Japan, 2009.
[4] O. Abul, Osman, F. Bonchi, and M. Nanni, “Never walk alone: Uncertainty for anonymity in moving objects databases,” Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on. Ieee, 2008.
[5] M. Terrovitis, and N. Mamoulis, “Privacy preservation in the publication of trajectories,” Mobile Data Management, 2008. MDM’08. 9th International Conference on. IEEE, 2008.

[6] R. Yarovoy, F. Bonchi, L.V. Lakshmanan, W. Wang, “Anonymizing moving objects: How to hide a mob in a crowd?”, In Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (pp. 72-83). ACM, 2009.
[7] R. G. Pensa, A. Monreale, F. Pinelli, D. Pedreschi, “Pattern-preserving k-anonymization of sequences and its application to mobility data mining,” PiLBA’08 Privacy in Location-Based Applications, 44, 2008.
[8] D. Hilbert, “Ueber die stetige Abbildung einer Line auf ein Flächenstück”, Mathematische Annalen 38.3 (1891): 459-460.
[9] C. Romero-Tris, D. Megías, “User-centric Privacy-Preserving Collection and Analysis of Trajectory Data,” Lecture Notes in Computer Science. Pag. 245-253, 2016.
[10] Twitter, 2018. https://twitter.com/

Técnica Anti-Forense para la Anonimización de Vídeos de Dispositivos Móviles

Carlos Quinto Huaman, Ana Lucila Sandoval Orozco, Luis Javier García Villalba*

Grupo de Análisis, Seguridad y Sistemas (GASS),
Departamento de Ingeniería del Software e Inteligencia Artificial (DISIA)
Facultad de Informática, Despacho 431, Universidad Complutense de Madrid (UCM)
Calle Profesor José García Santesmases, 9, Ciudad Universitaria, 28040 Madrid, España
Emails: cquinto@ucm.es, {asandoval, javiergv}@fdi.ucm.es

***Resumen*—El vertiginoso aumento de las redes sociales y la cantidad de vídeos e imágenes digitales, principalmente generados por dispositivos móviles, que se comparten por este medio es evidente. Esta información puede ser objeto de manipulación intencionada para culpar o eximir a una persona. En este sentido, los vídeos digitales cobran gran importancia para la ciencia forense, que verifica la autenticidad de estos vídeos como evidencia en procesos judiciales y así contribuir en la toma de decisiones acertadas. Sin embargo, los atacantes responsables de dichas manipulaciones podrían aplicar técnicas anti-forenses para manipular vídeos sin dejar rastro alguno dificultando la labor del analista forense. En este trabajo se propone una novedosa técnica anti-forense que permite anonimizar un vídeo MP4. Esta técnica elimina toda la información del sensor de la cámara de cada uno de los fotogramas del vídeo. Todos los experimentos ejecutados para evaluar el algoritmo se han realizado con vídeos provenientes de distintos modelos de dispositivos móviles.**

***Index Terms*—Análisis Anti-Forense, Análisis Forense, Anonimización de Vídeos, Identificación de la Fuente, PRNU, Ruido del Sensor, Transformada Wavelet, Vídeos Digitales.**

I. Introducción

De acuerdo al informe de Cisco Systems [1], se proyecta que para el año 2020 existirá 5.500 millones de usuarios de dispositivos móviles, es decir un 70 % de la población mundial, sabiendo que la población estimada para el 2020 será 7.800 millones de personas según información de las Naciones Unidas. Asimismo, Cisco Systems estima que en 2020 los teléfonos inteligentes, computadoras portátiles y tabletas generarán el 98 % del tráfico mundial de datos móviles que representará 366,8 exabytes anuales frente a 89 % que representó 44,2 exabytes en 2015. La generación de vídeos móviles será el elemento con mayor demanda frente a otras aplicaciones, aumentando el uso de dispositivos con conexiones 4G o 5G. Por tanto, se estima que en 2020 el 75 % del tráfico de las redes móviles será por transferencia de vídeos de alta resolución. Esto significa que se generarán 80 mil millones de imágenes (28 imágenes diarias) y 7 mil millones de vídeos (más de 2,5 vídeos diarios) por habitante en el mundo durante un año.

Las cámaras digitales de dispositivos móviles son requeridos por los usuarios, dada la capacidad de capturar escenas y grabar vídeos digitales de excelente calidad. Sin embargo, esta necesidad no se limita a dichas bondades sino además a características cada vez más exigentes. Esto obliga a los fabricantes a agregar más ventajas a los dispositivos móviles como: menor peso, uso más intuitivo de las funcionalidades y la inserción de dos cámaras principales. Esta última característica conlleva a duplicar el sensor y las lentes del dispositivo. El sensor para conseguir escenas y fotografías más nítidas incluso en condiciones de baja luminosidad, debido a que un sensor monocromático se encargará de captar el nivel de luz y el otro sensor a color procesará los componentes de color. Las lentes proporcionarán un ángulo de visibilidad mucho más amplio, hacer zoom sin perder calidad y capturar imágenes con gran angular.

En consecuencia, los dispositivos móviles son usados para producir gran cantidad de fotografías y vídeos, que posteriormente pueden ser almacenados en otros dispositivos, publicados en redes sociales, enviados por correo electrónico, etc. Esto hace que, desde un punto de vista forense, las acciones delictivas como robo de información, pornografía infantil, secuestros, espionaje, etc., prolifere diariamente.

En este sentido, la detección de anonimización de vídeos es un procedimiento necesario para determinar responsabilidades y mostrar evidencias que posteriormente serán utilizados como pruebas legales. Sin embargo, esta técnica forense es atacada por otras denominadas técnicas anti-forenses responsables de alteraciones a las imágenes y vídeos, con el objeto de obstaculizar el proceso del análisis forense. En concreto, intentan ocultar, eliminar o falsificar alguna evidencia. Esto hace necesario desarrollar técnicas forenses capaces identificar estos métodos y hacer frente a los obstáculos y engaños digitales derivados de su uso [2].

El conocimiento adquirido de las técnicas anti-forenses que anonimizan la identidad del dispositivo que generó una imagen o vídeo digital, conllevan al desarrollo de técnicas forenses de identificación de la fuente más robustas, fiables y eficientes que permitan en todo momento contrarrestar este tipo de ataques anti-forenses.

En este trabajo se presenta una técnica anti-forense para realizar la anonimización de vídeos con formato MP4, centrándonos concretamente en los vídeos generados por dispositivos móviles.

Este trabajo está estructurado en cinco secciones, siendo la primera la presente introducción. En la sección 2 se presenta un estado del arte de las técnicas de análisis forense y anti-forense para vídeos generados por dispositivos móviles. En la sección 3 se presenta la técnica anti-forense propuesto. En la sección 4 se describe los experimentos realizados y los resultados obtenidos. Por último en la sección 5 se presentan las conclusiones del presente trabajo.

II. Estado del Arte

Las técnicas de análisis forense sobre vídeos de dispositivos móviles son fundamentales para detectar delitos flagrantes y ser evidencias de hechos potenciales en procesos judiciales. Las técnicas forenses en vídeos se agrupan en herramientas forenses según el objetivo a cumplir: herramientas forenses para el análisis de adquisición, compresión y manipulación de un vídeo [3].

Las herramientas forenses para el análisis de adquisición de la fuente, obtienen información de marca y modelo de diversos dispositivos. Comúnmente, estas investigaciones se han realizado sobre imágenes, siendo muy limitado sobre vídeos digitales. La mayoría de las técnicas que se pueden aplicar a una imagen se pueden emplear con los fotogramas de un vídeo. En [4] se realiza una comparación detallada de los principales grupos de técnicas de identificación de fuente de adquisición. Estas se dividen en cinco grupos y están basadas en: metadatos, características de la imagen, defectos de la matriz *Color Filter Array* (CFA) e interpolación cromática, imperfecciones del sensor y las transformadas wavelet.

Las herramientas forenses para el análisis de compresión utilizan arquitecturas de codificación del vídeo que son más complejas que las adoptadas para imágenes. El contenido del vídeo suele estar disponible en un formato de compresión con pérdidas debido a la velocidad de bits de gran tamaño y es necesario representar las imágenes en movimiento, ya sea en un formato no comprimido o sin pérdidas. La compresión con pérdida deja huellas características, que pueden ser detectados por el analista forense. En [5] se propone un método para la detección de doble compresión *Moving Picture Experts Group* (MPEG) basada en el bloqueo de los artefactos. Se define un sistema de medición para calcular el *Block Artifact Strength* (BAS) de cada trama. La media BAS se calcula para la eliminación de secuencias obtenidas a partir de 1 a 11 tramas, obteniendo un vector de características de los valores BAS. Si la secuencia se ha manipulado previamente, sea eliminación de fotograma o re-compresión, el vector de características presenta un comportamiento característico.

Las herramientas forenses para el análisis de manipulaciones hacen frente a todo tipo de manipulación de un vídeo como la sustitución y eliminación de fotogramas, replicación de fotogramas y la inserción o eliminación de objetos de la escena. Los dispositivos móviles suelen dejar una huella característica en los vídeos grabados. Estas huellas normalmente son explotadas sólo para la identificación de dispositivos [6], [7], [8].

Sin embargo, el atacante puede tener mayor conocimiento sobre las técnicas forenses, desarrollando técnicas anti-forenses sofisticadas para manipular vídeos sin dejar rastro alguno, es decir ocultar la huella del procedimiento realizado y de esta forma confundir al analista forense y conllevar a tomar decisiones judiciales no acertadas.

En este sentido, la ciencia forense debe hacer frente a las técnicas anti-forenses analizándolas rigurosamente para que al momento de implementar dichas técnicas forenses se puedan aplicar medidas de prevención capaces de detectar este tipo de operaciones anti-forenses.

La mayoría de las investigaciones realizadas sobre el análisis anti-forense se han enfocado en imágenes. Sin embargo, alguna de las técnicas se pueden utilizar en el análisis de vídeos a partir de la extracción de fotogramas. Las técnicas anti-forenses dirigidas a vídeos son relativamente escasas por diversos factores como la complejidad y diversidad de arquitecturas de codificación del vídeo.

En [9] se estudian técnicas anti-forenses aplicables a los procesos de edición de vídeos con codificación H.264/AVC. El trabajo se basa en el hecho de que la excesiva predicción residual en la codificación puede aparecer en los fotogramas que se han codificado con predicción intra e inter fotograma. En el primer método implementado los autores evalúan el resultado residual después del filtrado de desbloqueo para revelar las operaciones de edición realizadas. En el segundo, se utiliza un mecanismo de control de frecuencia para comprobar los parámetros de cuantificación.

En [10] se propone una técnica anti-forense para suprimir la huella digital temporal que surge en secuencias de vídeo MPEG cuando se agregan o eliminan fotogramas después de la re-compresión. Para ejecutar esta técnica inicialmente los autores identificaron las propiedades de la huella digital temporal. Estas propiedades fueron utilizadas para modelar el efecto de supresión y adición de fotogramas. La técnica funciona aumentando el error de predicción en ciertos P-frames del vídeo, mediante el establecimiento de los vectores de movimiento de algunos macrobloques dentro de ese fotograma a cero; para posteriormente calcular el error de predicción del fotograma. Después de realizar sus experimentos los autores consiguen eficientemente eliminar la huella digital temporal de los vídeos MPEG que se han sometido a una eliminación o adición de fotogramas.

En cuanto a imágenes fijas, en [11], [12] realizan una clasificación de los ataques a las técnicas de análisis forense de imágenes según el objetivo a cumplir: ocultamiento de post-procesamientos maliciosos sobre la imagen, destrucción de la identificación correcta del origen de la imagen y falsificación del origen de imagen.

En [13], [12] los autores implementan algoritmos para ocultar los procesos al que hayan sido sometidos las imágenes mediante el análisis de un conjunto de rasgos, que vienen a ser características únicas que dejan los dispositivos móviles al momento de la captura de la escena o de vídeo.

En [12] se presenta una técnica para ocultar el proceso de re-muestreo *resampling*. El re-muestreo es el redimensionamiento con interpolación de las imágenes. Para detectar el re-muestreo los algoritmos realizan una búsqueda de dependencias sistemáticas y periódicas entre píxeles vecinos. Para ocultar el re-muestreo se elimina las equidistancias periódicas insertando distorsiones geométricas. El ataque consiste en generar una imagen (y) a partir de una imagen (x) aplicando el re-muestreo, para esto se calcula el componente de baja frecuencia de (x), se calcula el componente de alta frecuencia y por último se obtiene un la imagen (y) que es la suma de los dos pasos anteriores. Los autores obtuvieron una tasa de falsos positivos *False Acceptance Rate* (FAR) inferior al 1 %.

En [14] los autores implementan una técnica para detectar un ataque de falsificación de la huella de una imagen y la inyección de ésta en otra imagen, mediante el análisis de las diferencias entre las propiedades de un patrón de ruido copiado.

III. Técnica anti-forense propuesta

En este trabajo se propone implementar una técnica anti-forense para realizar la anonimización de vídeos con formato MP4. La técnica está basada en la descomposición del vídeo, imperfecciones en el sensor y la transformada wavelet.

El algoritmo de anonimización propuesto tiene como objetivo eliminar toda la información del sensor de la cámara del dispositivo móvil presente en los fotogramas del vídeo para impedir que se identifique el dispositivo usado para grabar un vídeo determinado. El algoritmo se compone, a su vez, de cuatro procesos: La extracción del flujo elemental AAC (audio), la extracción de fotogramas, la eliminación del ruido de cada fotograma extraído del vídeo y la composición del vídeo. En la Figura 1 se muestra gráficamente el proceso general para efectuar la anonimización.

El Algoritmo 1 representa gráficamente la anonimización de un vídeo. El proceso se inicia con la lectura del vídeo para verificar si la estructura cumple con las especificaciones. Se obtienen los datos necesarios de los átomos contenedores *mdat*, *moov*, *trak* y de sus respectivos átomos hijos. Seguidamente se extrae el flujo elemental *Advanced Audio Coding* (AAC) ($Audio$) desde el vídeo original ($Video_{original}$) conservando los datos en bruto para evitar re-comprimir. Después, se realiza la extracción de los fotogramas objetivos ($Fotogramas_{objetivo}$) del vídeo original ($Video_{original}$). A continuación, se elimina la huella del ruido *Photo Response Non Uniformity* (PRNU) presente en cada fotograma extraído del vídeo ($Fotogramas_{objetivo}$).

Como resultado se obtiene un conjunto de fotogramas sin ruido ($Fotogramas_{sinruido}$). Este proceso se basa en el trabajo presentado en [15]. Finalmente, se realiza la re-composición del vídeo con los fotogramas sin ruido, dando como resultado un vídeo anonimizado ($Video_{anonimizado}$). Este proceso necesita dos elementos fundamentales: Por un lado el audio extraído ($Audio$), y por otro el conjunto de fotogramas sin ruido ($Fotogramas_{sinruido}$).

El proceso de extracción del flujo elemental AAC, se inicia realizando una lectura de todos los átomos del vídeo, centrandonos en el atomo *trak* de audio: *stbl* (datos), *stsd* (canales, frecuencia), *stsc* (número de muestras en un *chunk*), *stco* (ubicación de la muestra), *stsz* (tamaño de cada muestra en un *chunk*), *mdat* (muestras de audio). Luego, se obtiene el número de entradas del átomo *stsc*. Seguidamente se obtiene el número de *chunk* y el número de muestras que contiene cada *chunk*.

Algorithm 1: Anonimización de un vídeo MP4

Input: $Video_{original}$: Es el vídeo objetivo MP4
Result: $Video_{anonimizado}$:
Es el vídeo anonimizado MP4

```
1: procedure AnonimizarVideo(Video_original)
2:     Leer Video_original;
3:     Audio← ExtraerFlujoAcc(Video_original);
4:     Fotogramas_objetivo ← ExtraerFotogramas(Video_original);
5:     foreach fotograma in Fotogramas_objetivo do
6:         Fotogramas_sinruido ← EliminarPrnu(fotograma);
7:     Video_anonimizado=Audio+Fotogramas_sinruido;
8: end procedure
```

El siguiente paso es obtener el desplazamiento para cada muestra. Luego, se obtiene el tamaño de cada muestra del *chunk* para escribir la cabecera de cada muestra con los siguientes elementos: Tamaño de la muestra, frecuencia del audio y el canal. La ubicación de cada muestra se desplaza según la información del átomo *stsc* para obtener el bloque de datos del átomo *mdat* del vídeo con formato MP4 y se guardan los datos en un fichero con extensión AAC.

El proceso de extracción de fotogramas, se incicia realizando una lectura de todos los átomos del vídeo, centrandonos en el atomo *trak* de vídeo: *stbl* (datos), *stsd* (datos de registro de configuración del decodificador), *stsc* (número de muestras en un *chunk*), *stco* (ubicación de la *muestra*), *stsz* (tamaño de cada muestra de un *chunk*), *mdat* (muestras de vídeo). Seguidamente se obtiene el número de entradas del átomo *stsc*. Los siguientes pasos son similares a los realizados en el algoritmo de extracción de audio hasta obtener el tamaño de cada muestra del *chunk*. De cada una de estas muestras, se obtiene el bloque de datos y se verifica el campo de longitud (*nalu size*), que comúnmente está constituido por 4 bytes. A continuación se obtiene la primera parte de la cabecera de la unidad de acceso (*access unit* del fichero H.264. En esta primera cabecera se insertan los siguientes datos: Un prefijo de inicio de la unidad de acceso (*start of access unit*) representado por (00 00 00 01). La cabecera de la unidad *Network Abstraction Layer* (NAL) que contiene los campos *forbidden_zero_bit*, *nal_ref_idc* (0) y *nal*_unit_type (9) identificado como *access unit delimiter*. El tipo de porción o (*slice types*): (E0) que representa cualquiera de los siguientes tipos de *frames* (I, P, B, SI y SP). Por último, un delimitador (00 00 00 01). Esta primera cabecera se añade al principio de cada unidad de acceso.

Figura 1. Proceso de general de anonimización de Vídeos MP4

Seguidamente, se obtiene la segunda cabecera de la unidad de acceso (*access unit* del fichero H.264. Para realizar esto extrae la secuencia de datos del *decoderConfigurationRecord* y se agregan estos datos seguidos por el delimitador (00 00 00 01) al fichero .H264.

Finalmente, se realiza la escritura de los datos que representan la información del vídeo. Una vez extraído el flujo elemental H.264 del contenedor MP4, se realiza la extracción de todos los fotogramas del fichero .H264.

IV. Experimentos y Resultados

Para realizar los experimentos se utilizó el algoritmo de identificación de la fuente de adquisición basado en el agrupamiento [16]. En la Tabla I se muestra detalladamente los teléfonos móviles clasificados por marca y modelo utilizados en los siguientes experimentos.

En el experimento 1 se realizó una clasificación mediante técnicas de agrupamiento de los fotogramas sin anonimizar para verificar que los fotogramas se agruparan en las categorías a las que se conocía a priori. Para el presente experimento se ha utilizado un conjunto de vídeos pertenecientes a una marca y cinco modelos. De cada uno de los cinco modelos se utilizaron cuatro vídeos y de éstos se obtuvieron 50 fotogramas de resolución de 1280x720 píxeles, resultando un total de 200 fotogramas para cada modelo de la marca analizada. Los resultados de la clasificación se detallan en la Tabla II.

Como se puede apreciar en la Tabla II al realizar la clasificación se obtuvieron 5 grupos con una tasa de acierto promedio del 97 %. Los resultados de este experimento servirá más adelante para evaluar la efectividad de la anonimización de los vídeos.

En el experimento 2, para evaluar la efectividad de la anonimización de vídeos MP4, se anonimizó un vídeo de cada modelo de los móviles detallados en la Tabla I. Una vez anonimizados los vídeos se ejecutó el siguiente experimento. En este experimento se utilizó 3 vídeos sin anonimizar y un vídeo anonimizado, para un total de 4 vídeos por modelo. De los 3 vídeos sin anonimizar se extrajeron 50 fotogramas (150 fotogramas por modelo) y del vídeo anonimizado por modelo se extrajeron 50 fotogramas, para un total de 200 fotogramas. En este experimento se realizó la clasificación con una resolución de fotogramas de 640x480 píxeles. En la Tabla III se muestra como se agruparon los 200 fotogramas de los diferentes modelos.

Del modelo Galaxy A3, de 150 fotogramas sin alteraciones en la identidad, se clasificaron 110 correctamente obteniendo un 73 % de acierto y los 50 fotogramas inicialmente anonimizados se agruparon dentro del grupo (no clasificados) alcanzando un 100 % de efectividad en la anonimización. Del modelo Galaxy Ace Style, de los 150 fotogramas, se clasificaron correctamente 120 con un porcentaje de acierto del 80 % y los 50 fotogramas inicialmente anonimizados se agruparon dentro del grupo (no clasificados) alcanzando un 100 % de efectividad en la anonimización. El resto de los modelos se interpretan de la misma forma alcanzando una tasa de acierto promedio del 73.2 %. Se observa que después de la clasificación, las 3 nuevas clases: Grupo 1, Grupo 2 y Grupo 3, agrupa un total de 260 fotogramas, alcanzando un 100 % de efectividad global en la anonimización, por lo que se puede concluir que la técnica anti-forense de anonimización tuvo éxito ya que 250 fotogramas fueron anonimizados. Esto muestra que los vídeos anonimizados no fueron asociados con su modelo real, se debe tener en cuenta que la baja resolución también afecta la clasificación.

El experimento 3, se detalla en la Tabla IV, donde muestra el resultado de la clasificación con fotogramas de una resolución de 1280x720.

Tabla I
Teléfonos móviles clasificados por marca y modelo

Marca	Modelo
Samsung	Galaxy A3
	Galaxy Ace Style
	Galaxy S5 Neo
	Galaxy S6
	Galaxy GT-I9000

Tabla II
Matriz de confusión de vídeos no anonimizados con 1280x720 píxeles

	Dispositivo Móvil					Tasa de
Samsung Galaxy	**A3**	**Ace Style**	**S5 Neo**	**S6**	**GT I9000**	**acierto**
A3	**192**	8	0	0	0	96 %
Ace Style	4	**192**	0	0	4	96 %
S5 Neo	0	0	**200**	0	0	100 %
S6	0	0	0	**200**	0	100 %
GT I9000	6	4	0	4	**186**	93 %

Tabla III
Clasificación de vídeos anonimizados de 640×480 píxeles

	Dispositivo Móvil					No clasificados			
Samsung Galaxy	**A3**	**Ace Style**	**S5 Neo**	**S6**	**GT I9000**	**G. 1**	**G. 2**	**G. 3**	**% Acierto**
A3	**110**	8	23	4	0	29	6	20	73 %
Ace Style	5	**120**	17	2	6	28	12	10	80 %
S5 Neo	16	9	**118**	0	0	33	10	13	79 %
S6	3	15	0	**121**	12	8	38	3	81 %
GT I9000	10	27	10	23	**80**	10	3	37	53 %

Tabla IV
Clasificación de vídeos anonimizados de 1280×720 píxeles

	Dispositivo Móvil					N/A		%
Móvil Samsung Galaxy	**A3**	**Ace Style**	**S5 Neo**	**S6**	**GT I9000**	**G. 1**	**G. 2**	**T/A**
A3	**142**	8	3	1	0	20	26	95 %
Ace Style	5	**140**	0	0	4	36	15	93 %
S5 Neo	0	3	**150**	0	0	29	18	100 %
S6	7	0	1	**144**	3	42	3	96 %
GT I9000	5	6	1	6	**142**	17	23	95 %

Del modelo Galaxy A3, de 150 fotogramas sin alteraciones en la identidad, se clasificaron 142 correctamente obteniendo un 95 % de acierto y los 50 fotogramas inicialmente anonimizados se agruparon dentro del grupo (no clasificados) alcanzando un 92 % de efectividad en la anonimización. Del modelo Galaxy Ace Style, de 150 fotogramas sin alteraciones en la identidad, se clasificaron 140 correctamente obteniendo un 93 % de acierto y los 50 fotogramas inicialmente anonimizados se agruparon dentro del grupo (no clasificados) alcanzando un 100 % de efectividad en la anonimización. El resto de los modelos se interpretan de la misma forma alcanzando una tasa de acierto promedio en la clasificación del 95.8 %. Como se puede observar en la tabla IV, dos nuevas clases aparecieron en la clasificación: Grupo 1 y Grupo 2. Del total de los 250 fotogramas anonimizados 229 no se clasificaron como parte del grupo al que pertenecían originalmente alcanzando un 92 % de efectividad global en la anonimización. La tasa de acierto promedio de la clasificación con respecto al experimento 1 disminuyó solo un 1,2 %. Se aprecia en la Tabla III que en la clasificación aparecen tres grupos desconocidos, a diferencia de los dos grupos desconocidos que aparencen en este experimento. Esto se debe a que la clasificación es más precisa cuando la resolución es más alta. Entre menos grupos desconocidos detectados, la anonimización es aún más efectiva ya que elimina características que facilitan su clasificación.

V. Conclusiones

El presente trabajo propone una técnica anti-forense para realizar anonimización de vídeos con formato MP4. Esta técnica está compuestas por una serie de procesos que están basados en la descomposición del vídeo, el ruido del sensor y la transformada wavelet. El método propuesto tiene como objetivo la extracción del audio y fotogramas del vídeo, la eliminación de la huella del sensor de cada fotograma y la composición del vídeo anonimizado. En el experimento 2, se alcanzó una tasa promedio de acierto de 73 %, clasificándose 3 nuevas clases con un total de 260 fotogramas que no se clasificaron como parte del grupo al que pertenecían originalmente, sabiendo que 250 fotogramas fueron anonimizados, representando un un 100 % de efectividad global en la anonimización. En el experimento 3, se alcanzó una tasa de acierto promedio de 95.8 %, clasificándose 2 nuevas clases con un total de 229 fotogramas que no se clasificaron como parte del grupo al que pertenecían originalmente, sabiendo que 250 fotogramas fueron anonimizados, representando un 92 % de efectividad global en la anonimización. La evaluación de la técnica anti-forense de anonimización se ha realizado con varios experimentos, alcanzando un porcentaje de acierto positivo, por lo que se puede concluir que la técnica anti-forense de anonimización tuvo éxito. La clasificación es más precisa cuando la resolución del vídeo es más alta. Entre menos grupos desconocidos detectados, la anonimización es aún más efectiva ya que elimina características que facilitan su clasificación.

Agradecimientos

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 700326. Website: http://ramses2020.eu

Referencias

[1] CISCO, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020," pp. 1–39, February 2016.

[2] M. Ramesh Chourasiya and A. P. Wadhe, "Video Anti Forensics - A Review," *IJRITCC*, vol. 3, no. 4, pp. 2110–2114, April 2015.

[3] P. Bestagini, M. Fontani, S. Milani, M. Barni, A. Piva, M. Tagliasacchi, and S. Tubaro, "An Overview on Video Forensics," in *APSIPA Transactions on Signal and Information Processing*, vol. 1. APSIPA, August 2012, pp. 1229–1233.

[4] A. Sandoval Orozco, D. Arenas González, J. Rosales Corripio, L. García Villalba, and J. Hernandez-Castro, "Techniques for Source Camera Identification," in *Proceedings of the 6th International Conference on Information Technology*, May 2013, pp. 1–9.

[5] W. Luo, M. Wu, and J. Huang, "MPEG Recompression Detection Based on Block Artifacts," in *Proc. SPIE*, February 2008, pp. 1–12.

[6] N. Mondaini, R. Caldelli, A. Piva, M. Barni, and V. Cappellini, "Detection of Malevolent Changes in Digital Video for Forensic Applications," in *Proceedings of SPIE*, February 2007, pp. 1–12.

[7] C. Hsu, T. Y. Hung, C. W. Lin, and C. Hsu, "Video Forgery Detection Using Correlation of Noise Residue," in *IEEE 10th Workshop on Multimedia Signal Processing*, October 2008, pp. 170–174.

[8] M. Kobayashi, O. Takahiro, and S. Yoichi, "Detecting Forgery From Static Scene Video Based on Inconsistency in Noise Level Functions," in *IEEE Transactions Information Forensics and Security*, December 2010, pp. 883–891.

[9] P. Su, P. L. Suei, M. Chang, and J. Lain, "Forensic and Anti-forensic Techniques for Video Shot Editing in H.264/AVC," *Journal of Visual Communication and Image Representation*, vol. 29, no. C, pp. 103–113, may 2015.

[10] M. C. Stamm and K. J. Ray Liu, "Anti-Forensics For Frame Deletion Addition In MPEG Video," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Maryland, USA, May 2011, pp. 1876–1879.

[11] J. Rosales Corripio, "Algoritmo de Identificación de Fuente en Imágenes Digitales de Dispositivos Móviles," Universidad Complutense de Madrid, Tesis de Máster, Julio 2013.

[12] T. Gloe, M. Kirchner, A. Winkler, and R. Bohme, "Can We Trust Digital Image Forensics?" in *Proceedings of the 15th International Conference on Multimedia*. ACM Press, September 2007, pp. 78–86.

[13] A. Popescu and H. Farid, "Exposing Digital Forgeries by Detecting Traces of Resampling," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 758–767, February 2005.

[14] M. Steinebach, H. Liu, P. Fan, and S. Katzenbeisser, "Cell Phone Camera Ballistics: Attacks and Countermeasures," in *Proceedings on Multimedia on Mobile Devices*. SPIE-International Society for Optical Engine, January 2010, pp. 75 420B–75 420B–9.

[15] A. L. Sandoval Orozco, L. J. García Villalba, D. M. Arenas González, J. Rosales Corripio, J. Hernandez-Castro, and G. S. J., "Smartphone Image Acquisition Forensics using Sensor Fingerprint," *IET Computer Vision*, vol. 9, no. 5, pp. 723–731, September 2015.

[16] L. J. García Villalba, A. L. Sandoval Orozco, and J. Rosales Corripio, "Smartphone Image Clustering," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1927–1940, March 2015.

Data Utility Evaluation Framework for Graph Anonymization

Jordi Casas-Roma
Internet Interdisciplinary Institute (IN3)
Universitat Oberta de Catalunya (UOC)
CYBERCAT – Center for Cybersecurity Research of Catalonia
Barcelona, Spain
jcasasr@uoc.edu

***Abstract*—Anonymization of graph-based data is a problem which has been widely studied over the last years and several anonymization methods have been developed. Information loss measures have been carried out to evaluate the noise introduced in the anonymized data. However, there is no consensus about how to evaluate perturbation and data utility in privacy-preserving and anonymization scenarios, where released datasets contain some noise to hinder re-identification processes. Thus, it is quite complex to compare different methods or algorithms in literature. In this paper we propose a framework to evaluate and compare anonymous datasets in a common way, providing an objective score to clearly compare methods and algorithms.**

***Index Terms*—Privacy-preserving, Anonymity, Evaluation framework, Data utility, Social networks, Graphs**

I. Introduction

Currently, data mining processes require large amounts of data, which often contain personal and private information of users and individuals. Although basic processes are performed on data anonymization, such as removing names or other key identifiers, remaining information can still be sensitive as well as useful for an attacker to re-identify users and individuals. To solve this problem, methods which introduce noise to the original data have been developed in order to hinder the subsequent processes of re-identification. However, the noise introduced by the anonymization processes may affect data by reducing its usefulness in subsequent processes of data mining. It is necessary to keep the main properties of data to ensure the data mining process is not altered by the anonymization process.

Anonymization processes should allow the analysis performed in the anonymized data to lead to results as equal as possible to the ones obtained when applying the same analysis to the original data. Nevertheless, data modification is contrary to data utility. The larger data modification, the less data utility. Thus, a good anonymization method hinders the re-identification process while causing minimal distortion to the data.

Owing to what we have mentioned in the previous paragraph, several measures have been designed to evaluate the goodness of the anonymization methods. Generic information loss measures evaluate to what extent the analysis on anonymized data differs from the original data. Each measure focuses on a particular property of the data and it is application-independent. We assume that if these metrics show little variation between original and anonymized data, then the subsequent data mining processes will also show little variation between original and anonymized data. Furthermore, we can also use measures specifically designed to quantify perturbation on real-world specific problems, such as community detection (i. e. clustering) or information flow.

However, there is no standard or common way to evaluate data utility or information loss. Usually, several authors use different generic information loss measures to quantify perturbation on anonymized data. Hence, it is very hard (or even impossible) to compare information loss and data utility among different methods and algorithms, since each work of literature uses specific (and usually different) metrics to evaluate the perturbation or noise in anonymous graphs.

In this paper we propose a common framework to evaluate data utility and information loss on privacy-preserving data publication processes. Specifically, we use generic information loss measures and clustering-specific ones on graph formatted data in order to provide a clear comparison among an original graph and several perturbed (i. e. anonymous) graphs.

A. Notation

Let $G = (V, E)$ be a simple graph, where V is the set of nodes and E the set of edges in G. We use $v_i \in V$ to denote node i and $(v_i, v_j) \in E$ to indicate an edge connecting nodes v_i and v_j. We define $n = |V|$ to denote the number of nodes and $m = |E|$ to denote the number of edges. We use $G = (V, E)$ and $\widetilde{G} = (\widetilde{V}, \widetilde{E})$ to indicate the original and the anonymized graphs, respectively.

B. Roadmap

This paper is organized as follows. We review the state of the art of anonymization in networks in Section II. Our framework is introduced in Section III. Then, we discuss metrics related to generic information loss measures in Section IV and a methodology to compare clustering-specific information loss measures in Section V. Finally, brief examples of our results are presented in Section VI, while Section VII concludes the research and points future work directions.

II. Related work

The two main objectives of an anonymization process are: (1) to preserve the privacy of users or individuals who appear in a dataset, hindering the re-identification processes, and (2) to preserve data utility on anonymized data, i.e., minimizing information loss.

Anonymization methods and graph assessment depend on the type of data they are intended to work with. In this paper, we will work with simple, undirected and unlabelled graphs.

A. Anonymization

We categorize anonymization methods on graph formatted data into three main categories [6]:

- Graph modification approaches: These methods anonymize a graph by modifying (adding and/or deleting) edges or nodes in a graph. There are two basic approaches: (1) The simplest way alters the graph structure by removing and adding edges randomly. It is called randomization or random-based approach. (2) Another way consists of edge addition and deletion to fulfill some desired constraints, i.e. anonymization methods do not modify edges at random, they modify edges to meet some desired constraints. For example, k-anonymity-based approaches modify graph structure (by adding and removing edges) in order to get the k-anonymity value for the graph.
- Generalization approaches (also known as clustering-based approaches): These methods cluster nodes and edges into groups. Then, they anonymize each group into a super-node to publish the aggregate information about structural properties of the nodes [12]. The details about individuals can be hidden properly, but the graph may be shrunk considerably after anonymization, which may not be desirable for analyzing local structures.
- Differentially private approaches: These methods refer to algorithms which guarantee that individuals are protected under the definition of differential privacy [8]. Differential privacy imposes a guarantee on the data release mechanism rather than on the data itself. The goal is to provide statistical information about the data while preserving the privacy of users.

We will focus on graph modification approaches, which preserve local structures and keep the details of the data for clustering processes.

Randomization methods are based on introducing random noise in the original data. For graphs, there are two main approaches: (a) Rand Add/Del that randomly adds and deletes edges from the original graph (this strategy keeps the number of edges) and (b) Rand Switch that exchanges edges between pairs of nodes (this strategy keeps the degree of all nodes and the number of edges). Naturally, edge randomization can also be considered an additive-noise perturbation. There are several works on graph randomization in literature, such as [11], [18], [19], [3].

Another widely adopted strategy for graph modification approaches consists of edge addition and deletion to meet desired constraints, usually to achieve a certain level of privacy. For instance, take the k-anonymity concept that was introduced by Sweeney [17] for the privacy preservation on relational data. It states that an attacker cannot distinguish among k different records although he managed to find a group of quasi-identifiers. Consequently, the attacker cannot re-identify an individual with a probability greater than $\frac{1}{k}$. The k-anonymity model can be applied using different concepts when dealing with networks rather than relational data like in our case. A widely used option is to consider the node degree as a quasi-identifier, which corresponds to k-degree anonymity [14]. It is based on modifying the network structure (by adding and removing edges) to ensure that all nodes satisfy this condition. There are several works in literature on constrained anonymization, for instance [21], [22], [12], [5], among many others.

B. Graph assessment

Several generic measures have been used to quantify the structure's properties in graph formatted data. Authors usually use these measures and compare the values obtained by the original and the anonymized data in order to quantify the noise introduced by the anonymization process. When we quantify the information loss as described above, we talk about generic information loss measure.

Hay et al. [11] utilized five structural properties from graph theory for quantifying network structure. For each node, the authors evaluate closeness centrality, betweenness centrality and path length distribution. For the graph as a whole, they evaluate the degree distribution and the diameter. The objective is to keep these five measures close to their original values, assuming that little distortion is involved in the anonymized data. Ying and Wu [18] and Ying et al. [19] used both real space and spectrum based characteristics to study how the graph is affected by randomization methods. The authors focused on four real space characteristics of the graph and on two important eigenvalues of the graph spectrum. The real space characteristics are: the harmonic mean of the shortest distance, the modularity, the transitivity, and the sub-graph centrality. Since graph spectrum has close relations with many graph characteristics and can provide global measures for some network properties, the authors also consider the following two spectral characteristics: the largest eigenvalue of the adjacency matrix and the second smallest eigenvalue of the Laplacian matrix. Alternatively, Zou et al. [22] defined a simple method for evaluating information loss on undirected and unlabelled graphs. The method is based on the difference between the original and the anonymized graph edges, $cost(G, \widetilde{G}) = (E \cup \widetilde{E}) - (E \cap \widetilde{E})$. Liu and Terzi [14] used clustering coefficient and average path length for the same purpose. Hay et al. [12] examined five properties commonly measured and reported on network data: degree, path length, clustering coefficient, network resilience and infectiousness.

III. Framework

It is important to emphasize that these generic information loss measures only evaluate structural and spectral changes between original and anonymized data. That is, these measures do not evaluate the data mining processes on anonymized data, and as such, they are general or application-independent. The analysis of specific and application-dependent quality measures is an open problem. We consider in this paper the case of an application in clustering.

Although it is possible to compare space and time complexity of these works using the Big $\mathcal{O}$ notation, it is not possible (or at least, not easy) to compare data utility and information loss among relevant works in literature, since they are evaluated using different methodologies and metrics. Our objective is to provide a common framework to quantify information loss on graph perturbation processes. To do so, we will use generic information loss (GIL) measures and

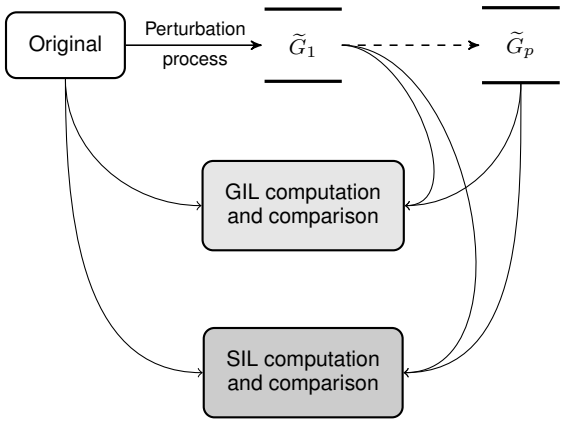


Fig. 1: Experimental framework. Each dataset is anonymized from $\widetilde{G}_1$ to $\widetilde{G}_p$ using some anonymization method. Next, we compare the original and perturbed data using GIL measures in order to quantify the noise introduced on the data. Then, we do the same with real clustering processes and SIL measures.

specific information loss (SIL) measures based on clustering processes.

As we have mentioned previously, in Section II, our framework is focused on graph modification approaches. All aforementioned methods share some properties that we use to evaluate the data utility in our evaluation framework. Specifically, the vertex set remains the same on perturbed graphs, i. e. $V = \widetilde{V}$ and $n = \widetilde{n}$. However, the edge set changes due to the randomization or constraint modification process, i. e. $E \neq \widetilde{E}$, while the number of edges is usually modified during the anonymization process, i. e. $m \neq \widetilde{m}$. The number of edges on perturbed graphs could be greater than the original one, i. e. $m < \widetilde{m}$, or smaller, i. e. $m > \widetilde{m}$. Mainly, it depends on algorithm's edge modification technique, which could be based on edge addition or edge removing.

Our experimental framework is shown in Figure 1. First, we apply perturbation to graph datasets using some graph modification approach. Each dataset is perturbed from $\widetilde{G}_1$ to $\widetilde{G}_p$ of edge set. Then, we evaluate original and perturbed data using GIL measures for quantifying network structure (details in Section IV). Next, we apply the clustering processes both on original and perturbed data and we use clustering-based specific measures (Section V) to evaluate the results.

IV. Generic Information Loss Measures

We use different generic measures for quantifying network structure. These generic measures are used to compare both the original and the anonymized data to quantify the noise introduced in the perturbed data by the anonymization process. These generic measures evaluate some key graph's properties, which are relevant according to [4]. They evaluate the graph structure, so they are general or, in other words, application-independent. Information loss was defined by the discrepancy between the results obtained from the original and the anonymized data. In our experiments we use several graph measures based on structural and spectral properties. In the rest of this section we review the measures used.

Average distance (AD) is defined as the average of the distances between each pair of nodes in the graph. It measures the minimum average number of edges between any pair of nodes. Formally, it is defined as:

$$AD(G) = \frac{\sum_{i,j} d_{ij}}{\binom{n}{2}} \tag{1}$$

where d_{ij} is the length of the shortest geodesic path from v_i to v_j, meaning the number of edges along the path.

Another used measure is **edge intersection** [22], [14] (EI). It is defined as the percentage of original edges which are also in the anonymized graph. Formally:

$$EI(G, \widetilde{G}) = \frac{|E \cap \widetilde{E}|}{max(|E|, |\widetilde{E}|)} \tag{2}$$

Clustering coefficient [14], [12], [10], [7] (C) is a measure widely used in literature. The clustering coefficient of a graph is the average:

$$C(G) = \frac{1}{n} \sum_{i=1}^{n} C(v_i) \tag{3}$$

where $C(v_i)$ is the clustering coefficient for node v_i. The clustering of each node is the fraction of possible triangles that exist. For each node the clustering coefficient is defined by:

$$C(v_i) = \frac{2T(v_i)}{deg(v_i)(deg(v_i) - 1)} \tag{4}$$

where $T(v_i)$ is the number of triangles surrounding node v_i, and $deg(v_i)$ is the degree of v_i.

Transitivity [18], [19], [7] (T) is the fraction of all possible triangles present in the graph. Possible triangles are identified by the number of triads (two edges with a shared node), as we can see in Equation 5.

$$T(G) = \frac{3 \times (number\, of\, triangles)}{(number\, of\, triads)} \tag{5}$$

Betweenness centrality [11] (BC) is a centrality measure, which calculates the fraction of number of the shortest paths that go through each node. This measure indicates the centrality of a node based on the flow among other nodes in the graph. A node with a high value indicates that this node is part of multiple shortest paths in the graph, which will be a key node in the graph structure. We define the betweenness centrality of a node v_i as:

$$BC(v_i) = \frac{1}{n^2} \sum_{s,t} \frac{g^i_{st}}{g_{st}} \tag{6}$$

where g^i_{st} is the number of geodesic paths from v_s to v_t that pass through v_i, and g_{st} is the total number of geodesic paths from v_s to v_t.

The second centrality measure is **closeness centrality** [11] (CC), which is described as the inverse of the average distance to all accessible nodes. Closeness is an inverse measure of centrality in which a larger value indicates a less central node, while a smaller value indicates a more central node. Formally, we define the closeness centrality of a node v_i as:

$$CC(v_i) = \frac{n}{\sum_j d_{ij}} \tag{7}$$

And the last centrality measure is **degree centrality** [11] (DC). It evaluates the centrality of each node associated with its degree. That is, the fraction of nodes connected to it. A higher value indicates greater centrality in the graph. The degree centrality of a node v_i is depicted in Equation 8.

$$DC(v_i) = \frac{deg(v_i)}{m} \tag{8}$$

The last three centrality measures described above evaluate the centrality of each node of the graph from different perspectives. These measures give us a value of centrality for each node. To assess the perturbation introduced in the graph by the anonymization process, we compute the vector of differences for each node between the original and the anonymized graph. Then, we compute the root mean square (*RMS*) to obtain a single value for the whole graph. We calculate the difference of the centrality measures between the original and the anonymized graph as follows:

$$\epsilon(G, \widetilde{G}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(g_i - \widetilde{g}_i)^2} \tag{9}$$

where g_i is the value of the centrality measure for the node v_i of G, and $\widetilde{g}_i$ is the value of the centrality measure for the node v_i of $\widetilde{G}$. In our experiments we use Equation 9 to compute a value representing the error induced in the whole graph by the anonymization process in the centrality measures.

We also focus on the **largest eigenvalue of the adjacency matrix A** (λ_1) [18] where λ_i are the eigenvalues of A and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. The eigenvalues of A encode information about the cycles of a graph as well as its diameter. The spectral decomposition of A is:

$$A = \sum_i \lambda_i e_i e_i^T \tag{10}$$

where e_i is the eigenvector corresponding to λ_i eigenvalue.

The number of nodes, edges and average degree are not considered parameters to assess anonymization process, since anonymization methods analysed in this work keep these values constant.

V. Specific Information Loss Measures

Variations in the generic graph properties are a good way to assess the information loss but they have their limitations because they are just a proxy to the changes in data utility we actually want to measure. For instance, the average distance or the diameter could remain constant while the topology of the network completely changes at the node level. What we are truly interested in is, given a data mining task at hand, quantify the disparity in the results between performing the task on the original network and on the anonymized one. We chose clustering because it is an active field of research, which provides interesting and useful information in community detection for instance. Therefore, the extracted clusters/communities of nodes are the data utility we want to preserve.

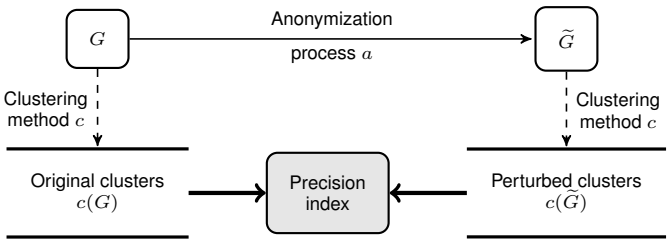


Fig. 2: Framework for evaluating the clustering-specific information loss measure.

A. Clustering

In this work we want to analyse the utility of the perturbed data by evaluating it on different clustering processes. Like generic graph measures, we compare the results obtained both by the original and the perturbed data in order to quantify the level of noise introduced in the perturbed data. This measure is specific and application-dependent, but it is necessary to test the perturbed data in real clustering processes.

We considered the following approach to evaluate the clustering assignment made by a given clustering method c using a particular graph perturbation method a: (1) apply a to the original data G and obtain $\widetilde{G} = a(G)$; (2) apply c to G and $\widetilde{G}$ to obtain the cluster assignments $c(G)$ and $c(\widetilde{G})$; and (3) compare $c(G)$ to $c(\widetilde{G})$, as illustrated in Figure 2. In terms of information loss, it is clear that the more similar $c(\widetilde{G})$ is to $c(G)$, the less information loss. Thus, clustering-specific information loss metrics should measure the divergence between both cluster assignments $c(G)$ and $c(\widetilde{G})$. Ideally, if the anonymization step was lossless in terms of data utility, we should have the same number of clusters with the same elements in each cluster. When the clusters do not match, we need to quantify the divergence.

For this purpose, we used the *precision index* [2]. Assuming we know the true communities of a graph, the precision index can be directly used to evaluate the similarity between two cluster assignments. Given a graph of n nodes and q true communities, we assigned to nodes the same labels $l_{tc}(\cdot)$ as the community they belong to. In our case, the true communities are the ones assigned to the original dataset (i.e. $c(G)$) since we want to obtain communities as close as the ones we would get on non-anonymized data. Assuming the perturbed graph has been divided into clusters (i.e. $c(\widetilde{G})$), then for every cluster, we examine all the nodes within it and assign to them as predicted label $l_{pc}(\cdot)$ the most frequent true label in that cluster (basically the mode). Then, the precision index can be defined as follows:

$$precision_index(G, \widetilde{G}) = \frac{1}{n}\sum_{v \in G} \mathbb{1}_{l_{tc}(v)=l_{pc}(v)} \tag{11}$$

where $\mathbb{1}$ is the indicator function such that $\mathbb{1}_{x=y}$ equals 1 if $x = y$ and 0 otherwise. Note that the precision index is a value in the range [0,1], which takes value 0 when there is no overlap between the sets and value 1 when the overlap between the sets is complete. To be consistent with the notion of error for the generic graph properties, we report 1 - $precision_index$ in the results tables so that the lower,

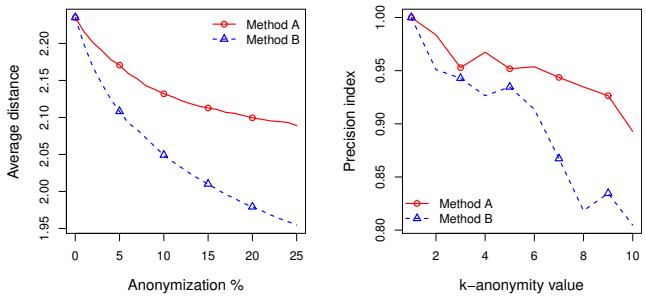


Fig. 3: Examples of our framework results. The horizontal axis presents the anonymization (randomization % or k-anonymity value), while vertical axis indicates the value of the original graph (leftmost point) and the evolution during anonymization processes.

the better.

Regarding the clustering methods c, we propose 4 graph clustering algorithms to evaluate the edge modifications techniques. All of them are unsupervised algorithms based on different concepts and developed for different applications and scopes. An extended revision and comparison of them can be found in Lancichinetti and Fortunato [13] and Zhang et al. [20]. The selected clustering algorithms are:

- *Fastgreedy* (FG) [9], a hierarchical agglomeration algorithm for detecting community structure based on modularity optimization.
- *Walktrap* (WT) [15] that tries to find densely connected sub-graphs, i.e. communities, in a graph via random walks.
- *Infomap* (IM) [16] that optimizes the map equation, which exploits the information-theoretic duality between the problem of compressing data and the problem of detecting significant structures in the graph.
- *Multilevel* (ML) [1], a multi-step technique based on a local optimization of Newman-Girvan modularity in the neighborhood of each node.

Even though some algorithms permit overlapping among different clusters, we did not allow it in our experiments by setting the corresponding parameter to zero, mainly for ease of evaluation.

VI. Application examples

In this section we briefly present some hypothetical results obtained by our experimental framework[1]. The framework expects two input graphs, G and $\widetilde{G}$, and it returns a score error for each GIL metric and a precision score for each selected clustering algorithm.

Usually, researchers are interested in comparing the original graph to a set of anonymous graphs obtained from the original one (i. e. $\widetilde{G}_1, \ldots, \widetilde{G}_p$), by applying different percentages of randomization or different k values in k-anonymity-based algorithms. Not only may this help to understand the behavior of the datasets, but also to choose the best parameter according to privacy and data utility requirements. An example is presented in Figure 3. For instance, the perturbation of average distance on anonymization percentage in range $[0, \ldots, 25\%]$ can be seen in Figure 3a. It is clear to see how this metric evolves during anonymization process. Therefore, it can help us to choose right algorithm's parameters to fulfill data utility and privacy constraints. A similar example is presented in Figure 3b, where the precision score clearly shows the behavior of two different methods over a set of $k \in [1, \ldots, 10]$.

[1]R Source code at: https://bitbucket.org/jcasasr/data-utility-framework/

VII. Conclusions

In recent years several anonymization algorithms have appeared to protect users' privacy. However, it is quite difficult to compare data utility among them, since each work usually uses different measures to compute and evaluate information loss. In this paper we have proposed a framework to evaluate data utility and information loss on privacy-preserving graph data. We claim that some generic information loss measures can be used to compute and evaluate information loss. Nevertheless, metrics related to application-specific real-world problems must be defined and used to compute and compare data utility among methods and algorithms in literature. Our framework provides a standard way to compute both metrics and can be easily used to perform comparisons among graph modification techniques (including random-based and constrained-based methods).

Many interesting directions for future research have been uncovered in this work. It would also be interesting to consider other specific information loss measures, such as those related to information flow or remaining ratio of top influential users. It would be also thought-provoking to extend this analysis to other graph's types (directed or labelled graphs, for instance).

Acknowledgements

This work was partly funded by the Spanish MCYT and the FEDER funds under grant TIN2014-57364-C2-2-R "SMART-GLACIS".

References

[1] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, Vol. 2008(10), P10008.

[2] Cai B-J, Wang H-Y, Zheng H-R and Wang H (2010) Evaluation repeated random walks in community detection of social networks. In: 2010 International Conference on Machine Learning and Cybernetics (ICMLC). IEEE Computer Society, Qingdao, pp 1849-1854

[3] Casas-Roma, J. (2014). Privacy-Preserving on Graphs Using Randomization and Edge-Relevance. In V. Torra (Ed.), International Conference on Modeling Decisions for Artificial Intelligence (MDAI) (pp. 204-216). Tokyo, Japan: Springer International Publishing Switzerland.

[4] Casas-Roma, J., Herrera-joancomartí, J., and Torra, V. (2015). Anonymizing Graphs: Measuring Quality for Clustering. Knowledge and Information Systems (KAIS), Vol. 44(3), pp. 507-528

[5] Casas-Roma, J., Herrera-joancomartí, J., and Torra, V. (2016). k-Degree Anonymity And Edge Selection: Improving Data Utility In Large Networks. Knowledge and Information Systems (KAIS), Vol. 50(2), pp. 447-474

[6] Casas-Roma, J., Herrera-Joancomartí, J., and Torra, V. (2017). A survey of graph-modification techniques for privacy-preserving on networks. Artificial Intelligence Review, 47(3), 341-366. http://doi.org/10.1007/s10462-016-9484-8

[7] Chakrabarti D and Faloutsos C (2006) Graph mining: Laws, generators, and algorithms. ACM Comput Surv 38(1):2:1-2:69

[8] Dwork C (2006) Differential Privacy. In: Proceedings of the 33rd International Conference on Automata, Languages and Programming (ICALP). Springer-Verlag, Berlin, pp 1-12

[9] Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. Physical Review E, Vol. 70(6), 66111.

[10] Girvan M and Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99(12):7821-7826

[11] Hay M, Miklau G, Jensen D, Weis P and Srivastava S (2007) Anonymizing Social Networks. Report, University of Massachusetts Amherst
[12] Hay M, Miklau G, Jensen D, Towsley D and Weis P (2008) Resisting structural re-identification in anonymized social networks. Proc VLDB Endow 1(1):102-114
[13] Lancichinetti, A., and Fortunato, S. (2009). Community detection algorithms: a comparative analysis. Physical Review E, Vol. 80(5), 56117.
[14] Liu K and Terzi E (2008) Towards identity anonymization on graphs. In: Proceedings of the ACM International Conference on Management of Data (SIGMOD). ACM Press, New York, pp 93-106
[15] Pons, P., and Latapy, M. (2005). Computing Communities in Large Networks Using Random Walks. In Computer and Information Sciences (ISCIS), Vol. 10, pp. 284293. Springer Berlin Heidelberg.
[16] Rosvall, M., and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences, Vol. 105(4), pp. 11181123.
[17] Sweeney L (2002) k-anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowl Based Syst 10(5):557-570.
[18] Ying X and Wu X (2008) Randomizing Social Networks: a Spectrum Preserving Approach. In: Proceedings of the SIAM International Conference on Data Mining (SDM). SIAM, Atlanta, pp 739-750
[19] Ying X, Pan K, Wu X and Guo L (2009) Comparisons of randomization and k-degree anonymization schemes for privacy preserving social network publishing. In: Proceedings of the 3rd Workshop on Social Network Mining and Analysis (SNA-KDD). ACM Press, New York, pp 10:1-10:10
[20] Zhang, K., Lo, D., Lim, E.-P., and Prasetyo, P. K. (2013). Mining indirect antagonistic communities from social interactions. Knowledge and Information Systems, Vol. 35(3), pp. 553583.
[21] Zhou B and Pei J (2008) Preserving Privacy in Social Networks Against Neighborhood Attacks. In: Proceedings of the 24th International Conference on Data Engineering (ICDE). IEEE Computer Society, Washington, pp 506-515
[22] Zou L, Chen L and Özsu MT (2009) K-Automorphism: A General Framework For Privacy Preserving Network Publication. Proc VLDB Endow 2(1):946-957

Anonimización de trayectorias por medio de intercambios

Julián Salas
Internet Interdisciplinary Institute (IN3)
Universitat Oberta de Catalunya (UOC)
Center for Cybersecurity Research of Catalonia (CYBERCAT)
Barcelona, Spain
jsalaspi@uoc.edu

David Megías
Internet Interdisciplinary Institute (IN3)
Universitat Oberta de Catalunya (UOC)
Center for Cybersecurity Research of Catalonia (CYBERCAT)
Barcelona, Spain
dmegias@uoc.edu

Vicenç Torra
School of Informatics
University of Skövde
Skövde, Sweden
vtorra@his.se

***Resumen*—El estudio de datos de movilidad puede ayudar a mejorar la toma de decisiones, así como planear los transportes en áreas metropolitanas o localizar los servicios en una ciudad. Sin embargo, si los datos geo-localizados son asociados a un individuo específico, pueden poner en en riesgo su privacidad, al permitir conocer todas sus localizaciones, sus horarios y revelar sus hábitos. Por estos motivos, antes de publicar datos de movilidad (trayectorias) es importante anonimizarlas, es decir, garantizar que no puedan ser asociadas a individuos específicos. En este trabajo, presentamos un método de anonimización perturbativa, que funciona de un modo similar al *rank swapping*, preserva la información agregada de los datos de movilidad, tales como el número de vehículos que viajan de un punto de interés a otro, durante un periodo específico de tiempo, y su vez, mantiene las identidades de los individuos y sus puntos de interés protegidos.**

Index Terms*—anonimización de trayectorias, privacidad in GIS/ datos espaciales, intercambio de trayectorias, *swapping

I. Introducción

Los servicios basados en la localización (*Location-Based Services*, en lo sucesivo, LBS) han sido posibles y van en aumento gracias a las técnicas de localización como son GPS, GSM o RFID. Pueden ser utilizados también en sistemas de transporte inteligente, por ejemplo, usar los vehículos como sensores para recoger información sobre la temperatura, las condiciones viales o los atascos de tráfico.

Por otra parte, los registros de localización comportan ciertos riesgos para la privacidad de los usuarios. Si su identidad real puede ser inferida en base a dichos datos, las preferencias, los hábitos e incluso los horarios de los individuos podrían ser revelados. Por este motivo, se han desarrollado diversos métodos de anonimización para proteger la identidad de los usuarios y sus ubicaciones exactas. El *survey* [1] revisa algunos ataques, garantías y transformaciones de los datos para proteger la privacidad en los LBS.

La privacidad en la minería de datos sociales tiene que ser protegida para promover la participación y garantizar los derechos de los sujetos de los datos. Algunos usos de la minería de datos sociales son: comprender el comportamiento humano a través de descubrir perfiles sociales individuales y de analizar el comportamiento colectivo, cf. [2]. Así como estudiar la propagación de epidemias, el contagio social y la evolución del sentimiento y la opinión.

La minería y el análisis de datos espacio-temporales es útil para publicidad, o para hacer políticas públicas, de salud o de movilidad. Al mismo tiempo, este tipo de datos puede revelar localizaciones sensibles, tales como la direcciones de las casas, los trabajos, o los sitios que los individuos visitan y no quieren hacer públicos porque esto revelaría sus preferencias privadas (como pueden ser políticas, religiosas o sexuales).

I-A. Literatura relacionada

Diversas soluciones han sido propuestas para anonimizar trayectorias antes de ser publicadas. Abul et al. [3], propusieron el modelo de (k, δ)-anonimato, que consiste en publicar un volumen cilíndrico de radio δ que contenga la trayectoria de un objeto en movimiento. Esta idea es una extensión del conocido concepto de k-anonimato para bases de datos [4].

Terrovitis y Mamoulis [5] consideran un dominio espacial discreto. En este caso, las trayectorias de los usuarios son expresadas como secuencias de POIs. Usan como ejemplo las tarjetas RFID de la compañía Octopus en Hong Kong[1], que recoge las historias de las transacciones de sus clientes, y está interesada en publicar las secuencias de transacciones que ha hecho una misma persona para extraer sus patrones de movimiento y de conducta. Sin embargo, si un usuario usa su tarjeta para pagar en distintas tiendas que pertenecen a la misma cadena, dicha compañía podría reidentificarlo si su secuencia de compras es única en la base de datos de trayectorias que ha sido publicada. En este caso la trayectoria se puede representar como una trayectoria en un grafo en el que los nodos representan todas las direcciones de las tiendas que aceptan tarjetas Octopus.

En [6] obtuvieron una aproximación similar que preserva los patrones secuenciales frecuentes (*frequent sequential patterns* [7]) por medio de transformaciones que agregan, eliminan o sustituyen algunos puntos de las trayectorias.

En [8] y [9], Hoh et al. se aborda el uso de datos de movilidad para planificación de transportes y para que aplicaciones que monitorizan el tránsito puedan proporcionar información sobre las condiciones del tránsito y del asfalto a los conductores. Para modelar las amenazas a la privacidad en dichos conjuntos de datos, asumen que un adversario no tiene información de cuales muestras pertenecen a un único usuario, aunque también asumen que usando algoritmos de

[1] http://www.octopuscards.com/

rastreo de objetos múltiples [10], las muestras posteriores se pueden vincular a un individuo que reporta periódicamente la información de su localización.

En [9] consideran el ataque de deducir las ubicaciones de las casas de los usuarios, por medio de heurísticas de clúster. Proponen técnicas de eliminación de datos como cambiar la frecuencia de muestreo (e.g., de 1 minuto a 2,4 y 10) para prevenir dichas inferencias. Mientras que en [8] proponen un algoritmo para perturbar las trayectorias de diversos individuos, haciéndolas parecer más cercanas para evitar que los atacantes descubran las trayectorias completas usando algoritmos de rastreo de objetos múltiples. Esto se lleva a cabo con una restricción de calidad, expresada como el error promedio entre la ubicación publicada y la ubicación real. Además, argumentan que se pueden garantizar niveles adecuados de privacidad si la densidad de usuarios es suficientemente alta.

Este trabajo está relacionado con las *mix-zones* que introdujeron en [11]. Las *mix-zones* son áreas en las que la localización de los usuarios no es accesible a las aplicaciones, de manera que cuando varios usuarios están presentes simultáneamente en una *mix-zone*, sus seudónimos pueden ser intercambiados para dificultar la asociación de una trayectoria que ha entrado a la zona con una que ha salido. El diseño que utilizan para proteger la privacidad de la ubicación intenta conservar las ventajas de los LBS y al mismo tiempo mantiene ocultas las identidades de los usuarios a las aplicaciones que reciben sus peticiones.

Para prevenir la reidentificación trivial, las comunicaciones entre los usuarios y las aplicaciones pasan por un intermediario de confianza y el seudónimo de un usuario cambia cuando cruza una *mix-zone*.

Para medir la privacidad en la ubicación, Beresford y Stajano [12] definen los conjuntos de anonimato como el conjunto de personas que visitan la *mix-zone* durante el mismo periodo de tiempo. Dado que la frontera y el tiempo en el que un usuario sale de una *mix-zone* está fuertemente correlacionado con la frontera y el tiempo en el que el usuario entra, esta información puede ser utilizada por un atacante, por este motivo, adaptan una métrica de teoría de la información que originalmente había sido propuesta por Serjantov y Danezis [13] para comunicaciones anónimas y considera las probabilidades de entrada y salida de mensajes a través de una red de *mix-nodes*.

Modelan estas probabilidades con una matriz de movimiento en la que se registran las frecuencias de entrada y salida de cada *mix-zone* en diversos momentos, y definen un grafo bipartito y con pesos, en el que los nodos representan los seudónimos de entrada y de salida y los pesos las probabilidades de que dos seudónimos representen a la misma persona. De esta manera un emparejamiento máximo en dicho grafo representa el mapeo más probable entre seudónimos de entrada y de salida.

Finalmente describen un método para llegar a soluciones parciales con la restricción de poder de computo, en un equilibrio entre la tratabilidad del problema y la precisión con la que se modela.

Otra aproximación que considera un sistema con un servidor no-confiable y clientes que se comunican en una red P2P para colectar trayectorias preservando la privacidad, fue propuesto en [14]. El objetivo es preservar el anonimato de los datos al guardarlos, transmitirlos o colectarlos. Esto se obtiene usando k-anonimato e intercambiando datos, sin embargo, aunque en cada fase los datos son anónimos, finalmente el servidor filtra los datos sintéticos y obtiene las trayectorias originales.

Una de las ventajas de que los usuarios sean quienes llevan a cabo la anonimización es que el proceso no es centralizado y los sujetos obtienen control, transparencia y más seguridad, como por ejemplo en [15].

II. Intercambio de trayectorias

En esta sección presentamos un método de intercambio de trayectorias para anonimizar datos de movilidad. Este método funciona de manera similar a las *mix-zones* pero en un espacio no delimitado, es decir, los usuarios que están cercanos pueden intercambiar sus trayectorias parciales sin necesidad de estar adentro de una *mix-zone*.

Aunque consideramos que el algoritmo se puede aplicar en tiempo real y P2P, en este trabajo definiremos una primera alternativa en la que los datos se anonimizan después de haber sido recogidos.

Al intercambiar sucesivamente las trayectorias parciales, estas se mezclan y cuando la trayectoria generada es recuperada por el servidor está formada por pequeñas fracciones de trayectorias de diversos individuos que se han encontrado durante el día, como en la Figura 1.

De este modo la relación entre los sujetos y sus datos se ofusca y al mismo tiempo se mantienen los datos agregados sin modificar, tales como el número de usuarios en un momento y una ubicación determinados.

En la siguiente sección presentamos definiciones y las suposiciones necesarias para formalizar nuestro método.

II-A. Definiciones

Suponemos que tenemos una tabla en la que la i-ésima observación es una tupla (ID_i, lat_i, long_i, t_i) que consiste en el identificador del individuo ID_i, la latitud (lat_i), la longitud (long_i) y el registro de tiempo (t_i).

La trayectoria T_x de un individuo x consistirá de todas las observaciones con identificador $ID_i = x$ ordenadas por sus registros de t_i. También se puede representar como $T_x = (x_1, x_2, \ldots, x_m)$ si hay m observaciones para el individuo x.

Diremos que dos individuos *se encuentran* o que sus trayectorias se cruzan en los puntos x_i y y_j, si han estado co-localizados, y lo denotamos con $x_i \approx y_j$. Estar *co-localizados* significa que han estado aproximadamente en el mismo lugar en el mismo momento, por lo tanto, estar co-localizados depende de los parámetros de proximidad χ y tiempo τ.

Definimos un *emparejamiento* como un subconjunto máximo de parejas de elementos de un conjunto

Denotamos por $Sw(T)$ a la trayectoria que resulta después de haber aplicado todos los intercambios a la trayectoria original T.

A continuación, definimos dos primitivas para nuestro algoritmo: *generar emparejamiento aleatorio* e *intercambiar*.

1. *Intercambiar (Swap):*
   Dadas dos trayectorias $T_x = (x_1, \ldots, x_i, x_{i+1}, \ldots)$ y $T_y = (y_1, \ldots, y_j, y_{j+1}, \ldots)$ que se encuentran en los puntos x_i y y_j, un swap de T_x

con T_y en los puntos x_i y y_j tiene como resultado $Sw(T_x) = (y_1, \ldots, y_j, x_{i+1}, \ldots)$ y $Sw(T_y) = (x_1, \ldots, x_i, y_{j+1}, \ldots)$.

2. *Generar emparejamiento aleatorio:* Dado un conjunto de elementos $S = s_1, s_2, \ldots, s_m$, generamos un emparejamiento aleatorio haciendo parejas entre los primeros $m/2$ y los siguientes $m/2$ números, después de haber hecho una permutación de todos los números m.

En caso de que el número de elementos m sea impar, para generar un emparejamiento es necesario dejar fuera un elemento al azar.

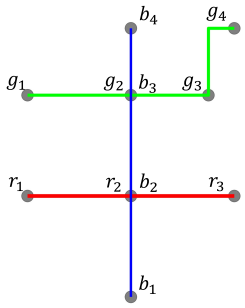
II-A1. Intercambio de trayectorias que se cruzan: Consideramos que las trayectorias de dos usuarios se cruzan si han estado colocalizados de acuerdo con los parámetros de proximidad χ y tiempo τ.

A continuación, simulamos el protocolo P2P en el que los usuarios intercambian sus IDs cuando están suficientemente cerca uno del otro.

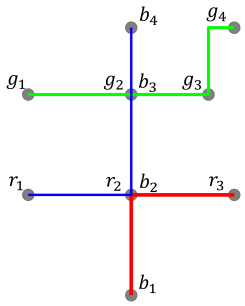
Calculamos los usuarios que están en contacto en un intervalo de tiempo determinado, y definimos un emparejamiento aleatorio entre ellos. El intercambio se lleva a cabo por parejas, sin embargo, podría haberse definido por una permutación como en [12].

Notemos que intercambiar seudónimos (IDs) es equivalente a intercambiar las trayectorias parciales.

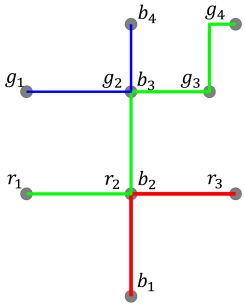
En la Figura 1, presentamos un sencillo ejemplo de tres trayectorias que se cruzan T_r, T_g, T_b. Suponiendo que se mueven de izquierda a derecha y hacia arriba, $T_r = (r_1, r_2, r_3)$, $T_g = (g_1, g_2, g_3, g_4)$ y $T_b = (b_1, b_2, b_3, b_4)$. También estamos suponiendo que la trayectoria azul se encuentra con la trayectoria roja primero ($b_2 \approx r_2$), y después con la verde ($b_3 \approx g_2$). En este pequeño ejemplo, podemos ver como los múltiples intercambios preservan partes de las trayectorias intactas, aunque al final cada trayectoria contiene partes de las otras con las que se ha cruzado, como por ejemplo la trayectoria verde que termina con un segmento de la azul, un segmento de la roja y un segmento de la trayectoria verde original, es decir, $Sw(T_g) = (r_1, r_2, b_3, g_3, g_4)$.



(a) Trayectorias originales



(b) Después del primer intercambio



(c) Después del segundo intercambio

Figura 1: Tres trayectorias antes y después del intercambio

II-B. Protección contra reidentificación

Es bien sabido que la de-identificación no necesariamente significa anonimización. Los mismos atributos que se utilizan para extraer conocimiento, pueden ser usados para encontrar a un individuo específico y distinguirlo únicamente, relacionando sus datos con su identidad.

En el contexto de datos geo-localizados, en [16], se mencionan los siguientes ataques a la privacidad de los individuos. Al identificar los POIs de un individuo, es posible inferir sus hábitos (e.g., hace deporte cada mañana, hace viajes largos en verano), las ubicaciones que visita frecuentemente y pueden estar relacionadas con sus creencias políticas o religiosas, o incluso con problemas de salud (como pueden ser clínicas y hospitales). También es posible deducir sus horarios, predecir sus ubicaciones y conocer sus ubicaciones anteriores, además de deducir sus relaciones personales observando con quien comparten frecuentemente sus ubicaciones. Asimismo, dichos hábitos y localizaciones pueden ser fácilmente utilizado para reidentificar a los individuos detrás de los datos, por ejemplo, los pares de ubicaciones casa/trabajo.

Golle y Partridge estudiaron en [17] individuos que revelaron las ubicaciones de sus casas y trabajos, con un ruido o una granularidad del tamaño de una manzana, un kilómetro o decenas de kilómetros (bloque censal, área censal o condado) y mostraron que los tamaños de los conjuntos de anonimato son respectivamente 1, 21 y 34980. Es decir, cuando la granularidad de los datos es del orden de un bloque censal, los individuos son únicamente identificables por sus ubicaciones de casa/trabajo, y para granularidades de área censal o condado, los individuos comparten el par de

Algorithm 1: algoritmo offline para intercambiar trayectorias

Input: Base de datos de trayectorias. Parámetros de tiempo τ y proximidad χ.
Output: identificadores de las trayectorias después de los swaps $Sw(T_i)$.
Partición de los registros de tiempo $t = \bigcup \tau_j$ en intervalos de longitud τ
for *Cada par de registros* i, j *en el intervalo* τ_j **do**
  if $dist(l_i, l_j) < \chi$ **then**
    agregar i, j a la lista de registros cercanos S_{τ_j} en el intervalo de tiempo τ_j
  end
end
generar un emparejamiento aleatorio con los registros en S_{τ_j}
ordenar todos los swaps en $\bigcup S_{\tau_j}$ por tiempo
for *para cada par* $i \approx j$ *en* $\bigcup S_{\tau_j}$ **do**
  intercambiar T_i con T_j
end
return *trayectorias intercambiadas* $Sw(T_i)$

ubicaciones casa/trabajo con 20 y 34979 individuos.

Zang and Bolot en [18], infieren las primeras N localizaciones de un usuario usando un registro de llamadas y las correlacionan con información accesible públicamente como datos del censo. Muestran que las primeras 2 localizaciones probablemente corresponden con la ubicación de casa y del trabajo, y posteriormente, que los conjuntos de anonimato se reducen drásticamente usando dichas localizaciones.

Por tanto, es crucial proteger las direcciones de casa y del trabajo de los usuarios para protegerlos contra la reidentificación y ofrecerles mínimas garantías de permanecer anónimos.

En nuestro caso, las trayectorias intercambiadas no permitirán seguir a un individuo específico y los sitios que visita, de manera que no permitirán ofrecerle información personalizada o clasificarlo, sin embargo, esto también es una forma de proteger su privacidad.

Nuestro algoritmo preserva el anonimato al disociar los segmentos de trayectorias del sujeto que los ha generado.

Aún en el caso que un atacante supiera algunos puntos que identifican a un individuo, no podría acceder a la trayectoria completa, puesto que la trayectoria publicada está formada por trayectorias de varios individuos

III. Evaluación

Hemos probado nuestro algoritmo en el conjunto T-drive de [19], [20], que contiene las trayectorias que han realizado 10,357 taxis del 2 al 8 de febrero de 2008 en Pekín. El número total de puntos es alrededor de 15 millones. Aunque, no todos los taxis aparecen cada día ni reportan sus posiciones con la misma regularidad. En promedio las reportan en un intervalo de alrededor de 177 segundos y 623 metros.

Asumimos que dos taxis han estado co-localizados si han estado a una distancia de menos de 111 metros ($\chi = 0{,}001$) en un periodo de 1-minuto ($\tau = 60$).

III-A. Reidentificación usando POIs y la ubicación de casa

En esta sección repasaremos métodos para inferir los POIs de un individuo, tales como la ubicación de su casa. Para probar que hemos protegido correctamente la ubicación de casa, compararemos la ubicación inferida antes y después de aplicar nuestro algoritmo.

Recordemos que uno de nuestros objetivos es proteger las ubicaciones relacionadas con los hábitos de los sujetos, y al mismo tiempo mantener los patrones agregados de movimiento.

Otro objetivo principal es proteger la asociación de las trayectorias a individuos específicos.

Como habíamos indicado anteriormente, si un atacante consiguiera asociar una trayectoria anonimizada con nuestro método a un individuo, no descubriría su trayectoria real, puesto que la trayectoria publicada estaría formada por segmentos de varias trayectorias. Por otra parte, aunque el atacante conozca algunos puntos reales y los intente utilizar para reidentificar a alguien, tendrá que encontrar una trayectoria anonimizada que también los contenga después del proceso de intercambio.

Para inferir los puntos de interés, en [21] consideran que, si un individuo permanece a menos de 200 metros de un punto determinado durante al menos 20 minutos, entonces es un punto de interés. En este caso, los parámetros para detectarlos serían distancia $\chi = 200$ metros y tiempo $\tau = 1200$.

En este trabajo, para obtener puntos de interés, discretizamos el espacio en celdas de 111 metros (0.001 grados decimales) y contamos las celdas más pobladas por cada individuo.

Definimos un punto de interés al centro una celda, si esta tiene al menos el 50 % de registros que la celda más poblada.

Analizando los tiempos, observamos que estos POIs también son POIs en el sentido de [21], aunque su definición podría considerar más puntos de los que hemos detectado.

Para inferir, la ubicación de la casa de un individuo, hemos seguido las ideas de [22] que discretizan el espacio en celdas de 25×25 km, y definen la ubicación de casa como la ubicación promedio de la celda más poblada. Ver también [23].

Igual que para los POIs, consideramos los puntos dentro de una celda de $t = 0{,}001$ grados de lado, es decir, aproximadamente 111 metros. Asignamos la ubicación de casa a la celda que tiene más puntos usando este parámetro.

Para validar que funciona nuestro método para deducir la ubicación de casa, generamos el histograma de las horas en las que cada individuo está cerca de la ubicación deducida de casa. Al analizar el histograma (Fig. 2), podemos observar que hay como mínimo alrededor de 5 % del total de taxis que están en (la ubicación deducida de) casa durante el día y este número incrementa consistentemente a partir de las 20 pm hasta las 6 am, que es cuando llega al máximo y decrece hasta alrededor de 5 % a las 10 am.

Este comportamiento es consistente con el comportamiento general de la población respecto a las horas que están en casa. Además, si asumimos que aquellas ubicaciones con más de 5 % de frecuencia relativa, son las ubicaciones correctas de casa, obtenemos que el porcentaje de ubicaciones correctas de casa es el 83,7 % del total, el cual es muy similar al que

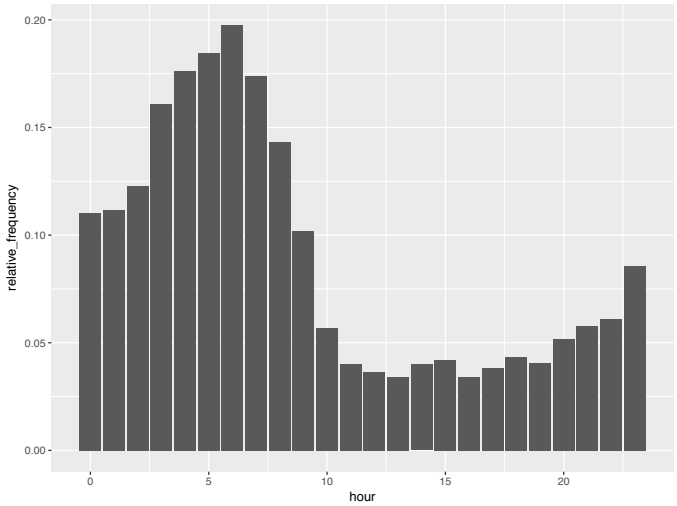


Figura 2: Histograma de las horas en que los taxis están en la localización deducida de sus casas, como porcentaje del total de mediciones para cada hora.

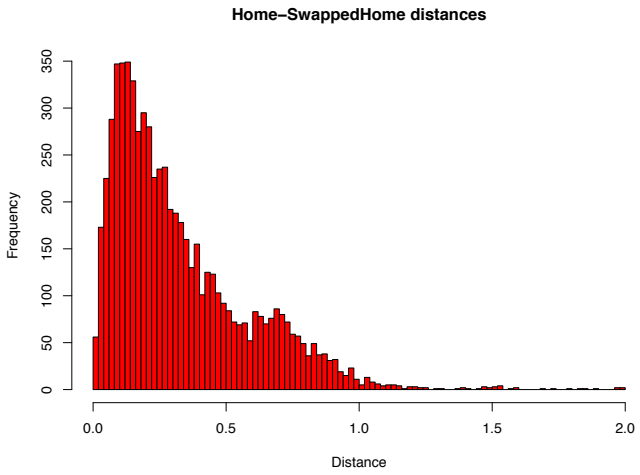


Figura 3: Histograma de distancias en grados decimales (i.e., $\times$111,32 km) entre la ubicación de casa deducida antes y después de aplicar nuestro algoritmo.

obtuvieron para su conjunto de datos en [22] por inspección manual (85 %).

Por ejemplo, usando este método deducimos que el usuario 7828 estuvo en casa desde las 00:13:44 a las 09:11:44 en la fecha 2008-02-03 (33 registros consecutivos). Posteriormente, desde las 23:32:01 en la fecha 2008-02-03 hasta las 10:11:59 en la fecha 2008-02-04 (70 registros consecutivos), llegó a casa a las 02:30:57 el día siguiente (55 registros) y salió a trabajar a las 10:10:56, y así sucesivamente. Así, podemos obtener no solo la ubicación de casa, sino también los hábitos o el horario de diversas personas.

III-B. Ubicación de casa después de los swaps

Para comprobar los resultados de nuestro método, intentamos deducir la ubicación de casa de los usuarios antes y después de aplicar nuestro algoritmo.

Considerando que la ubicación de casa es el POI más relevante y usando la técnica mencionada en la sección anterior para inferir POIs como los sitios más frecuentados, podemos concluir que nuestro método sirve para proteger no solamente la ubicación de casa, sino también los otros POIs.

En la Figura 3 mostramos el histograma de las distancias entre la ubicación de casa deducida antes y después de aplicar nuestro algoritmo.

Hemos obtenido que los usuarios para los cuales la distancia entre la ubicación de su casa deducida antes y después de aplicar nuestro algoritmo es menos de 0.001 grados son el 7674, 9197, 1, 533, 8205, 6590, 9223, 4801 y el 8427, las distancias respectivas en metros son 1.37, 10.24, 10.72, 17.48, 22.77, 26.98, 65.76, 74.8 y 106.34.

Únicamente los usuarios 1, 7674 y 9197 no han cambiado la ubicación de casa antes y después de intercambiar las trayectorias. Estos usuarios son un 0,02 % del total de usuarios.

El usuario 1 no se ha movido de casa durante el 2008-02-03 hasta las 18:35:14, por lo tanto, no había intercambiado su trayectoria a las 18:05, que es la hora que definimos para enviar las trayectorias intercambiadas al servidor. La misma situación se puede observar para los usuarios 7674 y 9197. Sin

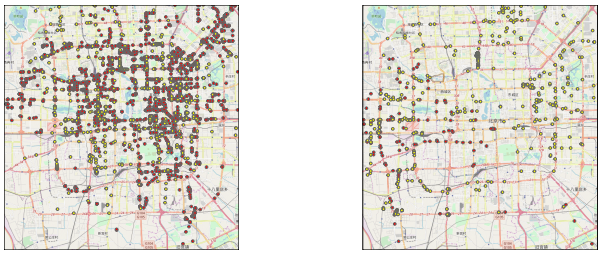


Figura 4: Puntos en las trayectorias originales y después de los swaps para los usuarios 7674 and 9197.

embargo, sus trayectorias han cambiado mucho después de los intercambios, como puede observarse en la Figura 4, en que la trayectoria original está representada en rojo y la modificada en amarillo. Por lo tanto, no hay una manera aparente en la que un atacante pudiera tener alguna certeza de que la ubicación de la casa que ha deducido sea correcta, además, hemos visto que acertaría con una probabilidad de a lo más 0,02 % en este ejemplo.

IV. Conclusiones

En este trabajo definimos un algoritmo para anonimizar trayectorias que no modifica los datos agregados de ubicación y tiempo. Preserva el número de individuos que hay en cada sitio, en cada momento, y modifica las trayectorias al intercambiar segmentos entre diversos usuarios. Para probar que las casas y los POIs son protegidos por nuestro algoritmo, utilizamos una heurística de clúster (basada en la frecuencia de visitas) para deducir las ubicaciones de las casas de los individuos en un conjunto con 15 millones de registros. Posteriormente, comprobamos que después de aplicar nuestro algoritmo de intercambio de trayectorias, no es posible deducir correctamente las ubicaciones de casa y otros POIs de los usuarios. Queda como trabajo futuro probar empíricamente otras medidas de utilidad y riesgo de los datos agregados.

Una posible herramienta para calcular la utilidad y el riesgo de reidentificación, será representar las trayectorias como un

"grafo de cruces", en el que los nodos representarán cuando se han cruzado dos o más individuos en una posición y un periodo de tiempo concretos y las aristas tendrán pesos que representen la cantidad de individuos que recorren el mismo trayecto entre dos cruces. Podemos adelantar que, usando esta herramienta, cualquier localización que no corresponda a un nodo en el grafo de cruces podría ser utilizada para identificar la trayectoria, puesto que será única. Además, sería interesante comparar la intersección de las trayectorias originales con las publicadas, para saber si aún en el caso de identificar una trayectoria, está podría no ser muy similar a la original según la distancia de Hamming o alguna otra métrica. El grafo de cruces que definiremos, también nos servirá para dar una aproximación al número de trayectorias que viajan de un sitio a otro.

Agradecimientos

Este trabajo está financiado parcialmente por el Ministerio de Economía y Competitividad a través del proyecto TIN2014-57364-C2-2-R "SMARTGLACIS". El tercer autor está financiado por el proyecto "Disclosure risk and transparency in big data privacy"(VR 2016-03346, 2017-2020). El primer autor agradece el apoyo de una beca postdoctoral UOC.

Referencias

[1] M. Terrovitis, "Privacy preservation in the dissemination of location data," *SIGKDD Explor. Newsl.*, vol. 13, no. 1, pp. 6–18, Aug. 2011. [Online]. Available: http://doi.acm.org/10.1145/2031331.2031334

[2] F. Giannotti, D. Pedreschi, A. Pentland, P. Lukowicz, D. Kossmann, J. Crowley, and D. Helbing, "A planetary nervous system for social mining and collective awareness," *The European Physical Journal Special Topics*, vol. 214, no. 1, pp. 49–75, Nov 2012. [Online]. Available: https://doi.org/10.1140/epjst/e2012-01688-9

[3] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," in *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, ser. ICDE '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 376–385. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2008.4497446

[4] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information (abstract)," in *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, ser. PODS '98. New York, NY, USA: ACM, 1998, pp. 188–. [Online]. Available: http://doi.acm.org/10.1145/275487.275508

[5] M. Terrovitis and N. Mamoulis, "Privacy preservation in the publication of trajectories," in *Proceedings of the The Ninth International Conference on Mobile Data Management*, ser. MDM '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 65–72. [Online]. Available: https://doi.org/10.1109/MDM.2008.29

[6] R. G. Pensa, A. Monreale, F. Pinelli, and D. Pedreschi, "Pattern-preserving k-anonymization of sequences and its application to mobility data mining," in *PiLBA*, 2008. [Online]. Available: https://air.unimi.it/retrieve/handle/2434/52786/106397/ProceedingsPiLBA08.pdf$\#$page=44

[7] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proceedings of the Eleventh International Conference on Data Engineering*, ser. ICDE '95. Washington, DC, USA: IEEE Computer Society, 1995, pp. 3–14. [Online]. Available: http://dl.acm.org/citation.cfm?id=645480.655281

[8] B. Hoh and M. Gruteser, "Protecting location privacy through path confusion," in *Proceedings of the First International Conference on Security and Privacy for Emerging Areas in Communications Networks*, ser. SECURECOMM '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 194–205. [Online]. Available: http://dx.doi.org/10.1109/SECURECOMM.2005.33

[9] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Enhancing security and privacy in traffic-monitoring systems," *IEEE Pervasive Computing*, vol. 5, no. 4, pp. 38–46, Oct 2006.

[10] D. B. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. 24, pp. 843–854, 1979.

[11] A. R. Beresford and F. Stajano, "Location privacy in pervasive computing," *IEEE Pervasive Computing*, vol. 2, no. 1, pp. 46–55, Jan. 2003. [Online]. Available: http://dx.doi.org/10.1109/MPRV.2003.1186725

[12] ——, "Mix zones: user privacy in location-aware services," in *IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second*, March 2004, pp. 127–131.

[13] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *Proceedings of the 2nd International Conference on Privacy Enhancing Technologies*, ser. PET'02. Berlin, Heidelberg: Springer-Verlag, 2003, pp. 41–53. [Online]. Available: http://dl.acm.org/citation.cfm?id=1765299.1765303

[14] G. Gidófalvi, "Spatio-temporal data mining for location-based services," Ph.D. dissertation, Faculties of Engineering, Science and Medicine Aalborg University, Denmark, 2007.

[15] C. Romero-Tris and D. Megías, "User-centric privacy-preserving collection and analysis of trajectory data," in *Data Privacy Management, and Security Assurance: 10th International Workshop, DPM 2015, and 4th International Workshop, QASA 2015, Vienna, Austria, September 21–22, 2015. Revised Selected Papers*, J. Garcia-Alfaro, G. Navarro-Arribas, A. Aldini, F. Martinelli, and N. Suri, Eds. Cham: Springer International Publishing, 2016, pp. 245–253. [Online]. Available: https://doi.org/10.1007/978-3-319-29883-2_17

[16] S. Gambs, M.-O. Killijian, and M. Núñez del Prado Cortez, "Show me how you move and i will tell you who you are," *Trans. Data Privacy*, vol. 4, no. 2, pp. 103–126, Aug. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2019316.2019320

[17] P. Golle and K. Partridge, "On the anonymity of home/work location pairs," in *Proceedings of the 7th International Conference on Pervasive Computing*, ser. Pervasive '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 390–397. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-01516-8_26

[18] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '11. New York, NY, USA: ACM, 2011, pp. 145–156. [Online]. Available: http://doi.acm.org/10.1145/2030613.2030630

[19] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: Driving directions based on taxi trajectories," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '10. New York, NY, USA: ACM, 2010, pp. 99–108. [Online]. Available: http://doi.acm.org/10.1145/1869790.1869807

[20] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 316–324. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020462

[21] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW '09. New York, NY, USA: ACM, 2009, pp. 791–800. [Online]. Available: http://doi.acm.org/10.1145/1526709.1526816

[22] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 1082–1090. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020579

[23] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks." *ICWSM*, vol. 11, pp. 329–336, 2011.

Validation of a SIR Epidemic Model for the Propagation of Jamming Attacks in Wireless Sensor Networks

Miguel López
Universidad de Málaga, E.T.S. Ingeniería de Telecomunicación, Dept. Ingeniería de Comunicaciones
Campus Teatinos 29071, Málaga, Spain
m.lopez@uma.es

Alberto Peinado
Universidad de Málaga, E.T.S. Ingeniería de Telecomunicación, Dept. Ingeniería de Comunicaciones
Campus Teatinos 29071, Málaga, Spain
apeinado@ic.uma.es

Andrés Ortíz
Universidad de Málaga, E.T.S. Ingeniería de Telecomunicación, Dept. Ingeniería de Comunicaciones
Campus Teatinos 29071, Málaga, Spain
aortiz@ic.uma.es

Abstract. This paper describes the utilization of an epidemic approach to study the propagation of jamming attacks, which can affect the physical and MAC/link layers of all nodes in a wireless sensor network, regardless of the complexity and computing power of the devices. We consider the deterministic epidemiological model *Susceptible–Infected–Recovered* (SIR), and analyse its accuracy using the results of public simulations and experiments based on protocols as MPH (*Multi-Parent Hierarchical Protocol*), AODV (*Ad-hoc On-demand Distance Vector*) and DSR (*Dynamic Source Routing*) over IEEE 802.15.4. The jamming term considers both the more classical approach of interfering signals, focusing on the physical level of the systems, and the cybersecurity approach that includes the attacks generated in upper layers, mainly in the MAC/link layer, producing the same effect on the communication channel.

***Index Terms*- Cyber security, jamming attacks, epidemiological models, wireless sensor networks.**

I. INTRODUCTION

The epidemiologic theory uses mathematical models to study and analyse the propagation of diseases. The current models comes from the proposed in 1927 by Kermack and McKendrick to describe epidemics in India [1]. The close relationship between the behaviour of biological and computer infections brought the researchers to apply these models in 1991 to predict computer virus propagation [2]. Since then, the models where constantly updated to describe malware propagation on different kind of networks, including wireless sensor networks (WSN) [3-5].

WSN consist of a large number of small nodes deployed to control or monitor critical infrastructure, industrial processes, environments and other applications based on the collection of data in real time. These networks are characterized by cooperation between nodes to create wireless communication paths. However, multicast nature of wireless technology, aggressive environments and other factors increase significantly the probability of executing various attacks against these networks [6]. On the other hand, the nodes in WSN are usually constrained devices in power supply, computation capability and memory. This fact can be considered a natural barrier against certain types of attacks, although some of them have proven their effectiveness such as those based on code injection or memory corruption vulnerabilities [7-8].

In any case, there are some attacks not based on malware that can be launched by corrupted nodes that can damage the network, downgrading its performance and functionalities, resulting in denial of service attacks (DoS). These nodes may interfere intentionally in the normal operation of the network wireless medium. At the physical or MAC level, they saturate the channel by injecting and continuously transmitting data packets, running down transmission and reception of data. This would cause in the affected nodes an extra effort in data communication reducing their energy, and therefore the overall lifetime of the network [9]. This type of DoS attack meets into the general category of jamming, (physical or link jamming) and could have devastating consequences even in the presence of a small number of attacker nodes [10].

Considering the physical layer of the IEEE 802.15.4 specification (the preferred wireless technology standard for real–time applications and generic WSN), it suffers from attacks generated at MAC layer, producing a DoS effect as it is reported in [11]. In addition, this standard is, at the same time, the least resistant against any type of jamming. The IETF has published in 2008 the IEEE 802.15.4e amended with multichannel scheme and new MAC procedures. However, most of devices currently in use is not updated [12].

In our previous work [21] we presented a first approach based on a *Susceptible–Exposed–Infected–Susceptible* (SEIS) model, in which individuals got the disease after an exposed period, and then the individuals became susceptible again, passing from one compartment to another continuously. In this paper, we present a different approach for the propagation of jamming attacks in WSN based on the deterministic epidemiological model *Susceptible–Infected–Recovered* (SIR), analysing whether the attack dies out or persists over time. In the SIR model, individuals who have no immunity get the disease, being infected if exposed, and after that infection period, they recover from the infection and become immune. To evaluate the model, we will use the data obtained in the experiment [22] and confirmed in [23], where the authors compare three routing protocol: MPH, AODV and DSR. For practical reasons, this model is intended to be applied over IEEE 802.15.4 standard, although –at this preliminary stage– is also applicable to any other wireless protocol, which go though from weaknesses at MAC level allowing to successfully perform a data-link jamming. It will therefore be of particular interest in the study, the analysis of the basic reproductive number R_0, which will represent the transition phase of non-equilibrium process of propagation of a disease. This parameter constitutes a key epidemic threshold in biological systems since if $R_0 < 1$ the infection dies out while if $R_0 > 1$ we may have a case of an epidemic disease.

The rest of the paper is organized as follows. In Section 2 introduces the background of the epidemiological theory in which the proposed model is based. Section 3 gives the analysis of the jamming attack under the proposed epidemiological approach. In section 4 we evaluate the model by comparing a case of study with our mathematical formulation; presenting the conclusions in Section 5.

II. FUNDAMENTALS OF EPIDEMIOLOGICAL THEORY

Modern epidemiological theory attempts to predict the propagation of a disease considering a particular population, by analyzing a threshold value to determine if the infection dies out in the population (*Disease-free Equilibrium*, DFE), or if it becomes epidemics (*Endemic Equilibrium*, EE) [13]. In this context, the basic reproductive number R_0 is a threshold that represents the average number of secondary infections that occur when the first infected individual (*patient zero*) is introduced into a population of N individuals fully susceptible [14]. Hence, if $R_0 < 1$ the infection dies out while if $R_0 > 1$ we may have a case of an epidemic disease. In this way, epidemiological theory is used to establish strategies and plans for prevention and control of diseases, such as vaccination or quarantine.

Most epidemiological models derive from the mathematical studies of Kermack–McKendrick [1], in which the population is divided into different compartments. Basics compartmental models are SIR (S: Susceptible, I: Infected, R: Recovered), SIS (S: Susceptible, I: Infected, S: Susceptible), SEIS (S: Susceptible, E: Exposed, I: Infected, S: Susceptible), etc. Diseases that confer immunity have a different model from diseases without immunity. For example, in SIR model susceptible individuals who have no immunity to the infectious agent, might become infected if exposed. If so, individuals who are currently infected can transmit the infection to susceptible individuals who they contact; and finally, recovered individuals are now immune to the infection, and do not participate in the transmission of the disease when they contact other individuals [14].

A. The SIR epidemic model

We will use the SIR epidemic model to describe a kind of disease that confers immunity against re-infection, hence the flow of individuals is from a compartment to another is, given by the flow chart:

$$S \xrightarrow{\beta} I \xrightarrow{\gamma} R$$

We are only interested in the study of the dynamics of a single epidemic outbreak, using three main hypotheses. Consider $S(t)$, $I(t)$ and $R(t)$ the number of individuals in Susceptible, Infected and Recovered compartments, respectively, at time t. Define β as the infection rate and γ as the period of recovery of infected individuals represented by $1/t_d$, where t_d is the expected duration of the disease. The first hypothesis is the *mass action incidence*, which assumes that an average member of the population is able to transmit the infection by making contacts with βN others member per time, where N is the population size. Since the probability that a random contact by an infective I is with a susceptible S (who can then transmit infection), the number of new infections (in unit time per infective) is $(\beta N)\cdot(S/N)$, giving a rate of new infections $(\beta N)\cdot(S/N)I=\beta S(t)I(t)$. The second hypothesis, determines that infected members of the population, leave the infective class at rate γI per unit time. From this assumption, we can write that the infected class should be $\beta S(t)I(t)-\gamma I(t)$ per unit time. Finally, the third hypothesis considers that there are no entries into or leaves from the population, except possibly through death from the disease, thus we have that $N(t)=S(t)+I(t)+R(t)$. This really meant, that the time scale of the spread of the disease is faster than the time scale of births and deaths so that demographic effects on the population may be ignored, and the population size is constant at any time [14].

According to the hypotheses, the algebraic expression of the model, can be written by the following differential equations system:

$$\frac{dS(t)}{dt} = -\beta S(t)I(t) \tag{1}$$

$$\frac{dI(t)}{dt} = \beta S(t)I(t)-\gamma I(t) \tag{2}$$

$$\frac{dR(t)}{dt} = -\gamma I(t) \tag{3}$$

This system of differential equations offers a view of the rate of change in the proportion of individuals in each group. The model is called autonomous invariant system since the infection rate β is assumed constant, *i.e.* does not change with time, so it depends on each disease; while the recovery rate γ is the inverse of the infectious period and depends only on the population of infected individuals who have at all times [14], when infectious diseases are considered.

In principle, we are unable to solve this system analytically but we can get a qualitative approach. To obtain the basic reproductive number, we seek for the solutions of the system by locating the steady states or equilibrium points, which are constant solutions that satisfy the conditions $dS(t)/dt=0$, $dI(t)/dt=0$ and $dR(t)/dt=0$. Considering that $S(t)+I(t)+R(t)=N$, we can find as possible solutions $I_0(t)=0$; or $I^*(t)>0$ with $S^*(t)=\gamma/\beta$. The rigorous mathematical analysis of local and global stability of these points implies the use of Lyapunov functions, or to apply linearization techniques, *Jacobian* operators, etc. [18, 19, 20], which is out of the scope of this paper. For our goal, we can assume the following theorem:

Theorem 1. Let $S(t)$, $I(t)$ and $R(t)$ be a set of solutions of the system of differential equations, then:

1) If $R_0 \leq 0$, then $\lim_{t\to\infty} I(t) = 0$ representing the *Disease-Free Equilibrium*, whereas
2) If $R_0 > 0$, then

$$\lim_{t\to\infty}[(S(t),I(t),R(t)] = \left(\frac{N}{R_0},\frac{\beta N}{\beta+\gamma}\left(1-\frac{1}{R_0}\right),\frac{\gamma N}{\beta+\gamma}\left(1-\frac{1}{R_0}\right)\right) \tag{4}$$

representing the *Endemic Equilibrium.*

It should be noted that if $R_0S(0)/N>1$, then there is an initial increase in the number of initial infected cases $I(t)$ addressing the epidemic state, but if $R_0S(0)/N\leq 1$, $I(t)$ decreases monotonically to zero, becoming a *Disease-Free Equilibrium*. The term $R_0S(0)/N$ is the *Initial Replacement Numb*er, and represents the average number of secondary infections produced by an infected individual during its infectiveness period at the outset of epidemics [19, 20]. In case of $R_0S(0)/N\leq 1$, the disease eventually disappears from

the population, but if the initial replacement number is greater than one, the population experiences an outbreak. Therefore, in the SIR epidemic model, the epidemic eventually ends.

Based on these definitions, we can find another parameter of interest about the dynamic of the disease which is the maximum number of infected. This is the number of infected when the derivative of I(t) is zero, that is, when $S = \gamma/\beta$. This maximum is given by:

$$I_{max} = S(0) + I(0) - \frac{\gamma}{\beta} lnS(0) - \frac{\gamma}{\beta} + \frac{\gamma}{\beta} ln\frac{\gamma}{\beta} \quad (5)$$

III. Epidemiological model for jamming propagation

Here we describe an epidemiological model to analyse the spread of jamming attacks from a single initial infected node, (the *zero patient*) which, in this case, will be the interfering node that generate the attack, and that further will produce the effects on the other nodes through the communication channel occupancy. Therefore, for modelling purposes, the effect produced by the jamming attack will be considered as the infectious disease, since it has the potential to affect progressively to other network nodes. In our model, the population under study is divided into compartments, taking assumptions about the nature and time rate to move from one compartment to another. The aim of this research is to study the influence of different infection rate β and recovery period γ, in the basic reproductive number of the system, that marks the difference between an epidemic process, which spread the attack throughout the network, or a process in which the attack dies out. The following sections describe the considered WSN model, the type of jamming to the study and the proposed epidemiological model.

A. Propagation model for Jamming attacks

Broadly speaking, jamming is a type of attack which aim is to interfere intentionally the normal operation of the wireless medium, at the physical and access level, saturating the channel by injecting and continuously transmitting data packets (with or without sense), causing abnormalities and errors in transmission and reception of data. However, knowledge of the communications protocol used in the WSN can deploy attacks from upper layers to achieve the same effect by sending valid frames, occupying the channel and thwarting the normal communication process. According to the taxonomy recently proposed by Lichtman et al. [15], such attacks falls into the category of cyber attacks usually related to the DoS, while other authors named it as link layer jamming [11].

As discussed before in both cases, physical and MAC layer, we can find several strategies to carry out such attacks with different levels of efficiency [10]. The simplest is the continuous emission of a signal to interfere a wireless channel (constant Jammer), so that legitimate traffic is completely blocked. However, there are techniques of signal modulation, which have proven resistant to this type of interference in the case of attacks at physical level. As for the flood attacks from the MAC level, while effective are easily detectable. Moreover, the strategy known as reactive jamming [17] has demonstrated to be more efficient in WSN environments. In this case, the jammer node is silent when there is no activity on the channel, and starts generating data packets as soon as a transmission is detected. Another strategy more efficient than the constant jamming is the random jamming. In this case, the attacker seeks to randomly inject signals or packets in the channel. Specifically, the attacker activates its radio interface during a time t_j and goes to sleep mode, resuming the attack again after a time t_s. Values for t_j and t_s can be random or fixed values.

This study focuses on the modeling of the random and reactive jamming originated at physical or MAC layer. Thus, it is considered the existence of a node into the network (patient zero) that is compromised to perform jamming attacks (acquires the disease) and then initiates it as described above. The effects of the attack (infection) will spread, first, through the nearest neighbor nodes, showing symptoms such as the increase in the number of packets forwarded between nodes, the loss of packets and extra consumption of resources due to the frames collisions, among others. These initially infected nodes will produce the same effect in their neighbors. It should be noted that in this model, nodes that overcome the stage of infection of the disease (the jamming attack) will be considered as recovered of the infection and, therefore, they do not participate in the transmission dynamics of the jamming effects in any way, when they contact other nodes.

B. Network model

In compartmental models, the patterns by which epidemics spread through groups of people is determined not only by the properties of the pathogen carrying it (contagiousness, the length of its infectious period, severity, etc.), but also by network structures within the population it is affecting. Assuming that the sizes of the compartments are large enough that the mixing of members is homogeneous, we consider a WSN consisting on N identical sensor nodes randomly distributed, following a uniform distribution in a region of area A, where $\rho = N/A$ is the average density. We also consider that all nodes are equipped with omnidirectional antenna which a maximum transmission range r_0. Therefore, the probability that k nodes are within the communication range of a particular node is given by:

$$p(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k} \quad (6)$$

Assuming a uniform distribution of nodes, we have that

$$p = \frac{\pi r_0^2 \rho}{N} = \frac{\pi r_0^2}{A} \quad (7)$$

where p is the probability of existence of a link at the physical level, that is, that at least two nodes are within their communication range, which can be assumed as a probability for the transmission of the disease.

C. Epidemiological analysis of the spread of jamming attack

The proposed model considers a stationary and uniformly distributed random population of N sensor nodes, with a density ρ. The process can be described as an epidemiological SIR model since thereby obtain a more accurate scheme of attack propagation, by including in the group of infected only those nodes that the attack effectively induces directly or indirectly malfunction that can spread throughout the network. Based on this premise, let us define $S(t)$, $I(t)$ and $R(t)$ as the number of individuals susceptible, infected and recovered at time t, where $S(t)+I(t)+R(t)=N$. Hence, taking as reference the homogeneous contact between individuals, and assuming an

infection rate given by the product $\lambda = p \cdot \beta$, the equations describing the model are:

$$\frac{dS(t)}{dt} = -\lambda S(t) I(t) \quad (8)$$

$$\frac{dI(t)}{dt} = \lambda S(t) I(t) - \gamma I(t) \quad (9)$$

$$\frac{dR(t)}{dt} = -\gamma I(t) \quad (10)$$

IV. Case Studies and Model Validation

To validate the proposed model, we will use the data obtained in the experiments [22, 23], where the authors compare the routing protocol MPH, with a couple of well-known protocols such as AODV and DSR. For this purpose, they considered a scenario with a single coordinator node that collects all of the information generated by the sensors. Hence, they tested the network without attacks and when it is subject to two different strategies of jamming attacks (random and reactive) considering several positions for the jammer. These experiments were based on a grid of 49 static nodes with 50 m of radio range, distributed uniformly in an area of 300 m x 300 m, with a coordinator node placed in the upper-left corner of the grid (functioning as a sink); and 48 nodes sending data with a rate of 1%–4% packets/s to the coordinator. The packet rate of the jammer node was set at 50 packets/s and 80 packets/s for random jamming attacks; while for reactive jamming is not necessary. The discovery neighbour time was stablished at 10 s, and the maximum data rate at 250 kbps, under a MAC layer with CSMA/CA protocol. For each experiment, the network performance was monitoring for 100 s in each case.

Regarding the experimental data, we are especially interested in the number of reached nodes by the coordinator. Hence, we have developed a simulation using MATLAB to represent the evolution of node's population over time using the data gathered by the experiment, and the evolution of node's population over time using the proposed SIR epidemic model. Note that the experiment's data is discrete, whereas the model is continuous. Also, we have extended the timeline to 200 s in order to simulate the steady state of the system by simply extending the data of the experiment repeating the values given by t = 100 s. Due to space limitations, we will only consider the example in which the jammer node is near of the coordinator. Tables I to III show the number of reached nodes in that scenario.

It should be noted that according to the proposed epidemic model the population of reached nodes at the steady state of the system is composed by the sum of susceptible nodes who have no immunity to the infectious agent, but eventually have not become infected (i.e. they have not been exposed to the jammer). Hence, recovered nodes, that overcome the stage of infection of the disease (the jamming attack) and they are able to receive and send packets to the coordinator in any way.

Figures 1 a 9 show the curves of the reached nodes, respect the curves of $S(t)$, $I(t)$ and $R(t)$. From the point of view of epidemic theory, the progress of the epidemic is managed by the contact network structure, the the infection rate β and the recovery period or expected duration of the disease γ (where $\gamma = 1/t_d$). It is observed that as time goes, the number of infected nodes $I(t)$ gradually increases reaching the maximum I_{max}, representing the top of the attack. After that time, the number of infected nodes $I(t)$ gradually decrease depending on the given parameters β and γ. For example, figures 1.a, 1.b and 1.c, show that spread of jamming attack at 50 packets/s from node to node is faster for AODV that than for DSR and MPH. Indeed, AODV and DSR have the same infection rate (β = 0.0290) but the expected duration of the disease t_d is about 15.5 sec. for AODV and 13.5 sec. for DSR, which yield a basic reproductive number R_0 = 1.877 for ADOV and R_0 = 1.634 DSR. Whereas for MPH, the infection rate β is 0.0255 and the expected duration of the disease is about 12.5 sec. thus the basic reproductive number is R_0 = 1.321. A similar situation is shown for the case of random jamming attack at 80 packets/s, and for reactive jamming regarding patterns of increasing, decreasing and prevalence for AODV, DSR and MPH.

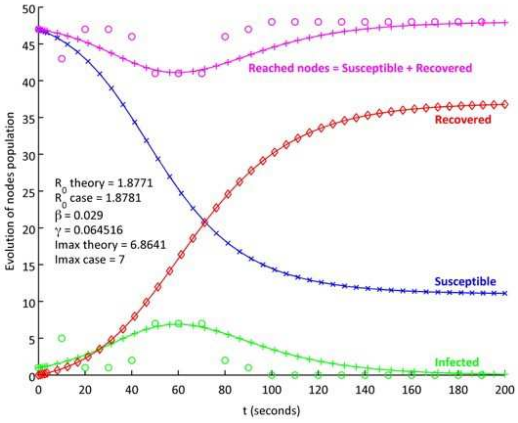


Fig. 1. AODV vs. random jamming at 50 packets/s.

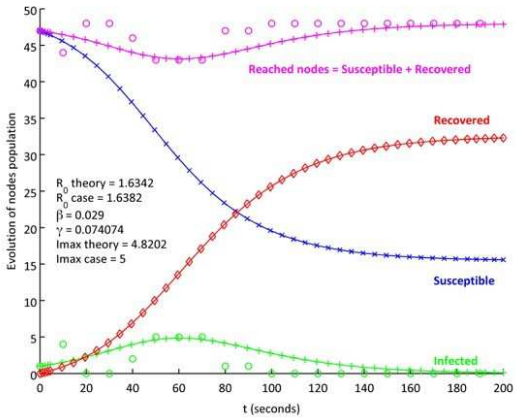


Fig. 2. DSR vs. random jamming at 50 packets/s.

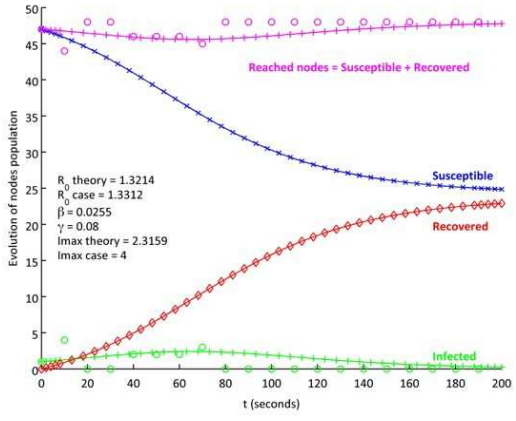


Fig. 3. MPH vs. random jamming at 50 packets/s.

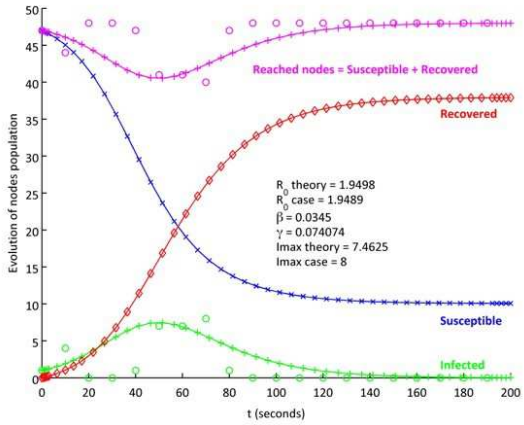


Fig. 4. AODV vs. random jamming at 80 packets/s.

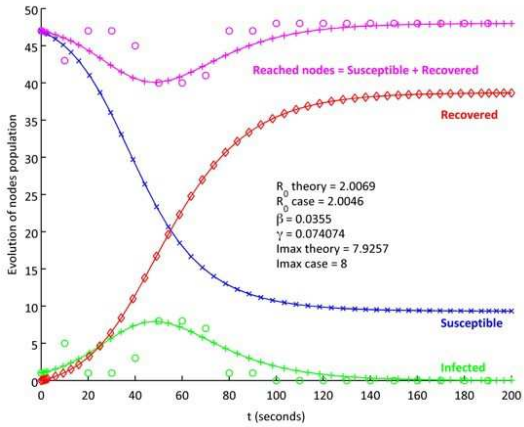


Fig. 5. DSR vs. random jamming at 80 packets/s.

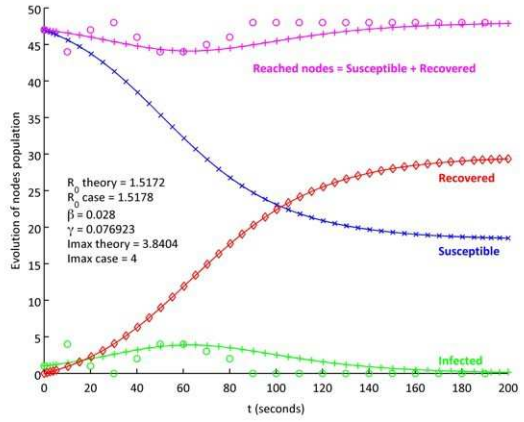


Fig. 6. MPH vs. random jamming at 80 packets/s.

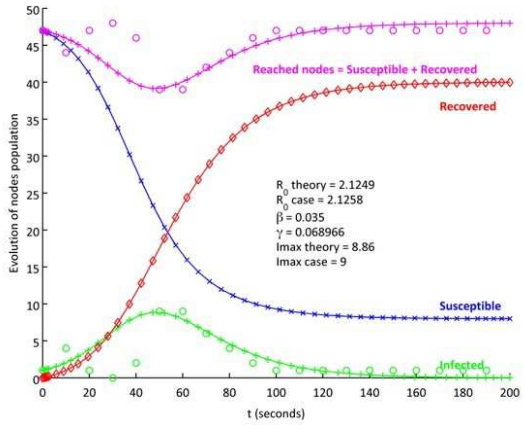


Fig. 7. AODV vs. reactive jamming.

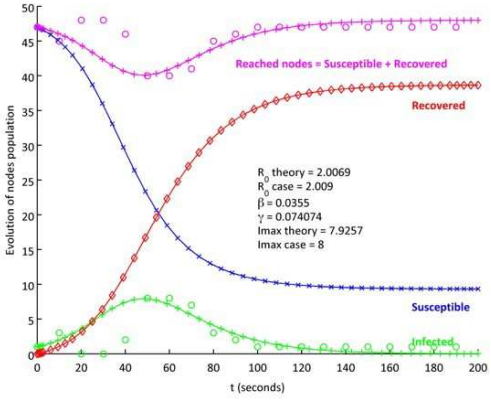


Fig. 8. DSR vs. reactive jamming.

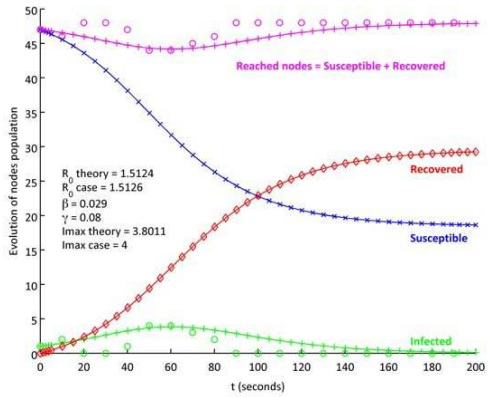


Fig. 9. MPH vs. reactive jamming.

V. Conclusion

This paper presents a first approach to apply the SIR epidemiological model to validate propagation of jamming attacks in physical and MAC/link layers in WSN. The values obtained provide a relation between the basic reproductive number R_0, and the severity of the attack. Indeed, the higher the value R_0, the more difficult it will contain the jamming attack. This analysis provides an important tool for predicting the attack effects and modelling possible countermeasures. For example, knowing the value R_0 associated with a particular type of jamming, we could determine a maximum transmission range r_0 to reduce the value λ, but guaranteeing the whole network connection. In addition, epidemiological models consider that to prevent the spread of a particular disease the population ratio of vaccination is $1–1/R_0$. This ratio could help us to develop a security mechanism that turn into sleep mode certain nodes, in order to mitigate the effects of the attack.

Acknowledgements

This work has been partially supported by Campus de Excelencia Internacional Andalucía Tech at Universidad de Málaga and by Ministerio de Economía, Industria y Competitividad (MINECO), Agencia Estatal de Investigación (AEI), and Fondo Europeo de Desarrollo Regional (FEDER, UE) under project COPCIS, reference TIN2017-84844-C2-1-R.

Table I
REACHED NODES WHEN THE JAMMER IS NEAR OF THE COORDINATOR: RANDOM JAMMING ATTACK

t (s)	Random jamming for 50 packets/s		
	AODV	DSR	MPH
10	43	44	44
20	47	48	48
30	47	48	48
40	46	46	46
50	41	43	46
60	41	43	46
70	41	43	45
80	46	47	48
90	47	47	48
100	48	48	48
110	48	48	48
120	48	48	48

Table II
REACHED NODES WHEN THE JAMMER IS NEAR OF THE COORDINATOR: RANDOM JAMMING ATTACK

t (s)	Random jamming for 80 packets/s		
	AODV	DSR	MPH
10	44	43	44
20	48	47	47
30	47	48	48
40	47	45	46
50	41	40	44
60	41	40	44
70	40	41	45
80	47	47	46
90	48	47	48
100	48	48	48
110	48	48	48
120	48	48	48

Table III
REACHED NODES WHEN THE JAMMER IS NEAR OF THE COORDINATOR: REACTIVE JAMMING ATTACK

t (s)	Reactive jamming		
	AODV	DSR	MPH
10	44	45	46
20	47	48	48
30	48	48	48
40	46	46	47
50	39	40	44
60	39	40	44
70	42	41	45
80	44	45	46
90	46	46	48
100	47	47	48
110	47	47	48
120	47	47	48
…	…	…	…

REFERENCES

[1] Kermack, W.O., McKendrick, A.G.: “Contributions to the mathematical theory of epidemics, part I”, in Proc. Roy Soc Edin A, vol. 115, pp. 700–721, 1927.

[2] Kephart, J.O., White, S.R.: “Directed-graph epidemiological models of computer viruses”, in Proc. of IEEE Symposium on Security and Privacy, pp. 343–359, 1991.

[3] Mishra, B.K., Keshri, N.: “Mathematical model on the transmission of worms in Wireless sensor network”, Applied Mathematical Modelling, vol. 37, pp 4103–4111, 2013.

[4] Xiaoming, W., Yingshu, L.: “An Improved SIR Model for Analyzing the Dynamics of Worm Propagation in Wireless Sensor Networks”, Chinese Jour. of Electronics, vol. 18, pp. 8–12, 2009.

[5] Martín del Rey, A.: “Mathematical modeling of the propagation of malware: a review”, Security and Communication Networks, vol. 8 (15), pp 2561–2579, 2015.

[6] Chowdhury, M., Kader, M.F., Asaduzzaman: “Security Issues in Wireless Sensor Networks: A Survey”, Int. Journal of Future Generation Communication and Networking, vol. 6, pp. 97–116, 2013.

[7] Habibi, J., Gupta, A., Carlsony, S., Panicker, A., Bertino, E.: Mavr: “Code reuse stealthy attacks and mitigation on unmanned aerial vehicles”, in Proc. IEEE 35th International Conference on Distributed Computing Systems, pp. 642–652, 2015.

[8] Braden, K., Crane, S., Davi, L., Franz, M., Larsen, P., Liebchen, C., Sadeghi, A.R.: “Leakage-resilient layout randomization for mobile devices”, in Proc. Network and Distributed Systems Security Symposium, 2016.

[9] Modares, H., Moravejosharieh, A., Salleh, R., Lloret, J.: “Security Overview of Wireless Sensor Network”, Life Science Journal, vol. 10, pp. 1627–1632, 2013.

[10] Znaidi, W., Minier, M., Babau, J.P.: “An Ontology for Attacks in Wireless Sensor Networks”, Unité de recherche INRIA Rhône-Alpes, Rapport de recherche Nº 6704, pp. 1–13, 2008.

[11] Sokullu, R., Korkmazy, I., Dagdevirenz, O., Mitsevax, A., Prasad, N.R.: “An investigation on IEEE 802. 15. 4 MAC layer attacks”, in Proc. of 10th International Symposium on Wireless Personal Multimedia Communications, 2007.

[12] Guglielmo, D., Brienza, S., Anastasi, G.: “IEEE 802.15.4e: A survey”, Computer Communications, 88, pp. 1-24, 2016.

[13] Hethecote, H.W.: “The Mathematics of Infectious Diseases”, SIAM Review, vol. 42, pp. 599–653, 2000.

[14] Brauer, F., van den Driessche, P., Wu, J.: “Mathematical Epidemiology”, Springer, 2008.

[15] Lichtman, M., Poston, J.D., Amuru, S., Shahriar, C., Clancy, T.C., Buehrer, R.M., Reed, J.H.: “A Communications Jamming Taxonomy”, IEEE Security & Privacy, vol. 14, pp. 47–54, 2016.

[16] Wei, Y., van Hoesel, L.L., Doumen, J., Hartel, P., Havinga, P.: “Energy Efficient Link Layer Jamming Attacks against Wireless Sensor Network MAC Protocols”, in Proc. SANS'05, 2005.

[17] Mohammadi, S., Jadidoleslamy, V: “A Comparison of Link Layer Attacks on Wireless Sensor Networks”, Int. journal on applications of graph theory in wireless ad hoc networks and sensor networks, vol. 3, pp. 35–56, 2011.

[18] Zhu, L., Zhao, H.: “Dynamical Analysis and Optimal control for a malware propagation model in an information network”, Neurocomputing, vol. 149, pp. 1370–1386, 2015.

[19] Hirsch, M.W., Smale, S., Devaney, R.L.: “Differential Equations, Dynamical Systems, and an Introduction to Chaos”, 3rd Ed., Academic Press, 2012.

[20] Perko, L.: “Differential Equations and Dynamical Systems”, 3rd Ed. Springer, 2010.

[21] López M., Peinado A., Ortiz A.: "A SEIS Model for Propagation of Random Jamming Attacks in Wireless Sensor Networks", International Joint Conference CISIS’16. Advances in Intelligent Systems and Computing, vol 527. pp. 668-677, Springer, 2017.

[22] Del-Valle-Soto, C., Mex-Perera, C., Monroy, R. and Nolazco-Flores, J. A.: “On the Routing Protocol Influence on the Resilience of Wireless Sensor Networks to Jamming Attacks”, Sensors, vol. 15, pp. 7619-7649, 2015.

[23] Del-Valle-Soto, C., Mex-Perera, C., Monroy, R., Nolazco-Flores, J. A.: "MPH-M, AODV-M and DSR-M Performance Evaluation under Jamming Attacks", Sensors 17, no. 7, pp. 1-26, 2017.

Ecosistemas IoT que preserven la privacidad

Juan Hernández-Serrano, Jose L. Muñoz
Universitat Politècnica de Catalunya

Arne Bröring
Siemens AG

Lars Mikkelsen
Aalborg Universitet

Wolfgang Schwarzott
Atos IT Solutions and Services GmbH

Oscar Esparza, Olga León, Miguel Soriano
Universitat Politècnica de Catalunya

Jan Zibuschka
Robert Bosch GmbH

***Abstract*—Hoy en día, ya están disponibles varias plataformas IoT (Internet of Things), dando acceso a miles de *cosas*. Iniciativas como *BIG IoT* son capaces de unir dichas plataformas IoT para formar ecosistemas. BIG IoT tiene como objetivo facilitar el desarrollo de aplicaciones entre diferentes plataformas, y establecer un marketplace centralizado. Dispondremos de una combinación de aplicaciones multiplataforma, con la posibilidad de gestionar de forma centralizada la contabilidad de los recursos. Sin embargo, esto también supone abordar desafíos cruciales para la seguridad y la privacidad. Este artículo analiza los requisitos de seguridad en los ecosistemas IoT y resume las soluciones a adoptar en el proyecto BIG IoT. También se presenta el análisis de seguridad y privacidad de un caso de uso de un asistente inteligente de transporte.**

***Index Terms*—Internet of Things, IoT, security, privacy**

I. Introduction

Hoy en día ya exiten multitud de plataformas IoT, y varias iniciativas para formar ecosistemas que permiten el desarrollo de aplicaciones cruzadas y entre dominios. Un ejemplo de estas iniciativas es BIG IoT [1] [1]. Por el momento, el proyecto BIG IoT comprende 8 plataformas IoT y creciendo. BIG IoT se centra en establecer la interoperabilidad entre plataformas, y tiene dos objetivos principales. El primero es definir una interfaz compartida, es decir, la llamada API BIG IoT, que comprende funcionalidades comunes tales como descubrimiento, acceso y manejo de eventos. Esta API debe ser compatible con todas las plataformas participantes, (además de su interfaz propietaria, si existe), como se ilustra en la Figura 1. El segundo objetivo es establecer un *Marketplace* centralizado donde registrar plataformas y servicios. En BIG IoT existen múltiples escenarios piloto involucrando varias plataformas IoT, servicios y aplicaciones del dominio Smart Cities. En la Sección IV explicaremos uno de estos escenarios.

Este artículo pretende explicar el análisis y acciones específicas de seguridad y privacidad en el proyecto BIG IoT. El resto de artículo se estructura de esta manera: la Sección II presenta un conjunto de requisitos de seguridad para BIG IoT, así como discusiones y acciones actuales para abordarlos. La Sección III describe las recomendaciones de privacidad para el ecosistema IoT. La Sección IV, analiza un ejemplo de caso de uso desde el punto de vista de seguridad y privacidad. Finalmente, la Sección V presenta las conclusiones.

II. Asegurando un ecosistema IoT

El marketplace BIG IoT (la API común a todas las plataformas/servicios/applicaciones) debe cumplir un conjunto de requisitos de seguridad. Después de un análisis de las necesidades de BIG IoT, se identificaron siete requisitos de seguridad y algunas soluciones que se enumeran a continuación.

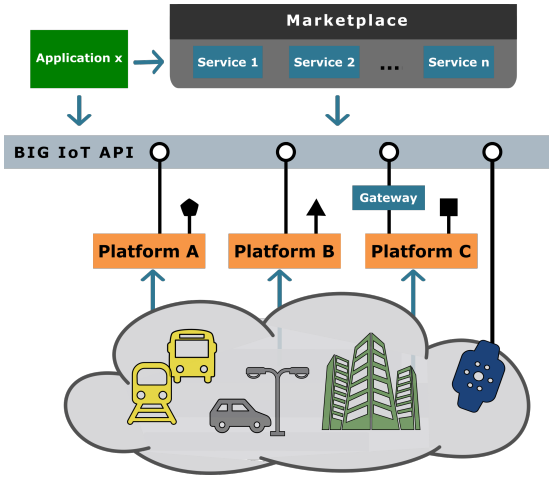


Figure 1. El enfoque BIG IoT de un ecosistema de plataformas IoT.[2]

A. Requisitos de seguridad

1) **Seguridad extremo a extremo**. Una posible solución para proporcionar seguridad en las comunicaciones sería dejar los mecanismos ya en uso para cada plataforma, y luego definir las políticas de adaptación entre plataformas. Esto aumentaría la complejidad cuando el número de plataformas creciera. La solución adoptada en BIG IoT es proporcionar seguridad en el nivel de la API, porque es común para todas las plataformas. Por lo tanto, no es necesario adaptar los mecanismos de protección entre las plataformas, ya que la API es extremo a extremo por naturaleza.
2) **"Batteries included but swappable"**[3]. Pueden aparecer nuevos ataques, cripto-sistemas, contramedidas, técnicas y topologías, pero el sistema IoT debe ser capaz de hacer frente a estas preocupaciones emergentes mucho después de que se haya implementado el sistema [2]. BIG IoT utiliza una implementación de seguridad que puede ser intercambiable, no codificada de forma rígida para protocolos/sistemas de seguridad específicos.
3) **Autenticación/Autorización flexible**. Los sistemas de autenticación y autorización utilizados en el ecosistema BIG IoT deben facilitar la gestión de identidades y permisos. Las diversas plataformas que forman BIG IoT pueden tener arquitecturas e interfaces de adminis-

[1] http://big-iot.eu
[2] Iconos obtenidos en Freepik http://www.flaticon.com
[3] https://blog.docker.com/2016/03/docker-networking-design-philosophy/

tración de usuarios muy diferentes. Además, los administradores de dichas plataformas pueden no querer exponer ciertas funciones de administración debido a preocupaciones comerciales o de seguridad. Por lo tanto, es necesario dar soporte a varios modelos, incluida la autenticación descentralizada, federada o delegada.

4) **Transferencia de propiedad**. BIG IoT apoya la transferencia segura de propiedad, incluso si un componente se vende o transfiere a un competidor; algo que sucede a menudo durante la vida útil de dichos componentes.
5) **Contabilidad y cobros**. BIG IoT debe implementar una contabilidad segura del consumo de recursos. Esta contabilidad generarmente se realiza en base a registros de datos de cobro (CDR). Como resultado de una política de cobro, puede ser necesario un sistema de facturación que pueda generar facturas para los consumidores del servicio. Todos estos sistemas deben ser lo suficientemente flexibles como para implementar diferentes modelos de negocios y estrategias de monetización de los servicios que pueden implementarse en el ecosistema BIG IoT. El mercado BIG IoT debe admitir cargos y facturación en línea y fuera de línea.
6) **Seguridad continua**. El sistema BIG IoT debería estar listo para responder a participantes hostiles, nodos comprometidos y cualquier otro evento adverso. Es necesario implementar mecanismos y/o herramientas para volver a emitir credenciales, excluir participantes, distribuir parches de seguridad, actualizaciones, etc.
7) **Desarrollo seguro**. La seguridad debe ser parte esencial durante el diseño del software BIG IoT, pero un diseño seguro es inútil si hay errores de desarrollo que den pie a ataques y/o vulnerabilidades. El uso de un ciclo de vida de desarrollo de software seguro (S-SDLC) y el análisis de código fuente seguro (SCA) ayudaría a los desarrolladores a construir un software más seguro que cumpla los requisitos de seguridad.

B. Abordando los requisitos de seguridad en BIG IoT

Estas son algunas de las acciones que se han adoptado en BIG IoT para abordar los requisitos anteriores.

El *Requisito 1* se cumple directamente ya que la API BIG IoT está basada en HTTP(s), por lo que es extremo a extremo. Además, para cumplir con el *Requisito 2*, la API debe ser lo suficientemente flexible como para manejar cualquier protocolo y/o contenido. BIG IoT define una API muy genérica, con anotaciones semánticas y descripciones sintácticas de cada servicio y plataforma registrados, que se utilizan luego para aclarar los detalles sobre cómo establecer la comunicación entre estos componentes.

El *Requisito 3* establece que es necesario proporcionar autenticación flexible en el ecosistema de IoT. Para resolver esto, BIG IoT utiliza un enfoque que es similar a los utilizados por otras iniciativas de IoT ampliamente conocidas (por ejemplo, [3]): tokens firmados. Un cliente presenta un token firmado a un servidor para demostrar que es capaz de realizar una acción determinada. Cuando el servidor recibe el token firmado puede confiar en los contenidos porque el token está firmado por un nodo raíz de confianza que es común.

Muchas tecnologías usan tokens firmados. La mayoría de soluciones para la Web usan codificaciones JSON, CBOR o XML y confían en JSON Web Encryption (JWE) [4], JSON Web Signature (JWS) [5], XML Encryption (XML-Enc) [6] o XML Signature (XML-Sig) [7]. La posición de BIG IoT es adoptar tecnologías de seguridad ya probadas. Dado que la API BIG IoT se basa en HTTP REST, posibles candidatos son SAML [8], Oauth 1 [9], OAuth 2 [10], o OpenID Connect [11] (basado en OAuth 2), apoyando la autorización delegada y la autenticación/identificación.

El *Requisito 4* también se verá afectado por la elección anterior, ya que el sistema de autenticación/autorización debe definirse centrándose en la administración sencilla de identidades y permisos. Esto incluye la transferencia segura de recursos y la respuesta rápida a topologías dinámicas con frecuentes entradas y salidas del sistema.

El *Requisito 5* establece que una contabilidad adecuada es clave para desarrollar sistemas de facturación y cobro, tanto fuera como en línea. Un sistema de cobro fuera de línea sólo almacena un CDR que contiene la información contable y de cobro relevante. En ambos enfoques (fuera y en línea), debería ser deseable tener pruebas de no repudio.

El *Requisito 6* obliga al marketplace a disponer de un repositorio donde descargar de forma segura actualizaciones y parches. El éxito de una solución como eśta depende de la facilidad de uso tanto para los usuarios finales como para los desarrolladores. Los markets de Apple, Google y Amazon son buenos ejemplos, pero BIG IoT apunta a un enfoque más abierto para este componente.

El *Requisito 7* requiere el uso de S-SDLC. Los desarrolladores de BIG IoT siguen las directrices para el desarrollo de software seguro con el proyecto de seguridad de aplicaciones web abiertas (OWASP) [12]: (1) la organización misma tiene que cumplir con el marco de OWASP Software Assurance Maturity Model (SAMM) [13]; (2) las aplicaciones cumplen con los requisitos del estándar de verificación de seguridad de la aplicación OWASP (ASVS) [14]; (3) el código fuente de la aplicación se analizará de acuerdo con las directrices de análisis de código fuente seguro (SCA) de OWASP [15]; (4) finalmente, la aplicación se probará en busca de vulnerabilidades y defectos de diseño de acuerdo con las directrices de prueba de OWASP [16].

OWASP ASVS define acualmente 19 requisitos de seguridad para aplicaciones y servicios, con tres niveles de verificación de seguridad: ASVS Nivel 1 ”Oportunista”, para todo el software y su cumplimiento afecta a las vulnerabilidades de seguridad de las aplicaciones que son fáciles de descubrir; ASVS Nivel 2 ”Estándar”, destinado a aplicaciones que contienen datos confidenciales, incluyendo aquellos que procesan información de salud; y ASVS Nivel 3 ”Avanzado”, diseñado para aplicaciones críticas, que requieran del más alto nivel de confianza. Las responsabilidades incluyen controles para garantizar la confidencialidad, integridad, disponibilidad, autenticación, no repudio, autorización y auditoría. Cada nivel de ASVS contiene una lista de requisitos de seguridad, cada uno con características y capacidades específicas de seguridad para que los desarrolladores los incorporen en el software. En BIG IoT, los desarrolladores deben (al menos) seguir los requisitos de ASVS nivel 2.

III. Buenas prácticas de privacidad en ecosistemas IoT

Los ecosistemas IoT implican tal cantidades de datos, a menudo sensibles, que su uso podría ser una amenaza para la privacidad de los usuarios. La FTC publicó en 2015 una guía que contiene las mejores prácticas para la privacidad en IoT [17] que se resume con la siguiente declaración: mientras que la flexibilidad en términos de recopilación de datos es clave para innovar en torno a nuevos usos de datos, la cantidad de datos a almacenar debe ser equilibrado para asegurar la privacidad y limitar los riesgos de seguridad para los consumidores. Estas recomendaciones son útiles y válidas en el ámbito europeo. Sin embargo, son genéricas y deben complementarse siempre con un análisis específico de cada caso de uso (se proporciona un ejemplo en la Sección IV). A continuación, proporcionamos las ideas principales detrás de las recomendaciones de la FTC.

A. Minimización de datos

La minimización de datos es un principio de protección de privacidad [18] que implica que un *data controller*[4] debe limitar la recopilación de información personal a lo que es directamente relevante y necesario para lograr un propósito específico. La minimización de datos es clave para fomentar el ecosistema de IoT, y puede ayudar a proteger contra dos riesgos relacionados con la privacidad: (1) el almacenamiento de grandes volúmenes de datos aumenta la probabilidad de que se produzca una violación de datos, ya que el daño potencial derivado es mayor; (2) la recopilación y el almacenamiento de grandes cantidades de datos también aumenta el riesgo de utilizar los datos de una manera que se aparta de las expectativas razonables de los consumidores. Para minimizar estos riesgos, las organizaciones deben desarrollar políticas y prácticas de minimización de datos que brinden respuestas a preguntas como qué tipos de datos se están recopilando, para qué fin y cuánto tiempo se deben almacenar.

En la UE, el principio de minimización de datos se deriva del artículo 6.1 (b) y (c) de la Directiva 95 / 46 / EC [18] y el artículo 4.1(b) y (c), del Reglamento EC 45/2001 [19], que afirman que los datos personales se deben ”recopilar con fines específicos, explícitos y legítimos” y deben ser ”adecuados, relevantes y no excesivos en relación con los propósitos para los cuales se recopilan y/o procesan”.

Cuando una empresa necesita recopilar y almacenar datos confidenciales con un fin comercial, debe considerar si puede hacerlo con un conjunto de datos no identificados. Los datos no identificados pueden reducir el daño potencial al consumidor y al mismo tiempo promover usos sociales beneficiosos. Una clave para una no identificación efectiva es asegurar que los datos no puedan ser razonablemente re-identificados, incluso con fuentes externas cruzadas. Esto generalmente requiere eliminar identificadores o pseudoidentificadores. El reconocimiento de identificadores no evidentes es todo un desafío que a menudo debe afrontarse de manera manual.

En BIG IoT, para cada caso de uso específico, se ha realizado un análisis de posibles identificadores entre los datos y metadatos almacenados/intercambiados. Como directriz en BIG IoT, siempre se recomienda la minimización de datos, teniendo en cuenta datos cruzados no solo desde otra plataforma/servicios, sino también desde cualquier otra fuente externa. En el caso de uso de la Sección IV se hace uso de técnicas de minimización de datos.

A menudo se cae en el error de pensar que la minimización de datos supone un coste adicional. De hecho, la minimización de datos reduce la sensibilidad de los datos y, por lo tanto, se requiere de medidas de seguridad menos costosas, lo cual supone generalmente un ahorro en el coste de desarrollo para cumplir el nivel de ASVS.

B. Responsabilidad

Dado que existe la posibilidad que los datos no identificados puedan ser re-identificados, se deben tener mecanismos de responsabilidad. En este contexto, la FTC ha declarado que las compañías que declaran que mantienen datos anónimos o identificados deben cumplir con tres acciones: (1) tomar medidas razonables para no identificar los datos, incluso manteniéndose al día con los avances tecnológicos; (2) comprometerse públicamente a no volver a identificar los datos; y (3) tener contratos vigentes con terceros con quienes comparten los datos, exigiendoles que también se comprometan a no volver a identificar los datos. Este enfoque asegura que si los datos se vuelven a identificar, los reguladores puedan responsabilizar a la empresa. Por tanto, las plataformas, servicios y aplicaciones BIG IoT deben proporcionar mecanismos adecuados para registrar de forma segura cualquier acción por parte de cualquier actor que maneje datos confidenciales.

C. Transparencia y fácil acceso

La legislación principal a nivel de la UE en el ámbito de la protección de datos es la ”Directiva de Protección de Datos” [18], que se implementa en los Estados miembros a través de leyes nacionales. Esta directiva tiene como objetivo proteger los derechos y libertades de las personas al establecer pautas que determinan cuándo el procesado de datos personales es legal. Las pautas se relacionan principalmente con la calidad de los datos, la legitimidad del procesado, las categorías especiales de datos, la información que debe darse al interesado, el derecho de acceso de los interesados a sus datos, el derecho a oponerse al procesado de datos, la confidencialidad y seguridad del procesado y la notificación del procesado a una autoridad supervisora. La Directiva también establece principios para la transferencia de datos personales a países terceros y establece autoridades de protección de datos en cada Estado miembro de la UE.

Las personas necesitan mejor información sobre las políticas de protección de datos y sobre qué sucede con sus datos. Como resultado, la UE exigirá a las organizaciones europeas políticas de protección de datos fácilmente accesibles. En este contexto, el uso de iconos simples en sitios web o aplicaciones podrían explicar de manera visual y sencilla cómo, por quién y bajo qué responsabilidad se procesarán los datos personales. A modo de ejemplo, la Figura 2 muestra un conjunto de iconos de privacidad propuestos por Aza Raskin [20]. Los servicios y/o aplicaciones BIG IoT deben usar iconos como estos para indicar claramente a los usuarios finales cómo se procesan sus datos.

[4]Una entidad que procesa y/o almacena información personal

Figure 2. Iconos de privacidad propuestos por Aza Raskin [20]

IV. Caso de uso: asistente inteligente de transporte

En esta sección describimos un caso de uso BIG IoT, un asistente de transporte inteligente, con el objeto de analizar y discutir los aspectos principales de seguridad y privacidad. Un suscriptor de la aplicación está en casa y quiere ir a un lugar específico. La aplicación le asistirá proporcionando información sobre transporte privado y público. Si decide usar su vehículo privado, le ayuda con información de navegación, condiciones de tráfico y lugares de estacionamiento disponibles. La aplicación también puede sugerir varias formas de transporte público como alternativa al vehículo privado. Por ejemplo, la aplicación permite al usuario seleccionar una línea de autobús. El usuario puede ver información sobre el próximo autobús que llega a la parada seleccionada, dónde se encuentra el autobús actualmente, si está retrasado, número de personas en el autobús y un pronóstico del número de personas en el autobús cuando llegue a su parada. Este caso de uso se va a implementar en dos pilotos diferentes del proyecto BIG IoT, uno centrado en el transporte privado que se desplegará en Barcelona, y el otro principalmente centrado en el transporte público, que se desplegará en Wolfsburg.

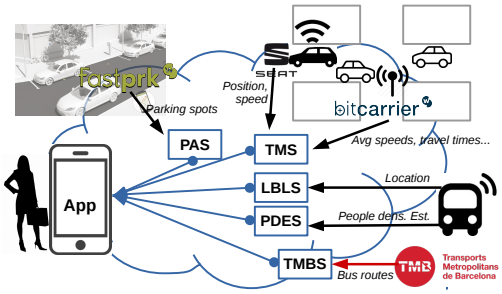


Figure 3. Caso de uso: un asistente de transporte en el ecosistema BIG IoT

Desde el punto de vista del usuario, toda la funcionalidad se accede con una sola aplicación en su teléfono. Sin embargo, la aplicación está consumiendo datos de varios servicios, que a su vez consumen datos proporcionados por sensores físicos operados por diferentes plataformas (ver Figura 3)

Los servicios son una capa de abstracción importante en BIG IoT porque permiten la reutilización del código y simplifican el proceso de creación de aplicaciones BIG IoT. Por ejemplo, un servicio puede agregar datos adquiridos de diferentes plataformas y luego presentar un conjunto de datos unificado a las aplicaciones. Otro servicio podría administrar el historial de datos, permitiendo al usuario acceder a datos pasados (datos no disponibles actualmente en la plataforma fuente). Otro servicio podría crear pronósticos a partir de datos (la adquisición de datos desde una plataforma o desde otro servicio). Finalmente, otro uso interesante podría ser un servicio que anonimiza un conjunto de datos subyacente. Esto podría permitir niveles más bajos de seguridad ASVS en la capa superior (aplicación). A continuación describiremos los componentes BIG IoT involucrados y discutiremos sus aspectos de seguridad y privacidad.

Plataforma 1: bitcarrier de antenas WiFi/Bluetooth [21]

Esta plataforma se alimenta de datos recopilados por antenas WiFi y Bluetooth ubicadas en cruces de calles. La tecnología detecta vehículos/usuarios por sus direcciones MAC únicas y proporciona tiempos de viaje promedio, velocidad, e incluso medidas de congestión de la calle. Las direcciones MAC son datos muy confidenciales, que podrían usarse para rastrear o registrar los hábitos de los vehículos (e incluso de los usuarios). Este posible ataque a la privacidad debe evitarse, así que todas las partes involucradas deben aceptar los contratos legales necesarios en los que aceptan utilizar correctamente los datos almacenados/intercambiados. Además, estos datos deben estar apropiadamente asegurados/anonimizados para evitar cualquier tipo de fuga.

El enfoque BIG IoT es anonimizar inmediatamente las

direcciones únicas usando una función de hash criptográfico unidireccional. Esta función utiliza como entradas: (1) la dirección a anonimizar y (2) una clave p, que se actualiza periódicamente. De esta manera, un dispositivo no puede ser rastreado por más de un período. La frecuencia con la que se actualiza la clave (desde minutos hasta días) es parte de la política de privacidad. El uso de funciones de este tipo permite anonimizar los datos, pero manteniendo la posibilidad de volver a identificar vehículos/usuarios. El operador de la plataforma debe mantener en secreto las claves temporales utilizadas para anonimizar los identificadores. Sin embargo, los operadores pueden verse obligados a revelar estas claves en algunas circunstancias, p.ej. un requerimiento legal. Esta plataforma debe cumplir al menos con la línea de base recomendada ASVS nivel 2 "estándar", mientras que la gestión de las claves de anonimización debe cumplir con ASVS nivel 3 "avanzado".

Plataforma 2: los coches de SEAT

SEAT ha instalado varios coches con sensores integrados en BIG IoT, que envían su posición y velocidad actual. Proporcionar estos datos puede permitir rastrear un vehículo y por lo tanto se puede considerar una amenaza a la privacidad. Se aplica el mismo razonamiento que para la plataforma 1: si los identificadores no se pueden eliminar de los datos proporcionados, al menos deben anonimizarse adecuadamente, por ejemplo, con una función de hash. Los mismos requisitos de seguridad de ASVS se aplican tanto a la plataforma de Bitcarrier como a los automóviles SEAT, es decir, los automóviles deben cumplir al menos con el nivel ASVS 2, pero el administrador de claves requeriría el nivel 3.

Plataforma 3: Fastprk, estacionamiento en la calle [22]

Esta plataforma puede proporcionar el estado individual de los lugares de estacionamiento en un área supervisada. Por ejemplo, para los casos de uso de BIG IoT Barcelona, esta plataforma actualmente ofrece información para 600 plazas de estacionamiento en la calle. Este tipo de datos conlleva riesgos de privacidad específicos debido a la correlación con otras fuentes: si un atacante sabe dónde alguien ha estacionado su coche, puede monitorizar cuando se va simplemente comprobando el estado de dicha plaza.

Una contramedida sería, por ejemplo, proporcionar los espacios libres en un segmento de calle (un lote virtual). Este enfoque garantizaría el k -anonimato (un individuo determinado no puede diferenciarse de otros $k-1$) de vehículos/usuarios supervisados siendo k el número de vehículos estacionados en el mismo segmento. Esto cumple con el requisito de anonimización/desidentificación de la información confidencial establecida anteriormente. Cuanto mayores son los segmentos, más anónimo es el servicio, pero el PAS proporcionará información menos específica, y potencialmente menos útil.

Ofrecer al usuario el segmento en lugar de la posición exacta de estacionamiento libre puede ser igualmente útil en este caso. Sin embargo, buscar el equilibrio adecuado entre privacidad y usabilidad puede ser complejo. Dado que los datos almacenados por la plataforma se pueden usar de alguna manera para seguir/monitorear a los usuarios finales, recomendamos asegurar esta plataforma BIG IoT siguiendo el nivel de verificación 2 de ASVS. Sin embargo, si la plataforma simplemente almacena y proporciona lugares de estacionamiento gratuitos en un segmento/lote predefinido, se podría considerar el nivel de verificación 1 de ASVS.

Plataforma 4: sensores de wifi probe en autobuses

Estos sensores están colocados en autobuses que recopilan las *probe request* a las wifi. Dichas peticiones contienen las direcciones MAC de los usuarios, que son únicas y deben ser anonimizadas. De hecho, el funcionamiento de esta plataforma es similar al funcionamiento de la plataforma 1 y se aplican las mismas recomendaciones de seguridad y privacidad.

Plataforma 5: sensores de localizacion en buses

Estos sensores, ubicados en los autobuses, proporcionan datos de ubicación y marcas de tiempo. Los datos recopilados no se almacenan en el sensor, ya que una ubicación obsoleta no sería de mucha utilidad. Como las posiciones de los autobuses públicos no son privadas, no se deben tomar medidas de privacidad específicas. El desarrollo de software debe cumplir con el estándar ASVS nivel 2 y, si está debidamente justificado, incluso con ASVS nivel 1.

Servicio 1: Servicio de monitorización de tráfico (TMS)

Este servicio proporciona rutas a vehículos según las condiciones de tráfico. TMS consume datos de las plataformas 1 y 2, así como un mapa de la ciudad. Suponiendo que los datos proporcionados por las plataformas 1 y 2 son anonimizados, no se requieren requisitos específicos en términos de privacidad. Con respecto a la seguridad, el desarrollo de software debe cumplir con el estándar ASVS nivel 2 y, si está debidamente justificado, incluso con ASVS nivel 1.

Servicio 2: Servicio de estacionamiento disponible (PAS)

PAS se alimenta del estado de estacionamiento en la calle que proporciona la plataforma 3. PAS puede almacenar estadísticas/histórico de dichas plazas de estacionamiento, así que se aplican las mismas recomendaciones de privacidad que para la plataforma Fastprk. Tanto el servicio como sus conexiones deben estar protegidos con ASVS nivel 2 o ASVS nivel 1 si se proporciona un nivel aceptable de k-anonimato.

Servicio 3: Servicio externo de datos de ruta de buses TMB (TMBS)

Transports Metropolitans de Barcelona (TMB) es el principal operador de transporte público en el área metropolitana de Barcelona. Ya tiene una API de datos abiertos [23] donde obtener rutas a destinos con diferentes transportes públicos: trenes, metro, autobuses. El TMBS hace de API de TMB en el ecosistema BIG IoT. Dado que todos los datos involucrados en el servicio son públicos, no se requieren requisitos específicos en términos de privacidad. Con respecto a la seguridad, a menos que esté debidamente justificado, el desarrollo de software debe cumplir con el estándar ASVS nivel 2.

Servicio 4: Servicio de estimación de densidad de personas en bus (PDES)

El PEDS consume datos de los sensores de wifi probe y proporciona información sobre la cantidad de personas en los autobuses. Los datos incluidos son la identificación del bus, estimación del número de personas, indicador de precisión e indicación de la hora. Dichos datos se almacenan durante una

duración fija en el servicio, lo que significa que se pueden solicitar dentro de esta duración. Después de ese tiempo, sólo se almacena información histórica más general, que también está disponible. Se recomienda que el servicio cumpla con el nivel 2 de ASVS, pero incluso podría ser el nivel 1 de ASVS ya que no se manejan datos específicos del usuario. Sin embargo, para proteger un posible uso comercial del servicio se recomienda el nivel 2 de ASVS.

Servicio 5: Servicio de localización de bus en vivo (LBLS)

Este servicio consume datos de los sensores de ubicación en los autobuses y proporciona información a la aplicación de ubicación. Los datos proporcionados consisten en la identificación del sensor, la ubicación y la marca de tiempo. Los datos se almacenan en el servicio durante un breve período de tiempo fijo, después del cual se eliminan. El servicio no maneja ningún dato específico del usuario, por lo que se recomienda cumplir con el nivel 1 de ASVS.

App móvil para usuarios finales

La aplicación consume datos proporcionados por los 5 servicios. Se espera que las aplicaciones en BIG IoT ofrezcan funciones de transparencia, tal como se menciona en las recomendaciones de privacidad anteriores, por ejemplo mediante iconos de privacidad. Los desarrolladores también deberían considerar los requisitos de privacidad más allá del alcance de la infraestructura BIG IoT, como el consentimiento informado o el derecho de acceso a la información personal, la rectificación o la eliminación. Con respecto al desarrollo seguro, las aplicaciones deben seguir el nivel 2 de ASVS.

V. Conclusiones

El proyecto BIG IoT tiene como objetivo establecer la interoperabilidad entre plataformas para generar un ecosistema IoT. Los pilares tecnológicos básicos de BIG IoT son una API común, así como un marketplace para todos los participantes del ecosistema IoT. La clave de su éxito es definir niveles adecuados de seguridad y privacidad.

Con respecto a la seguridad, en este documento hemos identificado siete requisitos que deben seguirse al crear y/o implementar componentes BIG IoT. Dichos requisitos afectan el diseño de la API BIG IoT y el marketplace, así como cualquier software en el ecosistema BIG IoT. Después de este análisis, hemos delineado cómo estos requisitos afectarán el enfoque arquitectónico de BIG IoT.

Con respecto a la privacidad, hemos propuesto tres recomendaciones: 1) minimización de datos, es decir, que un controlador de datos debe limitar la recopilación de información personal a lo que es directamente relevante y necesario para lograr un propósito específico; 2) una fuerte responsabilidad, es decir, proporcionar mecanismos para registrar de forma segura cualquier acción por parte de cualquier actor que se ocupe de datos confidenciales; y 3) transparencia y fácil acceso, es decir, cualquier controlador de datos debe publicar sus políticas de protección de datos para que sean fácilmente accesibles.

Finalmente, se presenta y analiza la seguridad y privacidad de un caso de uso: un asistente inteligente de transporte. Este caso de uso se está implementando en dos pilotos del proyecto BIG IoT, en Barcelona y Berlín/Wolfsburg.

Agradecimientos

Este trabajo se ha financiado a partir del proyecto BIG IoT (European Commission's Horizon 2020, No 688038),y los proyectos MINECO/FEDER ANFORA (TEC2015-68734-R) y ARPASAT (TEC2015-70197-R), y por el grupo de calidad de la Generalitat de Catalunya 2017-SGR-782.

References

[1] A. Bröring, S. Schmid, C.-K. Schindhelm, A. Khelil, S. Kaebisch, D. Kramer, D. Le Phuoc, J. Mitic, D. Anicic, and E. Teniente, "Enabling IoT Ecosystems through Platform Interoperability," *IEEE Software*, no. Software Engineering for the Internet of Things, 2017, forthcoming.
[2] OWASP Internet of Things project. Principles of iot security. [Online]. Available: https://www.owasp.org/index.php/Principles_of_IoT_Security
[3] Allseen Alliance. Alljoyn Framework. [Online]. Available: https://allseenalliance.org/framework
[4] M. Jones and J. Hildebrand. JSON Web Encryption (JWE). [Online]. Available: https://datatracker.ietf.org/doc/rfc7516/
[5] M. Jones, J. Bradley, and N. Sakimura. JSON Web Signature (JWS). [Online]. Available: https://datatracker.ietf.org/doc/rfc7515/
[6] T. Imamura, B. Dillaway, and E. Simon. XML Encryption Syntax and Processing. [Online]. Available: https://www.w3.org/TR/xmlenc-core/
[7] M. Bartel, J. Boyer, B. Fox, B. LaMacchia, and E. Simon. XML Signature Syntax and Processing (Second Edition). [Online]. Available: https://www.w3.org/TR/xmldsig-core/
[8] Organization for the Advancement of Structured Information Standards (OASIS). Official Wiki of the OASIS Security Services (SAML) Technical Committee. [Online]. Available: https://wiki.oasis-open.org/security/FrontPage
[9] IETF OAuth WG. OAuth 1. [Online]. Available: https://oauth.net/1/
[10] ——. OAuth 2.0. [Online]. Available: https://oauth.net/2/
[11] OpenID Foundation. OpenID connect. [Online]. Available: http://openid.net/connect/
[12] Open Web Applications Security Project (OWASP). [Online]. Available: https://www.owasp.org/
[13] The OWASP Software Assurance Maturity Model (SAMM). [Online]. Available: https://www.owasp.org/index.php/Category:Software_Assurance_Maturity_Model
[14] OWASP. Application security verification standard 3.0.1. [Online]. Available: https://www.owasp.org/images/3/33/OWASP_Application_Security_Verification_Standard_3.0.1.pdf
[15] OWASP Code Review Project second edition guideline. [Online]. Available: https://www.owasp.org/index.php/Category:OWASP_Code_Review_Project
[16] M. Meucci and A. Muller. OWASP testing guideline version 4. [Online]. Available: https://www.owasp.org/index.php/OWASP_Testing_Project
[17] FTC Staff. (2015, Jan.) Internet of things: Privacy and security in a connected world. [Online]. Available: https://www.ftc.gov/system/files/documents/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf
[18] EU Legislation. (1995) Directive 95/46/ec. [Online]. Available: https://secure.edps.europa.eu/EDPSWEB/edps/site/mySite/pid/74#data_directive
[19] ——. (2001) Directive 45/2001/ec. [Online]. Available: https://secure.edps.europa.eu/EDPSWEB/edps/site/mySite/pid/86#regulation
[20] A. Raskin. Privacy icons. [Online]. Available: https://www.flickr.com/photos/azaraskin/5304502420/sizes/o/
[21] Worldsensing's Bitcarrier. [Online]. Available: http://www.bitcarrier.com/
[22] Worldsensing's Fastprk. [Online]. Available: http://www.fastprk.com/
[23] Transport Metropolitans de Barcelona. TMB Open Data. [Online]. Available: https://www.tmb.cat/en/web/tmb/about-tmb/open-data
[24] T. Berners-Lee, J. Hendler, O. Lassila *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

Pruebas Basadas en el Entorno para la Detección de Ataques de Relay en Accesos a Zonas Restringidas

Carles Anglès-Tafalla*, Jordi Castellà-Roca*, Alexandre Viejo*,
M. Magdalena Payeras Capellà†, Macià Mut Puigserver†
*Departament d'Enginyeria Informàtica i Matemàtiques, UNESCO Chair in Data Privacy, Universitat Rovira i Virgili,
Av. Països Catalans 26, E-43007 Tarragona, Spain
{carles.angles,jordi.castella,alexandre.viejo}@urv.cat
†Departament de Ciències Matemàtiques i Informàtica, Universitat de les Illes Balears,
Ctra. de Valldemossa, km 7,5. E-07122 Palma, Spain
{mpayeras,macia.mut}@uib.es

***Resumen*—Las zonas de bajas emisiones (LEZ) son áreas urbanas donde el acceso de vehículos contaminantes está restringido con el fin de reducir las emisiones y mejorar la calidad del aire. Aunque muchas grandes ciudades han aplicado este método, los sistemas de control de acceso que implementan dichas zonas han levantado alarma social por su naturaleza intrusiva. Como consecuencia, se han propuesto sistemas que respetan la privacidad de los usuarios siempre que estos se comporten honestamente. No obstante, esto abre las puertas a ataques de confabulación, como ataques de "relay", que pueden llevar a suplantar de identidad de usuarios. Para paliar esta problemática, este artículo propone un sistema de generación de pruebas distribuido basado en el entorno en el que los usuarios cercanos, de forma privada, acreditan entre sí su posición y sus tiempos de respuesta con el fin detectar un ataque de "relay". El conjunto de pruebas generadas permiten determinar cuándo las credenciales de un usuario están siendo transmitidas en tiempo real y determinar si dicho usuario está confabulado con el atacante, caso en el que es posible revocar su privacidad.**

***Index Terms*—LEZ (Low Emission Zones), Ciudades inteligentes (Smart Cities), Privacidad (Privacy), Seguridad (Security), Pruebas de localización (Location proofs), Ataques de "relay" (Relay Attacks).**

I. Introducción

Los altos niveles de contaminación ambiental, derivados en gran parte de la congestión de tráfico urbano, se han convertido en un grave problema para las grandes ciudades en todo el mundo. En los núcleos urbanos de estas grandes áreas metropolitanas, los niveles de contaminación superan con creces algunos de los límites establecidos por la Organización Mundial de la Salud [1] suponiendo un peligro para la salud de sus ciudadanos. Para abordar este problema, las administraciones gubernamentales han empezado a aplicar medidas para fomentar un uso racional de vehículos, que incluyen, entre otros, restricciones en la circulación de vehículos contaminantes, carriles para vehículos de alta ocupación (VAO) o la delimitación de zonas de bajas emisiones (LEZ).

De entre las mencionadas, la implantación de LEZs, que consiste en un área donde se aplican restricciones o peajes a sus usuarios de acuerdo con las emisiones de sus vehículos, es una de las medidas que más ha proliferado. Suecia, Italia, Holanda, Reino Unido o Alemania son ejemplos de países que implementan este tipo de zonas en sus grandes ciudades. Ante esta tendencia, surge la necesidad de implementar sistemas de control de acceso a las LEZ que permitan gestionar las restricciones o peajes a los que deben someterse los vehículos que circulan por estas zonas. El caso más citado en la literatura es, sin duda, la LEZ de Londres y su controvertido sistema de control [2]. Una red integrada por más de 300 cámaras es el que compone la columna vertebral de dicho sistema, cuyo fin consiste en fotografiar indiscriminadamente las matrículas de todos los vehículos que circulan por su interior, para posteriormente identificar si dichos vehículos abonan sus correspondientes tasas.

Sistemas de cariz tan invasivo como el mencionado han propiciado la aparición de propuestas diseñadas en torno a la privacidad de los usuarios. Sistemas de control de acceso como los presentados en [3], [4] promueven respetar la privacidad de los usuarios en todo momento, y solo revocarles ese derecho en el momento en que se comportan de forma deshonesta. De este modo, la matricula del vehículo solo es fotografiada en caso de que el usuario omita, total o parcialmente, el proceso de autenticación con la infraestructura de control de acceso del sistema.

El enfoque de privacidad revocable aplicado a los sistemas de control de acceso supone un gran aliciente para los usuarios de dichos sistemas, pero también abre nuevas vías de ataque en las que el usuario puede intentar suplantar a otros usuarios amparado por su anonimato. Este tipo de ataques cobran especial sentido en LEZs, donde los vehículos que menos contaminan tienen tasas más bajas o incluso nulas.

Esto puede llevar a usuarios a realizar acciones fraudulentas dirigidas a confabularse o suplantar a usuarios que dispongan de credenciales para vehículos de bajas emisiones.

Para evitar ataques basados en el robo de credenciales y posterior suplantación de los usuarios, los servicios basados en localización vehicular (VLBS) hacen uso de hardware específico a prueba de manipulaciones, como una unidad de a bordo (OBU) o similar.

No obstante, esta medida es insuficiente contra los ataques basados en la confabulación, en los cuales se suplanta la identidad de un usuario simplemente redireccionando los paquetes recibidos hasta el vehículo que se desea suplantar y reproduciendo sus respuestas. Por medio de este ataque de "relay" o baliza, se hace creer a la infraestructura del sistema que el vehículo suplantado se encuentra accediendo a la zona restringida cuando en realidad se encuentra en otro lugar. (ver Figura 1).

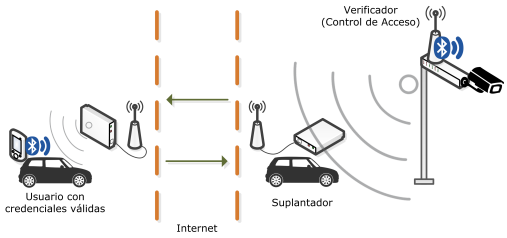


Figura 1. Ataque de baliza o "relay"

I-A. Contribución y organización del trabajo

Con la intención de abordar problemas todavía por resolver en la literatura actual, cuando no se puede asumir la integridad e irrefutabilidad de las ubicaciones de los dispositivos de localización, en este artículo afrontamos la problemática suscitada por el ataque de "relay" por medio de herramientas presentes en la literatura como pruebas de localización distribuidas y el análisis de tiempos de respuesta en las comunicaciones.

Con esto en mente, proponemos un nuevo sistema generador de pruebas distribuido, basado en el entorno y respetuoso con la privacidad de los usuarios, a partir de las cuales sea posible detectar un ataque de "relay". En dicho sistema, las entidades cercanas corroboran tanto la posición como los tiempos de respuesta de los usuarios con el fin de calcular un umbral del sobrecoste en la comunicación propio de un ataque de "relay". Las pruebas generadas actúan como garantía de que credenciales de un usuario no están siendo transferidas remotamente en tiempo real desde otra ubicación. Más concretamente, nuestras contribuciones son las siguientes:

- Proporcionar un sistema de generación de pruebas basado en el entorno cercano donde se respeta la anonimidad y privacidad de todos los usuarios implicados en el proceso. Las pruebas recolectan información sobre la posición y los tiempos de respuesta del entorno cercano con el objetivo de detectar comportamientos propios de un ataque de "relay".
- Permitir identificar al menos a uno de los usuarios que se han confabulado para perpetrar el ataque de "relay". Al generar pruebas que incluyen la ubicación, es posible determinar si el usuario que está siendo suplantado está confabulado con el atacante, en cuyo caso es identificado, o se trata de un usuario honesto víctima del ataque, con lo cual la prueba generada es desechada.

La contribución está organizada de la forma siguiente. En la Sección II se debate el estado actual de la literatura. En la Sección III se describe el modelo del sistema dando una visión general de la propuesta. La Sección IV contiene el protocolo en el que se basa el sistema presentado. La Sección V incluye un estudio del control de fraude de la propuesta. La Sección VI incluye un breve estudio sobre privacidad que ofrece el sistema. Finalmente, las conclusiones se presentan en la Sección VII.

II. Antecedentes

El uso de pruebas de localización o basadas en el entorno, que garantice la irrefutabilidad e integridad de las ubicaciones de los usuarios, es una estrategia que ya se ha considerado previamente en una gran cantidad de entornos sujetos a este tipo de servicios.

Este procedimiento, no obstante, abre gran número de cuestiones en relación a la seguridad y privacidad de los usuarios. Esto se debe a que, por un lado, es necesario garantizar la autenticidad de la prueba de localización vinculándola con la identidad del usuario. Por el otro, dicho vínculo puede revelar la localización del usuario a más partes del sistema de las requeridas. En los últimos años, numerosos trabajos se han hecho eco de esta situación y han presentado diferentes propuestas [5], [6], [7], [8], [9], [10] intentando abordar dicha problemática.

Las propuestas existentes pueden englobarse en dos grupos atendiendo a que entidad es la encargada de generar las pruebas de localización. En los trabajos [5], [6], [7] la generación de dichas pruebas es centralizada y recae sobre la infraestructura del sistema que actúa como autoridad de localización. Esto supone que las pruebas solo pueden generarse en puntos muy concretos lo cual limita su usabilidad.

En el segundo grupo, las propuestas [8], [9], [10] optan por un método de generación distribuida, las cuales se sustentan del entorno más cercano al usuario para generar pruebas de localización de forma privada. En este tipo de esquemas, los elementos cercanos, como por ejemplo otros usuarios, son los que, en conjunto, comprueban y generan las pruebas de localización. Aunque esta solución es más práctica que la centralizada, su naturaleza ocasiona nuevos problemas relacionados con la privacidad y la seguridad, ya que no se puede garantizar la honestidad de las entidades que generan las pruebas.

Para abordar esta problemática, en [8] se presenta un sistema basado en coartadas que permite que los móviles ubicados en una determinada área, llamados testigos, generen conjuntamente pruebas de localización de un dispositivo cercano. La identidad del usuario se mantiene oculta por medio de esquemas de compromiso criptográficos y solo es revelada en el momento de verificar la prueba. En [9] se presenta el sistema APPLAUS, que propone el mismo principio en la generación de pruebas de localización, pero usando pseudónimos que cambian periódicamente para conservar la privacidad de los usuarios. Además aborda problemas de confabulaciones entre usuarios y múltiples niveles de granularidad en la ubicación que no se consideraban en [8]. Mejoras sobre la detección de confabulaciones entre usuarios es lo que reivindica [10]. En este trabajo los autores presentan el sistema STAMP, al que integran un protocolo de límite de distancia con el fin de detectar usuarios que intenten generar pruebas remotamente.

Todos los sistemas anteriormente mencionados parten de la premisa que el usuario es el interesado en probar su localización, ya que en el entorno donde se integren dichos sistemas el usuario será recompensado de alguna forma al demostrar donde se encuentra físicamente. Ante este escenario, situaciones como intentar pasar inadvertido o suplantar a otro usuario ante la infraestructura del sistema, como sucede en un sistema de control de acceso a una LEZ, carecen de sentido y, por tanto, ninguno de los sistemas mencionados plantea o resuelve dichas cuestiones.

III. Modelo del Sistema

Nuestra propuesta plantea un nuevo sistema generador de pruebas basadas en el entorno cercano que permite detectar comportamientos propios de un ataque de "relay" a partir de la información que estas contienen. Aunque el esquema presentado en este artículo está orientado hacia un ámbito vehicular, como por ejemplo las LEZs, nuestro sistema generador de pruebas puede ser extrapolado a cualquier escenario donde se requiera controlar el acceso a un área restringida.

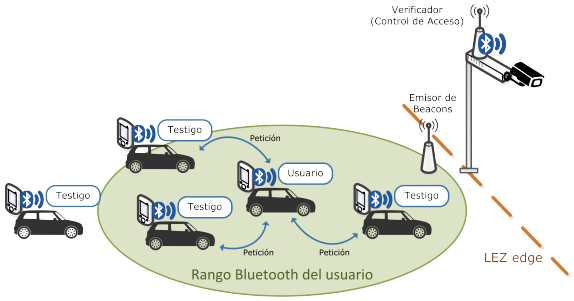


Figura 2. Arquitectura del sistema

La Figura 2 muestra un esquema general del sistema propuesto en este artículo. Se asume que los vehículos, tanto los que solicitan como los que generan pruebas, tienen integrado un dispositivo capaz de comunicarse inalámbricamente (p.e. Bluetooth o WIFI), de obtener su ubicación y conectarse a Internet en momentos puntuales.

Cuando un vehículo se acerca al Control de Acceso (CdA) de una zona restringida, una etiqueta BLE despierta al dispositivo encargado de ejecutar el protocolo. En ese momento se inicia la construcción de sus pruebas solicitando al resto de usuarios, dentro del alcance de su dispositivo de comunicación inalámbrica, la generación de una prueba. Para ello establece una comunicación segura con los usuarios dentro de su alcance para que individualmente acrediten su posición y sus tiempos de respuesta durante las comunicaciones. Cada uno de estos usuarios actúa como testigo, comprueba si la información que ha recibido es coherente con la suya y, en caso afirmativo, responde con una prueba criptográfica que lo corrobora. El conjunto de pruebas recibidas de todos los testigos integran la coartada del usuario con la que es posible demostrar su ubicación y que no está transmitiendo sus credenciales mediante la confabulación con otros usuarios.

En el momento de acceder a la zona restringida, el usuario transmite esta coartada al CdA que actúa como verificador. El CdA comprueba la validez de la coartada del usuario verificando cada una de las pruebas que le han generado el resto de usuarios. Si todas las pruebas son válidas, se asume que la coartada es válida y que el usuario está siendo honesto. En caso contrario, el CdA tomará las medidas pertinentes para revocar el anonimato del usuario en cuestión.

Durante este proceso, gracias a la generación basada en el entorno y en base a la información contenida en las coartadas, el CdA es capaz de generar temporalmente un mapa de los pseudónimos de usuarios que se encuentran en sus cercanías. Este mapa no solo determina las interacciones entre pseudónimos sinó que, debido a que las pruebas contienen información sobre ubicación y tiempo, también determina la distancia física y el tiempo de respuesta entre ellos.

Teniendo en cuenta que para perpetrar un ataque de "relay" hace falta un atacante que actúe de baliza y retransmita las señales Bluetooth hasta el usuario a suplantar, el CdA, gracias al mapa generado, calcula el umbral del tiempo de respuesta a partir del cual el sobrecoste de comunicación indica que hay una retransmisión de paquetes propia de un ataque de "relay". En base a esto también toma las medidas pertinentes para revocar la privacidad de los usuarios implicados.

IV. Descripción del Protocolo

En esta sección en primer lugar se describe de forma breve los actores, y a continuación las fases o operaciones que se pueden dar en el sistema.

IV-A. Actores

Nuestro sistema está compuesto por cuatro entidades: i) demostrador, ii) testigos, iii) Verificador y iv) Autoridad Certificadora.

- El demostrador (D) es un usuario del sistema (conductor) que dispone de un dispositivo, como una OBU, con GPS integrado que obtiene pruebas de su entorno cercano para demostrar su ubicación y tiempos de respuesta con dicho entorno a una tercera parte. En el escenario que proponemos D sería cualquier conductor que accede a una LEZ o zona de acceso restringido.
- El testigo (T) es otro usuario del sistema, con un dispositivo que integra GPS, que genera y avala pruebas de un D próximo físicamente cuando este lo solicite.
- El verificador (V) es la entidad que verifica las coartadas que presentan los Ds. En nuestro escenario, V sería la infraestructura que controla el acceso a la LEZ y que, en el momento en que un vehículo acede a la zona restringida, verifica que usuario no está transfiriendo sus credenciales en tiempo real por medio de un ataque de "relay".
- La Autoridad certificadora (AC) es una institución dedicada a generar y gestionar los materiales criptográficos a todas las entidades implicadas, p.e. claves públicas/privadas, certificados y pseudónimos.

IV-B. Fases

Las fases u operaciones que se realizan en el sistema propuesto son las siguientes: i) Inicialización; ii) Generación de pruebas; iii) Verificación; y iv) Control del Fraude.

IV-B1. Inicialización: Durante la fase de inicialización la AC genera y provee al resto de entidades que participan en el protocolo de los elementos criptográficos requeridos.

- Cada V con identificador ID_V consigue una clave privada sk_V, una clave pública pk_V y su correspondiente certificado $Cert_V^{AC}$.
- El dispositivo de cada usuario, independientemente del rol que desempeña como D o T, establece una conexión segura con la AC y recibe un pseudónimo PS_D, una clave privada sk_D, una clave pública pk_D y su correspondiente certificado $Cert_D^{AC}$. Dicho certificado solo contiene PS_D en el campo CommonName para evitar la identificación del usuario. Cada cierto intervalo

de tiempo, definido por el usuario, vuelve a repetirse este proceso a fin de generar un nuevo pseudónimo para el usuario. Una vez el usuario ha recibido sus nuevas claves y certificados, las anteriores son revocadas.

IV-B2. Generación de pruebas: La fase de generación de pruebas comienza en el momento en que un usuario requiere que su entorno cercano confirme su ubicación. El primer paso para iniciar el proceso es obtener un token del entorno que V emite periódicamente. Para ello V realiza las siguientes operaciones:

- Genera un token donde $V_{token} = \{ID_V, t_0, t_f\}$. Donde ID_V es el identificador de V, t_0 y t_f son timestamps que determinan el periodo de validez del token.
- Firma digitalmente el token $Sig^V(V_{token})$
- Emite V_{token} y $Sig^V(V_{token})$ periódicamente hasta la generación de un nuevo token.

Cuando un usuario D capta una emisión de V_{token} puede iniciar la construcción de sus pruebas basadas en entorno, para ello realiza el siguiente proceso:

- Verifica la firma $Sig^V(V_{token})$ y comprueba que el token no haya expirado.
- Enmascara su posición inicial $PosI_D$ en $\rho = Hash(PosI_D)$, donde $Hash$ es una función criptográfica de digest considerada segura.
- Construye una petición de generación de pruebas $D_{req} = \{V_{token}, \rho, t_1\}$. Donde t_1 es el timestamp del momento en que se genera la petición.
- Firma digitalmente los datos a enviar $Sig^D(D_{req})$.
- Hace un broadcast de D_{req} y $Sig^D(D_{req})$.

Cada T que recibe la petición D_{req} puede generar una prueba para D. Para ello realiza las siguientes operaciones:

- Verifica la firma digital de $Sig^D(D_{req})$ contenido en D_{req}.
- Verifica que los timestamps en D_{req} son consistentes. Para ello verifica que $t_0 < t_1 < t_2 < t_f$, donde t_2 es el timestamp del instante actual. Si la verificación no es correcta el proceso se detiene y T no responde a D.
- En caso afirmativo, genera una prueba $T_{res} = \{D_{req}, Pos_T, t_2\}$, donde Pos_T es la posición actual de T, y la firma digitalmente $Sig^T(T_{res})$.
- Envía T_{res} y $Sig^T(T_{res})$ a D.

Para cada prueba que recibe D, comprueba y almacena T_{res} realizando los siguientes pasos:

- Verifica la firma digital $Sig^T(T_{res})$.
- Verifica si la información contenida en T_{res} es coherente. Para ello comprueba que el D_{req} recibido coincide con el enviado, que Pos_T esté en un área cercana la posición actual de D $PosF_D$ y que $t_0 < t_1 < t_2 < t_3 < t_f$. Siendo t_3 la marca temporal actual. En caso que alguna de estas verificaciones falle, D desecha la prueba.
- Si las anteriores verificaciones son correctas, D almacena, junto al resto de pruebas derivadas de la misma petición D_{req}, la prueba recibida como $PR_i = \{PosI_D, PosF_D, Pos_T, t_2, t_3, Sig^T(T_{res})\}$.

IV-B3. Verificación: Cuando D se encuentra dentro del radio de comunicación de V, establece una comunicación segura e inicia la transmisión de las pruebas obtenidas. Para ello D realiza las siguientes operaciones:

- Recupera todas las pruebas PR_i a una petición D_{req}.
- Construye la prueba $D_{proof} = \{D_{req}, PR_1|...|PR_i\}$ y la firma digitalmente $Sig^D(D_{proof})$.
- Envía a V D_{proof} y su firma $Sig^D(D_{proof})$.

Cuando V recibe una coartada de un D comprueba la validez de la misma realizando los siguientes pasos:

- Verifica la firma de $Sig^D(D_{proof})$.
- Verifica si V_{token} se corresponde con el que estaba emitiendo durante t_0 y no ha expirado.
- Verifica la validez de cada prueba PR_i contenida en D_{proof}:
  - Verifica la firma $Sig^T(T_{res})$ con la información contenida en PR_i y D_{req}.
  - Comprueba que el pseudónimo PS_T contenido en el certificado de T es único en la actual D_{proof}
  - Calcula $Hash(PosI_D)$ y verifica que es igual a ρ contenido en D_{req}.
  - Comprueba que $PosI_D$, $PosF_D$ y Pos_T están contenidos en una área cercana a V.
  - Comprueba que los tiempos son coherentes y cumplen $t_0 < t_1 < t_2 < t_3 < t_f$.
- Si las comprobaciones son correctas se acepta PR_i como prueba válida. En caso contrario se descarta.

Si un determinado número de pruebas pasan estas verificaciones, la coartada D_{proof} se considera válida y se almacena. El umbral que determina cuantas pruebas son necesarias es dinámico y lo determina V en función del tráfico o de la frecuencia de accesos en ese momento.

IV-B4. Control del Fraude: A partir de las coartadas D_{proof} válidas almacenadas, V construye un mapa de interacciones entre pseudónimos. Por cada prueba PR_i contenida en D_{proof} realiza los siguientes pasos:

- Añade una arista en el mapa entre los pseudónimos de D PS_D y T PS_T.
- Calcula las distancias entre D y T usando $PosI_D$, $PosF_D$ y Pos_T.
- Calcula el tiempo de respuesta entre D y T a partir de t_1, t_2 y t_3. En base a esto, comprueba si el tiempo de respuesta supera la media de tiempo para generar una prueba para la distancia física entre D y T.
- Si el tiempo supera un umbral sobre la media se añade una incidencia en la arista del mapa entre PS_D y PS_T.

Periódicamente, V comprueba las relaciones entre los diferentes pseudónimos en el mapa. En el momento que un PS concentra una cierta cantidad de incidencias en una proporción elevada de sus relaciones en el mapa se le puede acusar de fraudulento. De la misma forma, también verifica que cada PS tenga más de una arista con otros pseudónimos, es decir, que haya generado pruebas como T además de presentar su D_{proof}. Si se detecta un PS egoísta que no ha generado pruebas para otros usuarios también se le acusa de fraude. En ambos casos V toma las medidas pertinentes e informa a AC para que revoque la privacidad del usuario implicado.

V. Estudio del Control de Fraude

En esta sección se definen los distintos ataques de "relay" que pueden darse en el entorno vehicular propuesto en este artículo. Para cada escenario se detalla cómo se defiende del

ataque el protocolo propuesto y cómo reacciona ante posibles contramedidas de los atacantes.

V-1. Caso normal: En la Figura 3 se muestra la generación de una prueba, por parte de un T honesto, para un usuario D en condiciones normales y sin que se produzca ningún ataque durante el proceso.

Figura 3. Caso normal

En este escenario, D, en el instante t_1, solicita la generación de una prueba. T recibe esta petición y entre los instantes t_2 y $t_{2'}$ genera la prueba con un timestamp t_2 y una posición Pos_T. Este tiempo de cómputo puede considerarse como negligible en comparación con el coste temporal de las comunicaciones Bluetooth de acuerdo con los estudios presentados en [4]. D completa la generación de la prueba añadiendo el instante de recepción t_3 y su posición final $PosF_D$.

Al no haber interferencias derivadas de actividades de usuarios deshonestos, el coste temporal de la generación de una prueba se obtiene de $t_3 - t_1$. V realiza este cálculo para cada prueba contenida en las coartadas que envían los Ds, con el fin de mantener una media actualizada del tiempo que les cuesta a los Ts responder con una prueba T_{res} a una petición D_{req}. Esta media se calcula para distintos rangos de distancias entre T y D para tener en cuenta el impacto de la distancia en la potencia de señal de las comunicaciones inalámbricas. V puede realizar estos cálculos porque todas las pruebas contienen $PosI_D$, $PosF_D$ y Pos_T. Estas medias sirven como referencia para detectar sobrecostes de comunicación que puedan indicar que se está cometiendo fraude por medio de un ataque de relay. Se asume que este tipo de ataques son casos excepcionales y que este tipo de procedimiento representa un porcentaje muy bajo del total de las pruebas que se generan.

V-2. Suplantación de T: En el escenario ilustrado en la Figura 4, un atacante (AT) confabulado con un T deshonesto, en una ubicación arbitraria, generan una prueba por medio de un ataque de "relay" para la coartada de D. Dichos actores se describen a continuación:

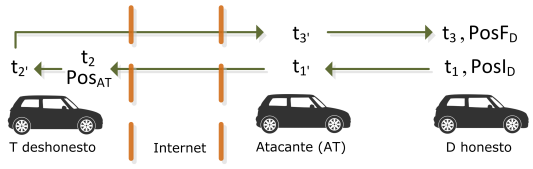


Figura 4. Ataque con suplantación de T

- D honesto con credenciales válidas que realiza una petición legítima a su entorno cercano para obtener una coartada.
- Un AT con la capacidad de hacer un ataque de "relay" para suplantar a un T deshonesto con el cual está confabulado. AT no tiene por qué ser usuario del sistema ni disponer de credenciales.
- Un T deshonesto, confabulado por con el atacante, que dispone de credenciales válidas.

Como puede apreciarse en la Figura 4, D en el instante t_1 emite una petición D_{req} para generar una coartada. AT, que retransmite las señales Bluetooth hacia T, recibe esta petición en el instante $t_{1'}$ y la envía a T que la recibe en t_2. En ese instante, T genera la prueba utilizando la posición Pos_{AT} del atacante en vez de la suya para evitar delatar el ataque de "relay". En $t_{2'}$ T inicia la transmisión de la prueba T_{res} hacia D repitiendo el proceso en sentido inverso y usando AT como baliza. D cierra la generación de la prueba añadiendo t_3 y su posición final $PosF_D$.

En el esquema descrito puede apreciarse que al ejecutar el ataque existe un sobrecoste debido a la comunicación extra entre AT y T. Este sobrecoste viene definido por $(t_2 - t_{1'}) + (t_{3'} - t_{2'})$ y representa el tiempo extra sobre el caso honesto descrito en la Sección V-1 que V valora para detectar un ataque de "relay".

Ante esta situación, T solo tiene maniobra para modificar el t_2 que incluye en T_{res}. No obstante, falsear dicho tiempo no tiene ninguna implicación en el tiempo de total de comunicación determinado por $t_3 - t_1$, ya que ambos valores son fijados por D que es honesto.

V-3. Suplantación de D: La Figura 5 muestra un escenario donde un D deshonesto, situado en una ubicación arbitraria, genera una coartada para un AT por medio de un ataque de "relay". Intervienen los siguientes actores:

- Varios Ts honestos que generan pruebas para un D cuando este lo solicita.
- Un AT con la capacidad de hacer un ataque de "relay" para suplantar a un D deshonesto con el cual está confabulado. AT no tiene por qué ser usuario del sistema.
- Un D deshonesto, confabulado con el atacante, que dispone de credenciales válidas.

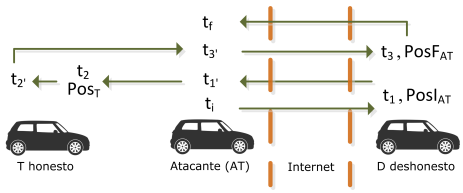


Figura 5. Ataque con suplantación de D

Como se puede ver en la Figura 5, el proceso lo inicia AT cuando en t_i envía V_{token} a D, quien se encuentra fuera del alcance de V y Ts. En t_1, D inicia el protocolo generando la petición D_{req}, usando la $PosI_{AT}$ para no delatar el ataque, y la envía a AT para que haga un broadcast a los Ts que le rodean. Cada T genera una prueba T_{res} en el instante t_2 y envía su respuesta a AT. AT de nuevo actúa de baliza y, en $t_{3'}$, envía a D las T_{res} que recibe junto con $PosF_{AT}$. Para cada prueba T_{res}, D introduce el tiempo de recepción t_3 y la nueva posición de AT. Finalmente construye D_{proof} con todas las pruebas recibidas. El proceso termina en t_f cuando AT recibe la D_{proof} y la envía a V haciéndose pasar por D.

En este escenario el ataque para suplantar a D también presenta un coste extra de comunicación. De acuerdo con la Figura 5, este sobrecoste se produce a causa de la comunicación extra entre AT y D y ahora se define por $(t_{1'} - t_1) + (t_3 - t_{3'})$. Hay que destacar que los costes de comunicación $(t_1 - t_i)$ y $(t_f - t_3)$ no pueden considerarse para detectar una suplantación de D, ya que $(t_1 - t_i)$ se produce antes de iniciar la construcción de la coartada y $(t_f - t_3)$ después de finalizarla. Aunque los tiempos que intervienen son distintos, el coste extra derivado del ataque en este caso es equivalente al de V-2. Esto es debido a que durante el ataque se envía la misma información entre D y T. Si bien es cierto que se genera una prueba por cada T, estos procesos son simultáneos e independientes.

Aunque queda patente que existe un sobrecoste de comunicación, D, en este caso, tiene capacidad para modificar los tiempos t_1 y t_3 que incluye en su D_{proof}. Si AT, con el que está confabulado, le facilita los tiempos t_i y $t_{3'}$, D puede calcular fácilmente el tiempo extra de comunicación y modificar los timestamps t_1 y t_3 para que el ataque de "relay" no pueda ser detectado por medio del tiempo de respuesta.

Aunque con este mecanismo D puede evitar ser detectado a partir de su D_{proof}, D seguirá interpretando el rol de T y generando pruebas T_{res} para otros usuarios. Ante esto, queda claro que las D_{proof} de los otros usuarios honestos acabaran delatando el ataque de "relay". Llegado a este punto, un usuario fraudulento puede evitar generar pruebas para otros usuarios a fin de no ser detectado. No obstante, el protocolo también considera fraudulento este comportamiento e igualmente emprende acciones contra el usuario.

Después de analizar los 3 escenarios posibles puede concluirse que, aunque los atacantes disponen de medios para intentar ocultar el sobrecoste de comunicación en ciertos casos, no tienen alternativa cuando generan pruebas como testigos para otros usuarios. Esto, sumado al hecho de que se considera fraudulentos a los usuarios egoístas que no colaboran en la generación de pruebas, garantiza que, con el protocolo propuesto, es posible registrar con seguridad el tiempo extra que consume un ataque de "relay".

VI. Privacidad

El sistema propuesto evita metodologías invasivas para la privacidad como el uso de cámaras para verificar la ubicación de los usuarios en el momento de acceder a una zona restringida. En vez de eso, el sistema hace uso del entorno cercano para generar pruebas firmadas criptográficamente que permitan verificar que sus credenciales no están siendo transferidas en tiempo real por medio de un ataque de "relay". Al usar esquemas de clave asimétrica que implican el uso de certificados que identifican al usuario, la privacidad de estos se protege por medio del uso de pseudónimos cuya relación con la identidad usuario solo la AC conoce. En el caso de los certificados, la AC los emite especificando únicamente el pseudónimo en el campo CommonName. Para evitar que un atacante o entidad del sistema pueda vincular un pseudónimo con su propietario, un usuario puede configurar cuando quiere solicitar un nuevo certificado y revocar el anterior; acto que conlleva la regeneración de sus claves. Finalmente, la inclusión de la posición de los usuarios en las pruebas que genera el sistema no tiene implicaciones para su privacidad. Esto se debe a que la generación de las mismas solo se produce durante los accesos y salidas de las zonas restringidas y por tanto no expone la ruta de los usuarios implicados.

VII. Conclusiones

El sistema propuesto permite generar pruebas, de forma privada, en colaboración con el entorno cercano para detectar ataques de "relay". La información sobre la posición y los tiempos de respuesta que contienen dichas pruebas garantiza que al menos a uno de los usuarios confabulados en el ataque pueda ser identificado y señalado como culpable. Más allá del entorno vehicular, la propuesta podría extrapolarse a cualquier entorno donde se detecte la necesidad de regular la entrada y salida de usuarios. Finalmente, el estudio de control de fraude realizado concluye que resulta imposible para un usuario deshonesto ocultar los indicios del ataque y que, por tanto, el sobrecoste que este produce en las comunicaciones siempre queda reflejado en las pruebas que se generan. Actualmente se está desarrollando el sistema que permita evaluar el verdadero alcance del sobrecoste derivado de un ataque de "relay" y poder verificar su aplicación en un entorno real. Debido al espacio disponible se ha dejado como trabajo futuro un estudio detallado de la seguridad y la privacidad que ofrece la propuesta.

Agradecimientos

This work was partially supported by the DGT under the project ISSUM SPIP2017-02250 and the Spanish Government under SmartGlacis TIN2014-57364-C2-R and Red de excelencia Consolider ARES TIN2015-70054-REDC projects. Some of the authors are members of the UNESCO Chair in Data Privacy, yet the views expressed in this paper neither necessarily reflect the position of the UNESCO nor commit with that organization.

Referencias

[1] World Health Organization. Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide. World Health Organization, 2006.

[2] G. Santos. Urban congestion charging: A comparison between London and Singapore. Transport Reviews 25 (5), 2005, pp. 511-534.

[3] J. Castella-Roca, M. Mut-Puigserver, M. Payeras-Capella, A. Viejo and C. Angles-Tafalla. Secure and Anonymous Vehicle Access Control System to Traffic-Restricted Urban Areas. In Computer Communication and Networks (ICCCN), 2017, 26th International Conference on IEEE.

[4] C. Angles-Tafalla, J. Castella-Roca, M. Mut-Puigserver, M. Payeras-Capella and A. Viejo. Secure and Privacy-Preserving Lightweight Access Control System for Low Emission Zones. To be accepted in Computer Networks.

[5] Saroiu, Stefan, and Alec Wolman. Enabling new mobile applications with location proofs. Proceedings of the 10th workshop on Mobile Computing Systems and Applications. ACM, 2009.

[6] Luo, W., and U. Hengartner. Veriplace: a privacy-aware location proof architecture. Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2010.

[7] Zhang, Y., Tan, C. C., Xu, F., Han, H., and Li, Q. Vproof: Lightweight privacy-preserving vehicle location proofs. IEEE Transactions on Vehicular Technology 64.1 (2015): 378-385.

[8] Davis, Benjamin, Hao Chen, and Matthew Franklin. "Privacy-preserving alibi systems."Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security. ACM, 2012.

[9] Zhu, Zhichao, and Guohong Cao. Toward privacy preserving and collusion resistance in a location proof updating system. IEEE Transactions on Mobile Computing 12.1 (2013): 51-64.

[10] Wang, X., Pande, A., Zhu, J., and Mohapatra, P. STAMP: enabling privacy-preserving location proofs for mobile users. IEEE/ACM transactions on networking 24.6 (2016): 3276-3289.

A Survey on Incident Reporting and Management Systems

Amna Qureshi, David Megías and Helena Rifà-Pous
Internet Interdisciplinary Institute (IN3),
Universitat Oberta de Catalunya (UOC),
CYBERCAT - Center for Cybersecurity Research of Catalonia
E-mail: {aqureshi,dmegias,hrifa}@uoc.edu

***Abstract*—Incident reporting and management systems are used to gather and process information about day-to-day occurring incidents such as road accidents, domestic violence, burglaries, harassment, etc. Nowadays, citizens can report such incidents through different means such as calling national emergency lines via telephones or mobile phones, or sending reports via Internet through web-browsers, or smartphones-based applications. A quick and appropriate response to the incidents can reduce both societal costs and fatality rates. Many smartphone-based applications and web-platforms exist that allow a person to send an image, a video, or map-based incident related reports to alert the concerned authority. A few reporting platforms and systems are also available to report the incidents anonymously to the relevant authorities to protect the eyewitnesses or victims. Our main contribution is the literature review of the incident reporting and management systems (based on smartphones and web-based) in terms of functionality, implementation details, technology, performance, security, and anonymity.**

***Keywords*—Incident Reporting, Anonymity, Efficiency, Smart phones.**

I. Introduction

The involvement of citizens in incident reporting has been in existence for a long time, e.g., almost all major cities have emergency service lines where citizens can report incidents such as road accidents, domestic abuse, fire, robbery/theft, shooting, hostage situation, etc. The emergency service number is typically a two (15 [1]) or a three-digit (911 [2], 112 [3]) number, that provides dispatch and communication support services for police, fire, ambulance, and related services to the caller. On average, it takes an emergency operator at least two to three minutes to collect the necessary information in order to respond to the caller [2]. A delay in collecting correct information about the incident can result in delayed incident response. Many traffic-car accidents result in deaths in developing countries every year due to the poor emergency response and reporting, e.g., according to the statistics of the India's department of highway and street transport, there were around 0.5 million road accidents in year 2000, which resulted in about 0.52 million injuries and more than 0.13 million deaths [4]. The most obvious reason for a person's death during traffic accidents is unavailability of the first aid provision, which is due to the delay in the information of the accident being reached to the ambulance or to the medics. Thus, response time is crucial for the timely delivery of emergency medical services to accident victims and is expected to have an impact on fatalities. Similarly in other life-endangering situations such as shooting, burglary, etc., a timely response from the concerned authorities could make a large difference in victims' survival rates. Therefore, communication plays an important role at the time of critical situations, where the loss of a few seconds can mean the difference between life and death.

Every year, about three hundred thousand people who call the emergency services cannot describe their location, because they may not know the exact location of the incident's occurrence [5]. In critical situations, knowing the exact location information in latitude and longitude of the incident reporter can help incident management authorities to react quickly and save lives. Today's mobile and smart devices are able to provide emergency services with accurate caller location via an SMS or data channel by using built-in Global Navigation Satellite System (GNSS) or Wi-Fi capabilities. Given the multitude of sensors (accelerometer, microphone, and high resolution cameras) and Internet connectivity present in modern smartphones, these devices can help in reporting accurate geo-coordinates of the incident location, and the extent and nature of the incident to the relevant authorities.

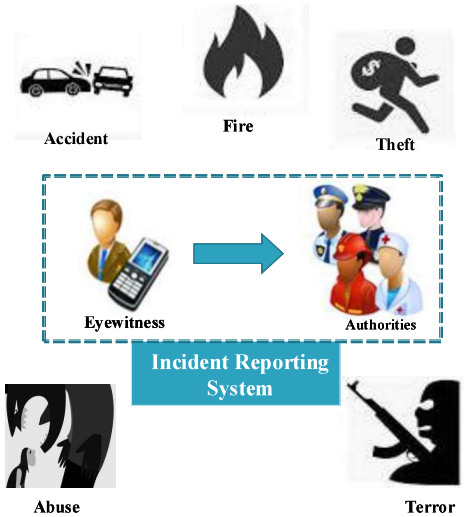


Fig. 1. Generic incident reporting system.

The increasing popularity of smartphones and the techno-

logical advancements in mobile communication and information systems have the most significant impact in the growth of the web-based and mobile incident management systems (Fig. 1). These systems or platforms provide citizens with the means to report incidents by specifying, for example, the location (such as selecting the location on a map), sending a precise location identification (such as using Global Positioning System (GPS)), or simply providing a proof of the incident (such as a photo) to the emergency responders, who, in return, respond quickly to the situation. For example, Elerts [6] is an effective and robust public safety communication platform based on smartphones. Using this application, the registered user witnessing the incident site can take a photo of it, which is automatically uploaded to a website, and other Elerts user gets the alert regarding such incident along with the location and the photo. In the wake of smartphones boom, social media platforms such as Facebook and Twitter, etc., are used at ever increasing level. The social media users use these platforms as communication tools to connect with one another, engage with news content, and share information, e.g., when an emergency strikes, people in the affected area start posting about it on social media from their mobile phones. By combining crowdsourced information from mobile devices as well as social media data, responsible officials can prepare for and respond to emergencies or incidents faster than ever before. Many incident reporting and management systems use crowdsourced information from the public to gain valuable forms of information from sources on the ground in and around the emergency or incident.

Also, many incident web-based and mobile reporting systems exist that allow citizens to communicate anonymously and confidentially any event or situation. For example, Fig. 2 presents a mobile-based app, Violens [7] that allows the eyewitnesses to take anonymous audio or video clips of an abusive incident, which are directly sent to the requisite authorities with a time and location stamp.

Fig. 2. Violens [7]: An anonymous reporting application.

Incidents such as road accidents, hate crime, harassment, and domestic abuse are critical issues that are often not reported because the witnesses do not want to be identified or reveal their specific location for personal reasons, or they are concerned about personal safety, or they fear the possibility of being considered as suspects of a crime. For example, in a recent road accidents survey in India [8], it was revealed that around 74 percent witnesses hesitate to report road accidents or help the victim due to legal hassles and fear of police harassment. Repeated questioning by the police, multiple summons from courts of law and even prosecution for unintentional accidental deaths prevent the witness from extending a helping hand to people in need, and reporting the accident to the relevant authorities. Thus, the systems should be designed in such a way that if an eyewitness desires anonymity, he/she can report the incident to the authority without any fear.

In this paper, we provide an overview of existing web-based and smartphone-based platforms and applications for incident reporting in terms of their features, benefits and implementation details. Also, these platforms and applications are compared in terms of reliability, usability, performance, security, and anonymity.

The rest of the paper is organized as follows. In Section II, we provide a brief overview of the existing incident reporting applications. Section III provides a comparative analysis in terms of performance and anonymity of the selected systems. Lastly, in Section IV, we present the conclusions of this work.

II. Incident Reporting and Management Systems

In this section, we highlight the features and benefits of the existing incident reporting and management applications. Also, details of a few recently proposed smartphone-based state-of-the-art incident reporting systems are presented. Although many emergency management systems exist in scientific literature, we include only the recent incident reporting systems that fulfill either one or all of the following requirements: (1) location-aware system that handles both reporting and management of the incident; (2) guarantee privacy-preservation of the eyewitnesses; and (3) provide cost-effective functionality in a simulated incident scenario.

1) Elerts [6] is a mobile emergency response communication application for reporting incidents all over USA. It is centralized and available on the market for each area such as airports, universities, health care, and corporate. Using the Elerts app, citizens can quickly report accidents, robberies, or any other incident by taking snapshots, which are automatically uploaded to the central website, that sends alerts and warnings to the respective service officers, and other users of Elerts. The main features of Elerts are bi-directional communication (once client or user gets warning or alert message on his/her cellphone, he/she can also reply back with other detail or elaborate information regarding the incident. The rescue groups or the service providers recommend ways to get out of this kind of situation), reports (the users of Elerts can send direct report or response to other users. A report could contain incident description, report type, photo/video and GPS location), photos (Elerts allows users to share live snapshots or

images with personal security or any other user), GPS and mapping (using Elerts, users can utilize advanced technology like Google maps to locate the emergency place. It can suggest or advice others to address the issues with the details of locations), mobile-platform support (Elerts works with iPhone and Android smartphones, and provides a simple interface in the form of a convenient drop-down menu to indicate report type, such as: robbery, medical emergency or suspicious activity).
The main benefits of Elerts are Elerts Attention Engine (a technology that transforms video surveillance cameras into active crime interception tools), National Incident Management System (NIMS) Compliant (Elerts provides the policies for how groups should plan, organize, train, and equip themselves in the event of a national security threat or disater), and integration (ELERTS enhances public safety by integrating the power of crowd-sourced incident reporting with video surveillance, access control and emergency mass notification systems).

2) A web-based platform in Brazil called WikiCrimes [9], allows users to register online information about criminal events in a specific geo-graphic location via a map. Since 2008, there have been over 200,000 crimes logged in this system [10]. Recently, the WikiCrimes mobile app for Android-based smartphones has been launched, which has more or less the same features as offered in desktop application.
The main features of WikiCrimes are accessibility (anyone can access WikiCrimes to add or consult information from geo-referenced database), common area of interaction (information about criminal incidents that go unreported by the police is shared in the community. Citizens can add links to videos, newspapers, photos or any other document to aid in credibility), violence hotspots tracking (allows users to keep track of hotspots and at-risk neighborhoods to make decisions in situations such as visiting an unknown neighborhood), report authenticity (WikiCrimes employs a user registration and a confirmation process to ensure the authenticity of the reported crime), incident confirmation (the reporting user should identify at least one person other than him/her, who can confirm that the registered incident is true), registration (the users need to register their name and valid email address for reporting the incident) and web-platform support (typical Web2.0 application).
The main benefits of WikiCrimes are false report detection (data mining algorithms aid in identifying irregularities and anomalies that can indicate false reports) and transparency (information is public and transparent).

3) AlertUs [11] is a mobile application that offers a panic button feature to the users, who can quickly and easily send incident reports to emergency managers.
The main features of AlertUs are geo-tagged incident reporting (users can report incidents with the user's current location, send a customized message in the message box, or select from multiple incident types with quick-action buttons via the panic button to deliver information to emergency management), custom branding (users can custom brand the home screen and notification details), public safety number (users can list safety numbers to easily contact emergency personnel or services), simple interface (AlertUs has a clean, intuitive interface that makes incident reporting easy and simple), and mobile-platform (users can easily download and install the AlertUs from iTunes [12] or Google Play [13] store).
The main benefits of AlertUs are auditing (detailed real-time reports track who receives/acknowledges each message and the current status of each client) and alerts notification (users instantly receive emergency alerts when alert notification is activated).

4) Ushahidi for crisis response [14] has recently launched a map-based mobile application that allows any citizen to gather distributed data via a text message (SMS), a tweet, an email or web forms, and visualize it on the map. Web-based Ushahidi has been used globally many times over the last years to facilitate the emergency management services in situations where there was a little or no support at all from the government or concerned authorities, e.g. for tracing events around Gaza strip. Ushahidi deployments are comprehensive software tools that allow organizations to build custom surveys, import data from third-party services, share it publicly in a map or timeline, and triage response, e.g., SafeCity [15] uses Ushahidi to crowd-source personal stories of sexual harassment and abuse in public spaces.
The main features of Ushahidi are data collection (receive reports from many sources, SMS, email, and Twitter), custom surveys (supports multiple data types with custom forms), multiple maps (provides map tiles including street and satellite provided by Open Street Maps, and MapQuest), filters and saved searches (crowd-sourced information can be filtered, tracked over time, and refined), report (Ushahidi saves data that contributors collect with it, including detailed GPS location data, and sends the report to the appropriate Ushahidi deployment), and mobile-platform support (available for both Android and iOS).
The Ushahidi app provides protection of sensitive information by introducing field level privacy controls, i.e. restricts visibility of specific fields in the form, to ensure that answers to those fields will only be accessible to certain people within the organization that has deployed Ushahidi at its end.

5) Aspire News [16] is a potentially life-saving app that allows victims of abuse or a person who knows about a person who is in an abusive relationship to call for help at the touch of a button. Aspire News is voted as one of the two "best apps" for domestic violence prevention in 2016.
The main features of Aspire News are quick escape button (a feature that the user can use to quickly change the screen to news if someone was to look over their

shoulder), trusted contacts (the user has more flexibility to choose a trusted contact. These contacts do not have to be individuals who are already in the phone's contact list), alert notification (trusted contacts who are added within Aspire News can be alerted with or without the user's location information), help section (24-hour hotlines and resources on domestic violence), and mobile platform support (Aspire News is available for both iPhones and Android-based smartphones).
The main benefit is an inconspicuous nature of the application. Aspire News appear as a news app, that contains summaries of top world, sports, and entertainment news. This feature is helpful to some survivors who want to use a personal safety app, but are concerned that the abuser may find the app on the phone.

6) SPEARS: Smart Phone Emergency and Accident Reporting System (SPEARS) [17] is a location-aware system that allows users of an online social network (Facebook or Twitter) in Thailand to quickly report emergency and accidents to the agencies such as police, fire department and hospitals, via a smartphone.
The main features of SPEARS are user registration (users must be identified by phone numbers and names before reporting an emergency or incident), quick response (Google map is used to display shortest routes of the relevant authorities to the occurred emergency in the quickest time), shortest route calculation (Dijkstra's algorithm is used to help finding the shortest path routing), GPS (the agencies can retrieve the current location of the emergency via GPS and send immediate help to the users involved in an emergency situation), user display (it shows locations of the nearest relevant agencies where the nearest incidents are reported), agency display (it displays locations where reported emergencies are received) and mobile-platform (only works in Android-based smartphones).
The main benefit of SPEARS is its integration with mobile social media such as Facebook and Twitter, that allows Thai users who are members of these social networks to report emergency.
7) STOPit [18] communities equip every citizen with a modern, anonymous, and safe way to share information directly with authorities about risks to public safety.
The main features of STOPit are two-way communication (anonymous communication between the user and the responsible officials. Using STOPit Messenger, the officials can respond to the report to ask questions, which can be replied by the user, while remaining completely anonymous), incident reporting (users can report inappropriate conduct or safety concerns such as harassment, bullying, ethics or compliance violations, weapons possession, hazing, safety hazards, threats, assault, or illegal activity by sending a report that may include text and photos or video), notifications (officials send users notifications such as updates or alerts), incident management (officials are equipped with a smart, easy backend system (anonymous messenger) to save time and conduct effective and efficient investigations) and mobile-platform support (available in IOS and Android devices).
STOPit is simple, fast and powerful tool which empowers individuals to protect themselves and others- all at the touch of a button.
8) A location-based emergency reporting system based on the Android operating system is proposed in [19] that allows the eyewitnesses to timely notify the authorized emergency authority about the incident while preserving their anonymity.
The main features of the system are eyewitness anonymity (the system employs an online social network in such a way that the eyewitness can create a dynamic and location-based social group to report the emergency information to the emergency management service (EMS) without revealing his/her real identity or any other personal information), revocable privacy (eyewitness's anonymity is preserved until he/she is found responsible of false reporting by the EMS), rewards and punishments (a game-theoretic design inspired by the co-privacy (co-utility) approach [20], is used to encourage legitimate information and discourage false reporting), GPS and mapping (users can utilize latest and advanced technology like Google maps to send the exact coordinates of the emergency place in the report), and report (incident description, GPS location, anonymous credentials, group ID).
The system provides security, k-anonymity, false report prevention, and cost-effective functionality.

III. Comparative Analysis

In this section, the presented systems are compared in terms of generic features that are desirable by the users of any smartphone-based incident reporting and management application or system. These generic features include platform support (Android, IOS, Web 2.0 and desktop versions), availability (free to download from Google Play, or iTunes store), usability (user friendly interface, issue-free experience), languages (single or multiple), performance (quick report submission and response, cost-effective), and location tracking (GPS or triangulation). Abuse of the emergency service providers, i.e. fake or false emergency calls is an expensive problem because it drains out the resources of the emergency services and, also, prevent people in need to get urgent assistance from the service providers. Thus, all the incident reporting and management systems desire reliability (verification of the reported incident, prevention of fake reports, report auditing/analysis post incident) so that people in need always get a top priority from emergency service providers. While users use incident reporting applications to report incidents that either involve them directly or they are eyewitness to a particular incident by sharing their personally identifiable information, they desire data security (report confidentiality, data integrity). In some applications that allow users to report hate crimes, domestic

TABLE I
COMPARISON OF PRESENTED INCIDENT REPORTING AND MANAGEMENT SYSTEMS

Group	Incident Reporting Systems/Applications	Platform	Availability	Reliability	Usability	Languages	Performance	Location Tracking	Data Security	Anonymity
Public	Elerts [6]	- Android 4.4 + - IOS 9.0 + - Desktop app for Windows and MAC	Yes	Poor	Medium	Multiple	Good	GPS	Yes	N/A
	AlertUs [11]	- Android 4.1 + - IOS 9.0 + - Desktop app for Windows and MAC	Yes	Good	Medium	English	Medium	GPS	Yes	N/A
	Ushahidi [14]	- Android 2.2 + - IOS 9.0 + - Desktop (Web 2.0)	Yes	Best	Good	English	Good	GPS	Yes	N/A
	SPEARS [17]	- Android	No	No	N/A	Thai	Good	GPS	No	N/A
Private	WikiCrimes [9]	- Android 1.5 + - Desktop (Web 2.0)	Yes	Good	Good	Portuguese	Poor	GPS	No	No
	Aspire News [16]	- Android 4.1 + - IOS 8.0 +	Yes	Poor	Poor	English	Medium	GPS	No	Yes
	STOPit [18]	- Android 4.0 + - IOS 8.0 +	Yes	Poor	Poor	Multiple	Poor	GPS	Yes	Yes
	Privacy-preserving EMS [19]	- Android	No	Yes	N/A	English	Good	GPS/ Triangulation	Yes	Yes

abuse, etc., anonymity (anonymous report submission) at the user end is a desirable feature because the victim or the eyewitness does not want to be identified.

In Table I, the applications are grouped into two categories. The first group (Public) contains applications in which users do not desire anonymity while reporting the incidents such as road accidents, wildfire, etc. The second category (Private) consists of the applications in which users require anonymity while reporting the incidents such as hate crimes, etc. In this table, a cell contains "No" when the corresponding attribute is not guaranteed by the system, and "N/A" when the corresponding attribute is not applicable to the system.

From Table I, following observations can be made:

- In terms of availability, almost all the presented systems can be downloaded freely either from Google Play [13] or iTunes [12] store except for SPEARS [17] and privacy-preserving EMS [19].
- Systems such as WikiCrimes [9], AlertUs [11], Ushahidi [14], and privacy preserving EMS [19] that require user registration, or report verification/confirmation, or fake report detection or prevention, provide good reliability as compared to other systems (Elerts [6], Aspire News [16], and STOPit [18]) that do not have such preventive measures.
- For usability, the user-based rating of each mobile application is obtained from the Google Play [13] and iTunes [12] stores. Depending on the average of users' ratings, we have scored each system in the following categories: Good (rating: 3.6-5.0), Medium (ratings: 2.6.0-3.5), and Poor (ratings: below 2.5).
- Almost all applications are available in English language except for WikiCrimes [9] (Portuguese) and SPEARS [17] (Thai).
- For performance evaluation, the user experience of each mobile application is obtained from the Google Play and iTunes store. Depending on the users' experiences, we have scored each system in the following categories: Good (efficient reporting, quick emergency/incident response), Medium (varies between quick and slow response from the officials), and Poor (slow or no emergency response at all). For the systems that have not been tested in real-world incident scenarios (SPEARS [17] and privacy-preserving EMS [19]), the experimental results are considered. SPEARS [17] calculates the shortest path on the Google map from the emergency authorities to the point of incident using Dijkstra's algorithm so as to help the authorities to reach the emergency location in time. However, this distance calculation is limited to the information within 30 kilometers of the point where the user is located and has reported an incident. In privacy-preserving EMS [19], the computationally expensive part of the system, i.e. threshold discernible ring signatures used to provide accountability and conditional anonymity is implemented on an Android emulator to compute the efficiency of the proposed system. Based on the experimental results, the proposed scheme can be implemented as a real-world mobile application to provide timely response to the affectees of the critical incident.
- In almost all the applications, location coordinates, i.e. longitude and latitude coordinates are calculated by GPS except in privacy-preserving EMS system [19] that in addition to GPS performs triangulation in case of poor GPS satellite coverage.
- The applications such as Elerts [6], AlertUS [11], Ushahidi [14], and STOPit [18] provide partial report confidentiality, i.e. these systems provide personally identifiable information such as name, email address, phone number and location to the service providers (police, emergency services, law enforcement authorities, etc. for the purpose of managing incident-related services such as incident verification), who may reconstruct, use and distribute user's personal information to the interested advertising parties. Privacy-preserving EMS [19] only shares the location with the relevant emergency authorities for report verification and dispatching relevant emergency service providers. Elerts [6] and Privacy-preserving EMS [19] provide data confidentiality, i.e. protects user

information against unauthorized access or modification. WikiCrimes [9], Aspire News [16] and SPEARS [17] do not provide data security.

- Elerts [6], WikiCrimes [9], AlertUS [11] and SPEARS [17] do not provide anonymous reporting, whereas Ushahidi [14], Aspire News [16] and STOPit [18] provide it. Privacy-preserving EMS [19] provides conditional anonymity, i.e. anonymity of an honest witness is preserved unless he/she is found responsible of false reporting by the emergency authority.

In the view of Table I and the above discussion, it can be observed that in the first category, Ushahidi [14] outperforms all other available applications in terms of performance, reliability, usability, and data security. In the second category, the only system that offers good performance, reliability, data security, and revocable privacy is privacy-preserving EMS system [19], but it has not been tested under real conditions.

IV. Conclusion

This paper presents a comprehensive survey and analysis of existing incident reporting and management systems. For this purpose, we have discussed the systems in terms of their features, implementation details, and benefits. Then, all the systems are compared in terms of features, functionality, security, and anonymity property. With the recent trend of growing number of applications based on crowd-sourcing, and increasing level of user participation in social media networks during disasters or incidents occurrence, governments and the concerned emergency response officials are able to respond effectively and efficiently to incidents. However, the main challenge lies in the effective and efficient response from the concerned incident response officials, while detecting/preventing false incident reports, and providing data security and revocable privacy options in the system to enable eyewitnesses/reporters to report the incident without fear of his/her safety.

Acknowledgments

This work was partly funded by the INCIBEC-2015-02491 “Ayudas para la excelencia de los equipos de investigación avanzada en ciberseguridad” and TIN2014-57364-C2-2-R “SMARTGLACIS”.

References

[1] I. Police. (2000) Rescue 15. http://islamabadpolice.gov.pk/Rescue15. Accessed on June 27, 2018.

[2] Tapshield, “How mobile and cloud technologies are reducing emergency response times,” Tapshield, White Paper, 2014, accessed on June 27, 2018.

[3] E. Committe, “False emergency calls,” http://www.eena.org/ressource/static/files/2011_03_15_3.1.2.fc_v1.0.pdf, European Emergency Number Association (EENA), Operations Document 3.1.2, 2011.

[4] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, “A survey of mobile phone sensing,” *Comm. Mag.*, vol. 48, no. 9, pp. 140–150, 2010.

[5] E. C. Committe, “Implementation of the european emergency number 112 - results of the tenth data-gathering round,” http://ec.europa.eu/newsroom/document.cfm?doc_id=43923, European Union Communication Committe, Operations Document, 2017.

[6] E. Corporation. (2010) Elerts. http://elerts.com/company/. Accessed on June 27, 2018.

[7] A. Aziz. (2015) Violens. https://www.microsoft.com/en-mv/store/p/violens/9nblggh6jv9w. Accessed on June 27, 2018.

[8] P. Jha. (2016) If no-one helps you after a car crash in india, this is why. http://www.bbc.com/news/magazine-36446652. Accessed on June 27, 2018.

[9] V. Furtado, L. Ayres, M. de Oliveira, E. Vasconcelos, C. Caminha, J. D’Orleans, and M. Belchior, “Collective intelligence in law enforcement - the wikicrimes system,” *Inf. Sci.*, vol. 180, pp. 4–17, 2010.

[10] A. Moore, M. de Oliveira, C. Caminha, V. Furtado, V. Basso, and L. Ayres, *Applying Geovisual Analytics to Volunteered Crime Data*. Springer Berlin Heidelberg, 2013, pp. 185–209.

[11] Alertus Technologies LLC. (2008) Alertus. https://www.alertus.com/alertusproducts/app. Accessed on June 27, 2018.

[12] Apple. itunes store. https://www.apple.com/lae/itunes/download/. Accessed on June 27, 2018.

[13] Google. Google play store. https://play.google.com/store?hl=en. Accessed on June 27, 2018.

[14] Ushahidi. (2017) Ushahidi. https://www.ushahidi.com/apps. Accessed on June 27, 2018.

[15] Red Dot Foundation Global. Safecity. http://safecity.in/. Accessed on June 27, 2018.

[16] The Robin McGraw Revelation Foundation. Aspire news. https://www.whengeorgiasmiled.org/aspire-news-app/#. Accessed on June 27, 2018.

[17] C. Namahoot and M. Bruckner, “Spears: Smart phone emergency and accident reporting system using social network service and dijkstra’s algorithm on android,” in *Mobile and Wireless Technology 2015*, ser. Lecture Notes in Electrical Engineering, vol. 310. Springer Berlin Heidelberg, 2015, pp. 173–182.

[18] STOPit. (2017) Stopit: Anonymous reporting and incident management. https://play.google.com/store/apps/details?id=com.stopitcyberbully.mobile&hl=en. Accessed on June 27, 2018.

[19] A. Qureshi, H. Rifà-Pous, and D. Megías, “Enabling collaborative privacy in user-generated emergency reports,” in *Privacy in Statistical Databases*. Springer International Publishing, 2016, pp. 255–271.

[20] J. Domingo-Ferrer, “Coprivacy: an introduction to the theory and applications of cooperative privacy,” *SORT-Statistics and Operations Research Transactions*, vol. 35, no. special issue: Privacy in statistical databases, pp. 25–40, 2011.

Sistema de comunicaciones seguras para el personal de servicios de emergencias

Alexandra Rivero-García, Candelaria Hernández-Goya, Iván Santos-González y Pino Caballero-Gil
Departamento de Ingeniería Informática y de Sistemas
Universidad de La laguna
{ariverog, mchgoya, jsantosg, pcaballe}@ull.edu.es

***Resumen*—Debido a la creciente concentración de personas en entornos urbanos, las ciudades son cada día escenario de múltiples eventos con diferentes niveles de riesgo. Puesto que los servicios de emergencia tienen el papel protagonista en lo referente a la protección de los ciudadanos, la comunicación eficiente y segura entre el personal de estos servicios debe ser una prioridad. En el caso de producirse eventos que impliquen la concentración de un gran número de ciudadanos en una determinada zona puede surgir un problema de saturación de las infraestructuras de comunicación tradicionales. Este documento propone un modelo de comunicación de bajo costo específicamente diseñado para solventar estas situaciones de manera que la información transferida quede también protegida. El sistema de comunicación descrito incluye un esquema de firma y cifrado basado en identidad que funciona teniendo en cuenta la ubicación y la identificación pública de los trabajadores de los servicios de emergencia con el fin de proporcionar integridad, confidencialidad, autenticación y no repudio en un solo paso y de forma eficiente.**

***Palabras clave*—Criptografía basada en Identidad; Dispositivos móviles; Android; Comunicaciones**

I. Introducción

En este trabajo se presenta un sistema de comunicaciones de bajo coste pensado para ser utilizado en situaciones de emergencia cuando se produce saturación en los canales de comunicación tradicionales por el acceso masivo a la red por parte de usuarios. Este es un esquema proactivo que genera un segundo canal de comunicación, que no depende de la infraestructura ni de la nube, mediante dispositivos móviles.

El objetivo es proporcionar un medio de comunicación adecuado a los servicios de emergencia (policía, bomberos, médicos, etc.) en caso de acontecimientos en los que se produzcan alertas o actividades específicas que requieran su participación (riesgo de inundación, marcha de protesta, concierto, incendio, etc.). En el esquema propuesto se localizan los eventos y se le asigna a cada un de ellos un color dependiendo del riesgo que supone el evento para los servicios de emergencia y para el correcto funcionamiento de la ciudad. Cuando se declara un evento, se debe asignar personal de diferentes servicios de emergencia a esa zona.

Normalmente, el primer paso cuando se preven este tipo de eventos es la asignación de diferentes trabajadores de los servicios de emergencia a áreas específicas para preservar la seguridad civil. En el sistema propuesto, este paso también se utiliza para generar y compartir previamente cierta información. Por eso se incluye un chat genérico de eventos, en el que los trabajadores pueden participar a través del teléfono móvil. En el caso de que se detecte congestión en la red, se declara el modo de emergencia. En este modo específico, las comunicaciones se realizarán directamente, en modo persona-a-persona (P2P, del inglés peer-to-peer) a través de sus dispositivos móviles y sin herramientas adicionales.

Un segundo objetivo de este trabajo es garantizar la seguridad de los datos compartidos, por lo que se utiliza un esquema de firma y cifrado basado en identidad, conocido como IBSC (Identity-Based SignCryption) basado en la ubicación y las identificaciones públicas de los trabajadores de los servicios de emergencia. Con este criptosistema todos los mensajes compartidos serán firmados y cifrados.

La mayoría de las comunicaciones que se desarrollan actualmente dentro de las secciones de emergencia se realizan por radiofrecuencia. En la propuesta presentada se ha desarrollado un prototipo basado en Nearby. Esta es la tecnología que propone Google para la comunicación de la Internet de las cosas (IoT, del inglés Internet of Things). Se basa en el envío de datos entre dispositivos cercanos, mediante el envío de mensajes, creación de buffers para el envío de datos en tiempo real e incluso para intercambiar mensajes modo beacon. Con esta propuesta Android se abstrae un nivel más en la implementación de las herramientas y decide automáticamente cual es la mejor tecnología para la comunicación Nearby en cada momento dependiendo de la latencia de las comunicaciones y tecnologías que tengamos en cada momento. Las comunicación con Nearby en nuestro prototipo se realizan a través de dos tecnologías diferentes disponibles en smartphones: Bluetooth Low Energy (BLE) [1] y Wi-Fi Direct [2].

Seguidamente procedemos a detallar la estructura del artículo. La sección II incluye una breve reseña de las publicaciones que justifican la novedad y adecuación del sistema propuesto. La sección III trata de los fundamentos de la Criptografía Basada en Identidad ya que la propuesta la utiliza para garantizar la confidencialidad, integridad y autenticidad de la información transferida. Una vista general del sistema se presenta en IV. La sección V trata de los detalles del esquema de comunicación y su descripción formal. La sección VI termina el documento resumiendo las principales conclusiones y contribuciones de la propuesta.

II. Estado del Arte

Generalmente, las comunicaciones actualmente desplegadas/utilizadas para emergencias se llevan a cabo por radiofrecuencia. Es una solución deficiente, ya que los trabajadores sólo pueden compartir audio en una frecuencia específica, no se permite agrupar y compartir datos de medios de forma efectiva. Se han propuesto múltiples soluciones basadas en Wi-Fi Direct y smartphones como en [3] donde se explica el potencial de Wi-Fi Direct en la implementación de sistemas P2P móviles. Este trabajo incluye algunos ejemplos del uso de Wi-Fi Direct compartir mensajes de texto, difundir información, etc. Utilizan un dispositivo intermedio para redes P2P para distribuir tablas hash para buscar pares. En [4], los autores explican la posibilidad de generar redes oportunistas sobre Wi-Fi Direct mediante el estudio de la latencia en la capa de enlace. Es una extensión de [5], donde se generaron múltiples grupos. Este trabajo presenta mediciones reales que confirman la idoneidad del Wi-Fi Direct para sistemas P2P.

Hay algo en común en todos estos sistemas, no se proponen elementos de seguridad. Nuestro enfoque se diferencia de otros en que tiene en cuenta la distribución, asignación y ubicación de los recursos humanos en múltiples eventos. Además, la seguridad de la información se aborda a lo largo de todo el desarrollo del sistema propuesto. En [6], los autores proponen el uso de Wi-Fi Direct como un sistema de comunicación alternativo para situaciones de emergencia, pero no para la comunicación entre servicios de emergencia. El objetivo principal de la aplicación desarrollada allí es compartir la geolocalización de las personas cuando están aisladas, sin señal y en dificultades.

III. Preliminares

A continuación, describimos las herramientas matemáticas básicas utilizadas. Por un lado el esquema de cifrado basado en identidad y por otro lado una de las bases dentro de este cifrado elegido, los grupos bilineares. El correcto entendimiento de la definición de los grupos bilineales son fundamentales para poder comprender los emparejamientos bilineares utilizados en el esquema de cifrado.

Sistema de firma y cifrado basado en identidad. En los sistemas basados en identidad (IBE, de su acrónimo del inglés Identity Based Cryptosystems) [7], el objetivo principal es el uso de cualquier cadena como clave pública válida. Estos esquemas a menudo utilizan como identificadores direcciones de correo electrónico, números de seguridad social, identificadores personales, etc. A partir de esta idea surgieron algunas modificaciones como los esquemas de firma y cifrado basados en identidad (IBSC) donde el objetivo principal es obtener integridad, confidencialidad, autenticación y no repudio en un solo paso y de manera eficiente. La idea básica es una composición de un esquema de cifrado con un esquema de firma [8]. Todos estos sistemas de cifrado basados en identidad cuentan con unos pasos básicos: *Inicialización*, para la generación de los parámetros del gestor central; *Extracción*, en este paso se generan las claves privadas de cada usuario basado en su identificador; *Cifrado con Firma*, este es el paso utilizado para la creación de mensaje y su posterior envío, como su propio nombre indica todos los mensajes enviados van cifrados y firmados; finalmente, para poder obtener los mensajes de forma adecuada se utiliza el paso de *Descifrado y verificación de firma* que se realiza en el lado del receptor del mensaje.

Grupos Bilineares. Se consideran dos grupos cíclicos $(\mathbb{G}, +)$ y $(\mathbb{G}_\mathbb{T}, \cdot)$ del mismo orden q. P es un generador de $\mathbb{G}$ y $\hat{e}$ se define como un emparejamiento bilineal de la forma $\hat{e} : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_\mathbb{T}$. Este satisface las condiciones siguientes: $Bilinearidad$, $\forall P, Q \in \mathbb{G}$ y $\forall a, b \in \mathbb{Z}$, $\hat{e}(aP, bQ) = \hat{e}(P, Q)^{ab}$. *No degeneración*, $\exists P_1, P_2 \in \mathbb{G}$ que $\hat{e}(P_1, P_2) \neq 1$. Esto significa que si P es un generador de $\mathbb{G}$, entonces $\hat{e}(P, P)$ es un generador de $\mathbb{G}_\mathbb{T}$. $Computabilidad$, existe un algoritmo para calcular $\hat{e}(P, Q), \forall P, Q \in \mathbb{G}$ [9].

IV. Sistema propuesto

Este documento presenta un sistema alternativo de comunicaciones para servicios de emergencia a través de teléfonos móviles en diferentes escenarios. Para evitar la congestión de la red, los usuarios deben compartir cierta información pública cuando se activa el modo de emergencia. Esta información es la identificación del usuario y se comparte a través de BLE usando el modo de baliza (véase Figura 1) para identificar a cada participante. Cada persona tiene una lista de identificadores (IDs) correspondientes a personas cercanas. Cuando un trabajador de servicios de emergencia encuentra a un compañero que comparte una identificación, almacena esta identificación. Más adelante, las identificaciones pueden utilizarse para comunicaciones entre pares.

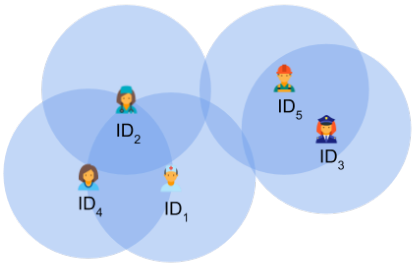


Figura 1. Compartición de identidades modo beacon.

El sistema soporta dos modos de comunicación diferentes: P2P y broadcast. En el primer modo, dos personas pueden intercambiar información directa y bilateralmente. En el modo de transmisión, todo el mundo en el área afectada puede recibir la notificación y simplemente haciendo clic en ella, pueden chatear para ayudar a su colega. El personal de emergencias puede compartir texto, imágenes, audio e incluso vídeos. La seguridad de los datos compartidos es crucial en el sistema propuesto. Así, en ambos modos de comunicación se utiliza un esquema de cifrado y firma basado en identidad (IBSC).

Para participar en las comunicaciones, cada trabajador del servicio de emergencia asignado a un área/evento también

debe tener alguna información. En primer lugar, desde el sistema central se generan eventos. Después, el controlador debe asignar diferentes tipos de personal a esa zona. También se proporciona información específica que permite la participación del personal en el sistema de chat, (véase Figura 2).

Figura 2. Panel de administración web.

A cada evento se le asigna un identificador único generado aleatoriamente, así como la información para su geolocalización. Cuando un miembro del personal de emergencia es asignado a un evento, el sistema genera credenciales específicas y las claves para compartir datos. Los usuarios pueden obtener su propia ubicación, la ubicación de sus pares y el área del evento desde la aplicación móvil (véase Figura 3).

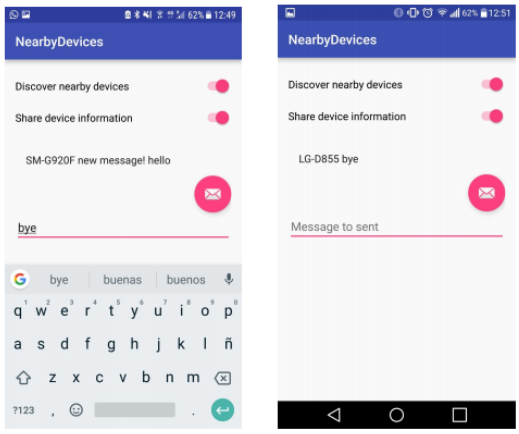


Figura 3. Aplicación móvil.

V. Esquema de comunicaciones de emergencia

Como se mencionó en la sección anterior, se utiliza un esquema de firma basado en la identificación. Este enfoque ofrece la ventaja de simplificar la gestión al no tener que definir una infraestructura de clave pública. Se eligió este tipo de esquema por su baja complejidad computacional, eficiencia en términos de memoria y usabilidad. El personal del servicio de emergencia puede compartir información con una sola persona en modo peer-to-peer, a través de una señalización basada en ID, y con múltiples usuarios en modo broadcast a través de un esquema de señalización multi-receptor basado en ID. En el esquema propuesto, el servidor central soporta el Generador de Clave Privada (PKG). Es una parte crucial de la propuesta, ya que es el servicio encargado de generar las claves privadas del personal de emergencias. El esquema de firma y cifrado utilizado es una combinación del esquema de IBSC propuesto en [10] y un esquema IBSC para múltiples receptores [11].

A continuación se presentan algunos elementos matemáticos básicos y notación utilizados en la descripción del sistema. También se necesitan varias funciones hash: $H_1 : \{0,1\}^* \rightarrow G^*, H_2 : \{0,1\}^* \rightarrow \mathbb{Z}_q^*, H_3 : \mathbb{Z}_q^* \rightarrow \{0,1\}^n, H_4 : \{0,1\}^n \rightarrow \{0,1\}^{|m|}, H_5 : G \times G \times \{0,1\}^n \times Z_q^* \times Z_q^* \times ... \times Z_q^* \rightarrow Z_q^*$, en donde n es el tamaño del mensaje. $x \xleftarrow{r} S$ hace referencia a la selección de un elemento x al azar del grupo S, $x \leftarrow y$ denota la asignación del valor y a la variable x,y finalmente, $||$ es utilizado para la representación de la concatenación de dos elementos. Los pasos necesarios para el esquema de firma y cifrado son los siguientes:

Inicialización: En este primer paso, el servidor inicializa los parámetros para generar sus propias claves: clave pública maestra (mpk) y clave secreta maestra (msk). Para lograrlo, se necesitan algunos datos privados: $k \in \mathbb{Z}$ para generar un número entero primo q, dos grupos $\mathbb{G}$ y $\mathbb{G}_\mathbb{T}$ de orden q y un emparejamiento bilineal $\hat{e} : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_\mathbb{T}$. Después, $P \in \mathbb{G}$ se selecciona aleatoriamente y se definen las funciones hash a utilizar. Finalmente, las claves del servidor se generan $msk \xleftarrow{r} \mathbb{Z}_q^*$ y $mpk \leftarrow msk \cdot P$.

Generación de claves maestras: Este algoritmo tiene por objetivo la identificación del personal. La clave pública $Q_{ID}enG$ se genera a través de una función hash aplicada al ID, $Q_{ID} \leftarrow H_1(ID)$. Por otro lado, la clave privada S_{ID}, usada para comunicaciones con el servidor $S_{ID}.enG$, se calcula teniendo en cuenta msk, $S_{ID} \leftarrow msk \cdot Q_{ID}$. Hay que tener en cuenta que el intercambio de claves entre el servidor y el personal se realiza utilizando el cifrado SNOW3G [12]. bajo la clave de sesión obtenida a través de un Eliptic Curve Diffie-Hellman (ECDH)[13].

Generación de claves de eventos: Este paso es obligatorio, y se realiza cuando se genera un nuevo evento en el sistema. Cada uno de los eventos generados tiene un identificador único, $ID_e \xleftarrow{r} \mathbb{Z}_q^*$ y algunas coordenadas de ubicación, lat y lon. En este paso, la clave pública para este evento $Q_{IDe} \in G$ se genera como: $Q_{IDe} \leftarrow H_1(ID||ID_e||lat||lon)$. Luego, la clave secreta para este evento $S_{IDe} \leftarrow msk \cdot Q_{IDe}$.

Cifrado simple: Todos los mensajes $m \in \{0,1\}^n$ serán cifrados y firmados. La clave pública del receptor se genera teniendo en cuenta su identificación y los datos precompartidos (ID_e, lat y lon): $Q_{IDe_b} \leftarrow H_1(ID_b||ID_e||lat||lon)$. Luego, se desarrollan algunas operaciones dando como resultado σ (una t-upla de tres componentes: c, T, U). T se genera como $x \xleftarrow{r} \mathbb{Z}_q^*$ y $T \leftarrow x \cdot P$. Luego, se realiza la firma usando la clave privada del remitente (S_{IDe_a}) que está en U, la cual se obtiene de la siguiente manera $r \leftarrow H_2(T||m)$, $W \leftarrow x \cdot mpk$ y $U \leftarrow r \cdot S_{IDe_a} + W$. Finalmente, el mensaje cifrado está en c y se genera como se muestra a continuación $y \leftarrow \hat{e}(W, Q_{IDe_b})$,

$k \leftarrow H_3(y)$, $c \leftarrow k \oplus m$.

Cifrado para múltiples receptores: En este modo de transmisión hay n receptores, por lo que el remitente se identifica con IDe_a y los receptores con $IDe_1, IDe_2, ..., IDe_n$. Todos los mensajes de difusión $m \in \{0,1\}^n$ serán cifrados y firmados. La clave pública del remitente se genera de la forma: $Q_{IDe_a} \leftarrow H_1(ID_a||ID_e||lat||lon)$. Luego se desarrollan algunas operaciones dando como resultado σ (una tupla de componentes: $c, T, U, V, W, X, a_0, ...a_n - 1$). Después el remitente selecciona al azar algunos números $r \xleftarrow{r} \mathbb{Z}_q^*$, $r' \xleftarrow{r} \mathbb{Z}_q^*$, $s \xleftarrow{r} \mathbb{Z}_q^*$ y $p \xleftarrow{r} \mathbb{Z}_q^*$ para luego operar $T \leftarrow r \cdot Q_{IDe_a}$, $U \leftarrow r \cdot P$, $X \leftarrow r' \cdot T$, $J \leftarrow r' \cdot mpk$. Las claves públicas de los receptores se generan teniendo en cuenta todas las identificaciones $ID_1, ID_2, ..., ID_n$, de la siguiente manera: $f(x) = \prod_{i=0}^{n}(x - v_i) + p(mod q) = a_0 + a_1 x + ... + a_{n-1}x^{n-1} + x^n$ con $Qe_i \leftarrow H_1(ID_i||ID_e||lat||lon)$, $y_i \leftarrow \hat{e}(Q_i, J)$ y $v_i \leftarrow H_2(y_i)$. Luego se calcula $V \leftarrow s \cdot H(p)$, la clave k de la forma $k \leftarrow H(s)$ y el mensaje cifrado c como $c \leftarrow k \oplus m$. Finalmente, se genera un autenticador h como $h \leftarrow H_5(c, X, U, V, a_0, a_1, ..., a_{n-1})$ y $W \leftarrow (r' + h)r \cdot S_{ID_a}$.

Descifrado simple: En primer lugar, la clave pública del remitente se genera teniendo en cuenta IDe_a y la información precompartida de la forma: $Q_{ID_a} \leftarrow H_1(IDe_a||ID_e||lat||lon_e)$. Luego σ se divide y filtra (c, T, U). Si todo está bien, se devuelve el mensaje $m \in \{0,1\}^n$. De lo contrario, si hay algunos problemas se devuelve la firma o el cifrado de m, $\perp$. La verificación es:

$$\hat{e}(U, P) == \hat{e}(Q_{IDe_a}, mpk)^r \cdot \hat{e}(T, mpk)$$

Por lo tanto, el usuario calcula $y \leftarrow \hat{e}(S_{IDe_b}, T)$, $k \leftarrow y$, $m \leftarrow k \oplus c$ y $r \leftarrow H_2(T||m)$.

Descifrado para múltiples receptores: En este paso se realizan dos verificaciones, pero en primer lugar σ se analiza como $c, T, U, V, W, X, a_0, ...a_n - 1$ y $h \leftarrow H_5(c, X, U, V, a_0, a_1, ..., a_{n-1})$. La primera verificación es la verificación pública para comprobar que el texto cifrado es válido:

$$\hat{e}(W, P) == \hat{e}(X + hT, mpk)$$

De lo contrario, el texto cifrado se ha dañado o no es válido y se devuelve $\perp$. La segunda verificación es:

$$\hat{e}(W, Qe_i) == \hat{e}(X + hT, S_{IDe_i})$$

Es para verificar si el usuario ID_i es uno de los receptores elegidos por el remitente y si el texto cifrado es válido. De lo contrario, el receptor abandonará el proceso de descifrado y el sistema devolverá $\perp$.

Para generar el mensaje se realizan algunas operaciones: $y_i \leftarrow \hat{e}(S_{IDe_b}, U)$, $v_i \leftarrow H_2(y_i)$, $p \leftarrow f(v_i)$, $s \leftarrow V \oplus H_3(p)$, $k \leftarrow H_4(s)$ y $m \leftarrow k \oplus c$.

VI. Conclusiones

Nuestro sistema genera automáticamente los datos precompartidos en función del evento al que el trabajador ha sido asignado; el objetivo principal es que los trabajadores de emergencias puedan compartir información entre ellos cuando diferentes eventos saturan la red. El sistema tiene una herramienta de apoyo de emergencia para contactar a sus compañeros a través de un chat. Se utiliza un Signcryption basado en ID para proteger la integridad, confidencialidad, autenticación y no repudio en las comunicaciones. Específicamente, el servicio de emergencia puede compartir información con una sola persona en modo P2P a través de un esquema de firma y cifrado basado en ID y con múltiples usuarios en modo broadcast a través de un esquema de firma y cifrado basado en ID para múltiples receptores. Como trabajo futuro, se añadirán más funcionalidades al servidor, como estadísticas, chats privados basados en roles, etc. Se ha implementado un prototipo beta con Nearby que sólo está disponible en Android 8. La adición de LTE-Direct [14] depende del Native Development Kit (NDK), porque ahora mismo este código es privado.

Agradecimientos

Este trabajo ha sido financiado con TESIS2015010102, TESIS2015010106, RTC-2014-1648-8, TEC2014-54110-R, MTM-2015-69138-REDT y DIG02-INSITU.

Referencias

[1] N. K. Gupta, *Inside Bluetooth low energy*. Artech House, 2016.
[2] W. Shen, B. Yin, X. Cao, L. X. Cai, and Y. Cheng, "Secure device-to-device communications over wifi direct," *IEEE Network*, vol. 30, no. 5, pp. 4–9, 2016.
[3] R. Motta and J. Pasquale, "Wireless p2p: Problem or opportunity?" in *Proceedings of the Second IARIA Conference on Advances in P2P Systems*, 2010, pp. 32–37.
[4] M. Conti, F. Delmastro, G. Minutiello, and R. Paris, "Experimenting opportunistic networks with wifi direct," in *Wireless Days (WD), 2013 IFIP*. IEEE, 2013, pp. 1–6.
[5] D. Camps-Mur, A. Garcia-Saavedra, and P. Serrano, "Device-to-device communications with wi-fi direct: overview and experimentation," *IEEE wireless communications*, vol. 20, no. 3, pp. 96–104, 2013.
[6] I. Santos-González, A. Rivero-García, P. Caballero-Gil, and C. Hernández-Goya, "Alternative communication system for emergency situations." in *WEBIST (2)*, 2014, pp. 397–402.
[7] M. Joye and G. Neven, *Identity-based cryptography*. IOS press, 2009, vol. 2.
[8] X. Boyen, "Identity-based signcryption," in *Practical Signcryption*. Springer, 2010, pp. 195–216.
[9] J. Groth and A. Sahai, "Efficient non-interactive proof systems for bilinear groups," *Advances in Cryptology–EUROCRYPT 2008*, pp. 415–432, 2008.
[10] J. Malone-Lee, "Identity-based signcryption." *IACR Cryptology ePrint Archive*, vol. 2002, p. 98, 2002.
[11] S. S. D. Selvi, S. S. Vivek, R. Srinivasan, and P. R. Chandrasekaran, "An efficient identity-based signcryption scheme for multiple receivers," Cryptology ePrint Archive, Report 2008/341, 2008, https://eprint.iacr.org/2008/341.
[12] I. Santos-González, A. Rivero-García, P. Caballero-Gil, and C. Hernández-Goya, "Alternative communication system for emergency situations." in *WEBIST (2)*, 2014, pp. 397–402.
[13] J. W. Bos, J. A. Halderman, N. Heninger, J. Moore, M. Naehrig, and E. Wustrow, "Elliptic curve cryptography in practice," in *International Conference on Financial Cryptography and Data Security*. Springer, 2014, pp. 157–175.
[14] S. Sesia, I. Toufik, and M. Baker, *LTE: the UMTS long term evolution*. Wiley Online Library, 2009.

Criptografía adversaria usando deep learning. Limitaciones y oportunidades

Alfonso Muñoz Muñoz
Head of Cybersecurity Lab - Blueindico BBVA
Avenida de Burgos 16D, 28036 Madrid
alfonso@criptored.com

José Ignacio Escribano
Security Researcher - Blueindico BBVA
Avenida de Burgos 16D, 28036 Madrid
joseignacio.escribano@i4s.com

***Resumen*—En la última década numerosos investigadores han intentado aplicar los avances en algoritmos de inteligencia artificial al campo de la ciberseguridad, desde un punto de vista ofensivo y defensivo. Aunque también antigua, su aplicación a la criptografía en la actualidad está siendo revisitada con nuevas propuestas que afirman poder crear algoritmos criptográficos (criptografía adversaria) sin conocimiento previo y sin la supervisión de un humano (criptógrafo/criptoanalista). En este trabajo se recopila el avance en esta dirección y se inicia un análisis crítico para poder llegar a inferir si la criptografía adversaria usando deep learning es una propuesta criptográfica seria a considerar. La evaluación de sus características y de las redes neuronales en juego no es para nada trivial y los resultados empíricos de momento la alejan de una utilidad práctica en escenarios reales.**

***Index Terms*—criptografía, redes neuronales, deep learning, criptografía adversaria**

I. Introducción

En los últimos años, tanto el *Machine Learning* (ML), y en particular, como el *Deep Learning* (DL) se han convertido en potentes herramientas con las que obtener patrones sobre grandes volúmenes de datos y estos avances han sido aplicados a numerosas áreas del conocimiento, entre las que se encuentran la visión por computador [1–3], la detección y predicción de enfermedades [4–7], entre otras. Un área donde el machine learning ha cobrado especial relevancia es el de la ciberseguridad, ya que es posible aplicarlo tanto a técnicas defensivas como ofensivas. En cuanto a seguridad defensiva, se han aplicado a esteganografía [8–10], detección de anomalías en tráfico de red [11–13], detección y clasificación de malware [14–18] o detección de vulnerabilidades [19–22], entre otros. En el campo de la seguridad ofensiva, se han publicado numerosos ataques contra algoritmos de machine learning [23–25], ya que numerosos algoritmos de machine learning se han visto vulnerados mediante pequeños cambios imperceptibles en los datos de entrada, conocidos como *ejemplos adversarios* (Figura. 1).

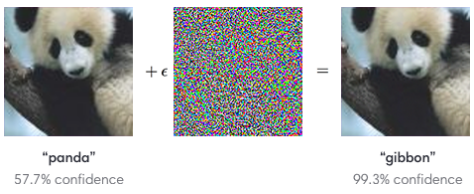


Figura 1. Ejemplo adversario sobre una imagen. Introduciendo una pequeña perturbación en una imagen es posible que un algoritmo de machine learning clasifique de forma incorrecta. En el ejemplo, un algoritmo de ML clasifica de forma correcta la imagen como un panda, pero al aplicarle cierto ruido el algoritmo lo clasifica de forma incorrecta como un gibón. Fuente: OpenAI[2].

Esto ha dado lugar a un área de investigación conocido como *Adversarial Machine Learning* [26], que surge de la necesidad de hacer los modelos de machine learning más robustos frente a ejemplos adversarios. Su objetivo es crear ejemplos adversarios (entradas con cierto nivel de ruido, indistinguibles del ojo humano, que hacen que los modelos de machine learning obtengan una respuesta incorrecta). Entre los métodos para generar ejemplos adversarios se encuentran [27–34]. Algunos de estos métodos se implementan en la herramienta Cleverhans disponible en GitHub[3].

I-A. Deep Learning

El machine learning es la rama de las ciencias de la computación que busca la creación de modelos, de forma que aprenda de forma autónoma a partir de un conjunto de datos. Se clasifica en tres tipos de aprendizaje: supervisado [35], no supervisado [36] y por refuerzo [37]. El objetivo del machine learning es minimizar una función de pérdida a partir de un conjunto de datos, que sirve para medir la bondad del modelo generado. El deep learning es un subconjunto del machine learning que trata de obtener representaciones complejas de los datos a través de distintas arquitecturas de redes neuronales. Las principales arquitecturas en la actualidad son:

- **Redes convolucionales:** consisten en una serie de capas a la que se aplican filtros convolucionales. Este tipo de red ha sido utilizada con éxito en procesamiento de imágenes.
- **Redes recurrentes:** consisten en capas conectadas en forma de grafo. Permiten el procesado de secuencias como texto o audio. Sus variantes más utilizadas son LTSM [38] y GRU [39].

I-A1. Generative Adversarial Networks: A parte de las anteriores, hay una nueva arquitectura llamada Generative Adversarial Network (GAN) [40] que consiste en dos redes neuronales que compiten entre sí, en un juego de ataque-defensa. Estas dos redes neuronales reciben el nombre de *Generador* (G) y discriminador (D) (Figura. 2). El generador, a partir de ruido y algunas variables, genera nuevos datos (sintéticos) que no se encuentran en el conjunto de datos original. El discriminador recibe este conjunto de datos junto con nuevos datos del generador y trata de distinguir entre los reales y los generados. Las dos redes neuronales compiten entre ellas, con el generador tratando de generar mejores datos "falsos", y el discriminador trata de distinguir mejor

[2]https://blog.openai.com/adversarial-example-research
[3]https://git.io/v5tzX

las muestras falsas de las reales. El juego acaba cuando se alcanza un equilibrio entre el generador y el discriminador en sus respectivas funciones de pérdida. La idea original de las redes *GAN* ha tenido gran impacto, haciendo que se hayan propuesto una gran cantidad de variaciones. Algunos de estos ejemplos se pueden ver en *The GAN Zoo*[4].

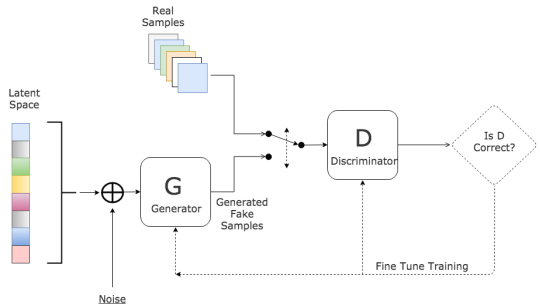


Figura 2. Estructura de una Generative Adversarial Network. Fuente: [41]

II. Estado del arte en Deep Learning y criptografía

El estudio de las redes neuronales para el cifrado y descifrado de datos, y criptoanálisis, conocido como *criptografía neuronal* (neural cryptography, en inglés) no es algo reciente, sino que sus principios tienen como base el artículo de *Sébastien Dourlens* [42], donde se aplican redes neuronales para el criptoanálisis del algoritmo DES. Más adelante, se presentó un sistema criptográfico basado en una red recurrente autoasociativa que presentaba algunas propiedades criptográficas que los sistemas clásicos no tenían [43]. En 2002, fue presentado un nuevo fenómeno entre redes neuronales y criptografía llamado *sincronización de redes neuronales* (*synchronization of neural networks*, en inglés), con la que podían generar claves efímeras para transmitir datos de forma segura [44]. Ese mismo año, se presenta un análisis detallado del sistema de sincronización de redes neuronales y cómo puede ser roto de tres formas distintas [45]. Más recientemente, se han hecho otras propuestas con distintas arquitecturas de redes neuronales para crear criptosistemas, de los cuales destacan dos: el primero presenta un criptosistema basado en una variación de la *Generative Adversarial Network* [46], que en las próximas secciones explicaremos más en profundidad, y el segundo, presenta un criptosistema basado en *Echo State Networks*, un caso muy sencillo de red recurrente que permite crear un criptosistema que cumple las propiedades de confusión y difusión, y que es aplicable a cualquier formato de datos que permitan una representación en forma de bytes [47]. Otro tema de interés creciente es el entrenar redes neuronales manteniendo la privacidad de los datos, de tal manera, que se pudiera delegar su entrenamiento en servicios en la nube como *Amazon Web Services*[5], *Google Cloud Platform*[6] o *Microsoft Azure*[7], entre otros, sin comprometer la privacidad de los datos. Este es un campo en estrecha relación con la criptografía homomórfica [48]. Se han propuesto distintas alternativas, quizás una de las más prometedoras sean las *CryptoNets* [49], una nueva arquitectura de redes neuronales que se pueden aplicar sobre datos cifrados, delegando el entrenamiento en la nube, sin la necesidad de descifrar los datos. Demuestran que usar este nuevo tipo de redes es posible, y lo prueban sobre la base de datos de imágenes del MNIST [50], consiguiendo una precisión del 99 % y pudiendo hacer 59 000 peticiones en una hora. Uno de los problemas de este tipo de redes es que no permiten mantener la privacidad de los datos sobre redes neuronales muy profundas. Para solucionar en parte este problema, que sigue abierto a día de hoy, se ha conseguido utilizar las *CryptoNets*, junto con *Batch Normalization* [51] con 6 capas no lineales. Con este enfoque, se obtiene una precisión sobre la base de datos del *MNIST* muy similar a sus versiones sin cifrar [52].

[4]https://git.io/vpemf
[5]https://aws.amazon.com
[6]https://cloud.google.com
[7]https://azure.microsoft.com

III. Caso de estudio: criptografía adversaria

En los últimos 5 años, el algoritmo que más relevancia ha tenido utilizando deep learning en el ámbito de la criptografía es [46] (Figura. 3).

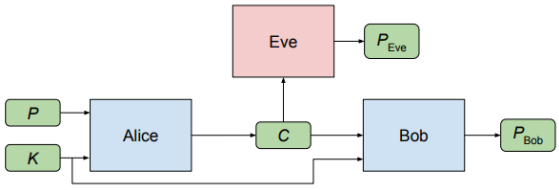


Figura 3. Esquema del criptosistema. Fuente: [46]

Una propuesta que sería capaz de crear algoritmos criptográficos sin conocimiento previo, lo cual podría suponer un hito en la defensa contra posibles atacantes. Conceptualmente, se podría llegar a crear algoritmos criptográficos de un solo uso con la robustez de un algoritmo creado por un humano, o al menos eso han supuesto muchos investigadores tras la publicación de ese trabajo. En nuestro artículo, se analizan las bondades e inconvenientes de esa propuesta desde un punto de vista empírico e iniciaremos un análisis de sus propiedades criptográficas que desmitificará su aplicación real con su definición actual.

La propuesta de los investigadores de Google [46] introduce un método de cifrado entre redes neuronales en presencia de un atacante, usando una variación de una *Generative Adversarial Network*, que intenta descifrar el texto plano original de la comunicación. Se puede realizar de forma simétrica o asimétrica. En nuestro caso, sólo consideraremos el caso simétrico (Figura. 3). En este esquema intervienen tres actores: *Alice*, *Bob* y *Eve*; todos ellos son redes neuronales. La entrada de *Alice* es un texto plano (P) y una clave (K), precompartida con *Bob*. Tanto P como K se representan en forma de bits en el rango [-1,1], considerando un bit 0 como -1 y un bit 1 como 1. Su objetivo es el de obtener un texto cifrado C (criptograma). En el caso de Bob, debe descifrar el criptograma C y obtener P (conociendo la clave K); y el de *Eve* es el mismo que Bob, salvo que no dispone de la clave K. Con este esquema, las tres redes neuronales "compiten" entre

ellas, de tal manera que *Alice* aprende a cifrar el mensaje P, Bob intenta descifrar el criptograma C (con el uso de la clave K), y por último, el de *Eve* es el descifrar de forma correcta sólo el 50 % de los bits. De esta forma se evita que invirtiendo los bits incorrectos por correctos se obtenga un mejor descifrado. Según esto, las arquitecturas de cada red neuronal planteadas son las siguientes:

- *Alice* tiene una entrada de 2N bits (los N primeros bits para el texto plano P y N bits para la clave K). Consta de 5 capas ocultas: una capa "fully connected" y cuatro capas convolucionales. Cada una de estas últimas cuatro capas ocultas tienen filtros [4,1,2], [2,2,4], [1,4,4] y [1,4,1], donde los valores son el tamaño de la ventana y, la profundidad de la entrada y la salida. Además, se le aplica a cada capa un *stride* de 1,2,1 y 1, respectivamente. En cuanto a las capas de activación, todas las capas hacen uso de la función sigmoide, salvo la última que usa una función tangente hiperbólica (tanh).[8]
- *Bob* y *Eve* son exactamente igual *Alice*, salvo que como entrada N bits en vez de 2N, que se corresponden con el criptograma C.

El sistema se entrena con 4 096 ejemplos de longitud de 2N bits (N bits del texto plano P y N bits de la clave K). La forma de generación de estos es a través de números aleatorios, tanto en la clave como en el texto plano. De esta forma se consigue que las redes neuronales no aprendan patrones específicos entre un texto plano o clave fija y el texto cifrado. En [46] se entrena el sistema con N=16, 32 y 64 bits. El código de los experimentos del artículo se encuentra disponible en GitHub[9]. Todo el código se encuentra escrito en Python[10] y requiere tener instalado Tensorflow [53] para su correcto funcionamiento.

IV. Análisis crítico del algoritmo

El sistema propuesto en [46] se apoya en una arquitectura GAN para que Alice y Bob alcancen un entrenamiento adecuado de sus redes neuronales que les permita cifrar y descifrar la información. Para que esto sea posible la presencia del atacante (Eve) y diversas funciones de pérdida son vitales. Eve medirá la diferencia entre el texto que es capaz de descifrar respecto a la entrada original de Alice. Bob medirá la diferencia entre el texto recuperado y el texto original introducido en la red neuronal de Alice. Utilizando cientos de miles de iteracciones en el proceso de entrenamiento y las funciones de pérdida, la capacidad de Eve de descifrado debería ser similar a tomar una decisión al azar y la red de Bob y Alice deben converger para poder funcionar una vez el sistema esté entrenado. Al menos todo esto en teoría, como se puede deducir de [46] esto no implica que una vez esté entrenado el sistema el atacante (red neuronal) no pueda descifrar ningún mensaje, dado que en realidad si podrá, este es un primer inconveniente reseñable. Para analizar diferentes cuestiones criptoanalíticas se ve necesario simular el criptosistema, utilizamos el código disponible en GitHub y un PC con una tarjeta gráfica NVIDIA GTX 1080, un procesador Intel Core i7-5960X a 3GHz y 128 GB de RAM. El tiempo de entrenamiento, usando GPU, para distintos valores de N se puede ver en la Tabla I, en el caso de nuestro análisis generalizamos su propuesta para tamaños de bloque y tamaño de clave similares a criptosistemas modernos, en nuestro caso, bloque y clave de 128 bits.

Tabla I
Tiempo de entrenamiento para distintos valores de N bits

Número de bits (P,K)	Tiempo (h)
16	2
32	4
64	6
128	11

Se puede observar que este criptosistema requiere una gran cantidad de tiempo para converger usando GPU. El tiempo de entrenamiento en un equipo convencional (sin GPU) se vería incrementado en varios órdenes de magnitud, por lo que a día de hoy, no parece un sistema que pueda ser utilizado por todo el mundo sin requerir una gran capacidad de cómputo. De igual manera, tampoco es escalable, la necesidad de computo es excesiva, para tamaño de bloque superiores a 128 bits. El código disponible en [46] no funciona en tiempo de evaluación (una vez entrenado). Esto quiere decir que no es posible evaluar la seguridad del sistema para múltiples textos y claves de entrada y sus correspondientes cifrados. Actualmente estamos analizando como variar el código para que esto sea posible. Las apreciaciones y análisis inicial que se justifica a continuación se infiere de la fase de entrenamiento.

Verificamos que los datos para entrenar el criptosistema es imprescindible que sean generados de forma aleatoria, pues, de otra manera, *Alice* sólo es capaz de generar textos cifrados con todos los bits 0, o todos los bits 1. Esto demuestra el hecho que se presupone en [46] y, es que, utilizando otra forma de generación de textos y claves que no sea de forma aleatoria hace que generen criptogramas triviales. En cuanto a los valores que cifra la red neuronal, se pueden ver en la Tabla II.

Se puede observar que entre los 25 valores más frecuentes de los bits cifrados se encuentran en torno a -0.348, la región lineal de la función tangente hiperbólica, la función de activación de la última capa convolucional de *Alice*. Esto parece indicar que el sistema se está entrenando de forma correcta evitando el problema de anulación del gradiente (vanishing gradient problem, en inglés) [54]. Con el sistema disponible para iniciar la fase de entrenamiento y que las tres redes neuronales evolucionen para realizar su trabajo adecuadamente en regla con sus funciones de pérdida (Alice, Bob y Eve) se decide medir alguna propiedad criptográfica que nos haga entrever si esta propuesta, además de razonablemente segura desde el punto de vista de Eve, lo es también de un criptoanalista humano.

Con el estado del código actual y los medios de computación disponibles se decide iniciar este camino analizando propiedades de aleatoriedad de los bits generados y propiedades de efecto avalancha (qué bits se modifican con mayor probabilidad).

1. **Propiedades estadísticas de los criptogramas generados por Alice**. Para la evaluación del sistema se

[8]Nótese que el rango de esta función es [-1,1], que se corresponde con el rango de entrada de las redes neuronales, haciendo que todo el esquema sea consistente.

[9]https://git.io/vx9Jr

[10]https://www.python.org

Tabla II
FRECUENCIA DE LOS 25 VALORES DE LOS BITS MÁS FRECUENTES

Valor de la red neuronal	Frecuencia
-0.34829	502776474
-0.34828	216508587
-0.34827	209106663
-0.34826	193375342
-0.34825	162464444
-0.34824	128640326
-0.34823	101358486
-0.34822	84499015
-0.34821	72750752
-0.34820	61695603
-0.34819	53539283
-0.34818	47248630
-0.34817	40715808
-0.34816	35536298
-0.34815	30915081
-0.34814	26694136
-0.34813	23212870
-0.34812	19937915
-0.34811	17616547
-0.34810	16071334
-0.34809	13828933
-0.34808	12300502
-0.34807	10892053
-0.34806	9681770

selecciona los parámetros definidos en entrenamiento en [46] pero con un tamaño de P y K de 128 bits. La fase de entrenamiento constará de 850 000 iteraciones, que se agrupan en 4 250 pasadas con 4 096 pares de P,K en cada una de ellas. En total, se han podido analizar 2 228 224 000 bits cifrados obtenidos de diferentes "rondas" del algoritmo.

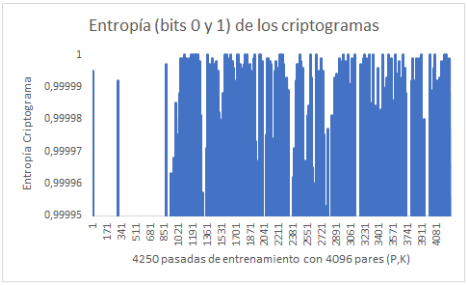


Figura 4. Entropía de criptogramas Alice

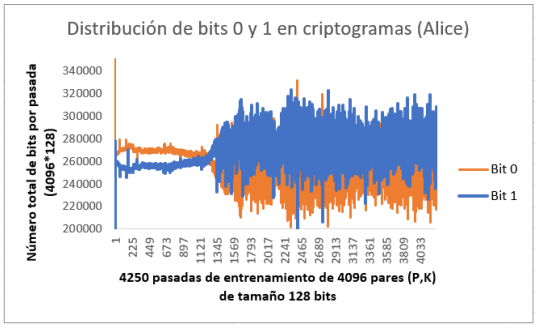


Figura 5. Distribución de bits 0 y 1 en criptogramas Alice

Para su evaluación se analizó sus valores con test estadísticos estándar utilizando herramientas habituales: NIST [55] o *dieharder* [56]. En general, la información cifrada generada no pasa ningún test de aleatoriedad del NIST lo cual no deja de ser interesante, ya que cabría esperar que el número de test estadísticos que se pasaran fueran mayor con un mayor número de iteraciones de entrenamiento del sistema. De las 4250 pasadas solo 44 (con sus criptogramas correspondientes) pasaron todos los test del NIST. Estos criptogramas no se encuentran al final de la fase de entrenamiento sino distribuidos de manera que todavía no somos capaces de razonar. En una iteracción puede que el criptograma tenga mejores propiedades aleatorias e iteracciones después puede que sean peores. Esta característica es común en sistemas de machine learning (mínimos y máximos locales) pero en su aplicación a la criptografía no deja de ser un problema evidente. Por tanto, por las pruebas realizadas, un criterio mínimo que se exigiría a un criptograma no se cumple (Figura. 4 y Figura. 5).

2. **Efecto avalancha. Modificación de bits en función de P y K**. El sistema planteado en [46] solo puede ser analizado en tiempo de entrenamiento, dado que como se justificó anteriormente en tiempo de evaluación no es posible introducir mensajes y claves de entrada de manera arbitraria, este hecho permitiría medir como afectan en el criptograma cambios específicos en posiciones de bits del mensaje y la clave de entrada. Por otro lado, en fase de entrenamiento los textos y clave de entrada tienen que ser aleatorios para que las redes neuronales se entrenen correctamente, este hecho introduce una serie de limitaciones a la hora de medir como afectan cambios en la entrada a la salida y cada cuantas iteracciones se produce el mayor cambio posible (efecto avalancha). Con estas limitaciones se decide analizar algún aspecto que al menos nos permita observar probabilidades o ciertas propiedades criptográficas que aprenda la red neuronal. Se decide analizar, dado un texto de entrada (aleatorio) y su correspondiente criptograma, el número total de bits modificados en cada ocasión y analizar cual es la probabilidad de que un cierto bit se modifique en el criptograma de tamaño 128 bits. Después de entrenar el sistema para que genere 17 408 000 criptogramas de tamaño 128 bits se observa que la probabilidad de modificación de los bits en el criptograma resultante no es equiprobable y esto no cambia aunque el sistema se entrene con más iteraciones. Nuestras pruebas indican que es más probable que sufran cambios los bits que se encuentran más a la izquierda que los que se encuentran más a la derecha en el criptograma generado. En concreto la probabilidad de modificación por bloques es: bits 1-16 un 69 %, bits 16-32 un 21 %, 32-48 un 0,2 %, 48-64 un 0,2 %, 64-80 un 0,2 %, 80-96 un 0,2 %, 96-112 un 1,2 % y 112-128 un 7,1 %. El valor medio de bits modificados por criptograma es 20,58, con un valor máximo de 127 (cambiamos todos los bits) y valor mínimo de 1. La varianza se calcula en 5,0179. Ante estos resultados cabe destacar buenas y malas noticias. La red neuronal infiere que la clave de entrada debe afectar a la modificación del mensaje de entrada y

lo consigue sin conocimiento previo. Lógicamente la manera en que afecta y las probabilidades observadas en los criptogramas resultantes permiten afirmar que estos no serían seguros desde el punto de vista de un criptoanalísta humano. Por otro lado, la red neuronal deduce al menos funciones criptográficas similares a un XOR (máscara de bits que permite modificar más de 1 bit en el criptograma) que aunque interesante carece de la seguridad requerida, por si sola, para un criptosistema moderno (Figura. 6).

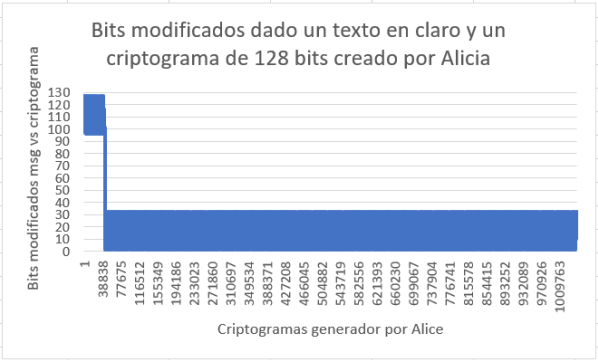


Figura 6. Modificación de bits en el criptograma generado por Alice respecto al mensaje en claro de entrada. Ejemplo de un millón de criptogramas

V. Conclusiones y trabajo futuro

En el presente artículo se destaca el interés actual en la ciencia del machine learning y deep learning en su aplicación a diversas áreas y, en especial, al ámbito de la ciberseguridad desde un punto de vista ofensivo y defensivo. En este contexto la aplicación de estas ideas al mundo de la criptografía podría ser provechoso. Existen numerosas propuestas intentando abordar esta cuestión desde diferentes prismas (criptografía neuronal, cryptonets, etc.). En este artículo se centra el interés en la criptografía adversaria. Diversas redes neuronales "descubrirán" un algoritmo común sin conocimiento previo de técnicas de criptografía ni supervisión humana, para ello necesitarán de un atacante que con sus funciones de pérdidas, por ejemplo diferencia de lo detectado por él con lo emitido por el emisor, permite entrenar el sistema y converger a un algoritmo común. Está idea es muy interesante, y al menos desde un punto de vista conceptual, segura desde el punto de vista del atacante (red neuronal). Nuestro trabajo inicia el camino para comprender la seguridad de estas propuestas desde el punto de vista del atacante humano no de red neuronal, aplicando los análisis criptoanalíticos tradicionales y verificando si estas propuestas podrían sustituir a algoritmos reales creados por criptógrafos. Hasta la fecha es fácil de mostrar que todavía la experimentación con este tipo de algoritmos es compleja, costosa computacionalmente y en algunas ocasiones difícil de razonar (una vez entrenado el sistema criptográfico actúa como una caja negra de la cuál no somos capaces de inferir el algoritmo deducido) lo cuál podría atraer consecuencias derivadas de seguridad por oscuridad (algoritmo no visible o difícil de comprender). El análisis actual demuestra que estos sistemas no son realistas para tamaño de bloque y claves actuales, no podría ejecutarse en cualquier dispositivo y sobre todo no podrían considerarse seguros según los estándares actuales. Las pruebas iniciales demuestran carencias de propiedades aleatorias en los criptogramas o una correcta distribución (confusión/difusión) de los cambios en la entrada en el criptograma final (efecto avalancha). No obstante, pensamos que podría ser una línea de investigación interesante en la cual profundizar tanto desde un punto de vista computacional (acelerar la convergencia y el entrenamiento) como del estudio del mejor tipo de redes neuronales que faciliten ciertas propiedades criptográficas. Adicionalmente, la arquitectura GAN de las redes neuronales permitiría introducir conceptos de criptoanálisis más avanzados en Eve, de forma que Alice y Bob entrenaran sus redes neuronales bajo esos supuestos (criptoanálisis lineal, diferencial, efecto avalancha, aleatoriedad, etc.), mucho más sofisticado que lo propuesto en [46]. Si esta línea de investigación supondrá un hito o no en la historia de la criptografía esta por ver. De momento, su utilidad real queda reducida a un ejercicio intelectual.

Referencias

[1] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[3] B. Hariharan, P. Arbelaez, R. B. Girshick, and J. Malik, "Simultaneous detection and segmentation," *CoRR*, vol. abs/1407.1808, 2014. [Online]. Available: http://arxiv.org/abs/1407.1808

[4] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub, "Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, pp. 68 EP –, Jan 2002, article. [Online]. Available: http://dx.doi.org/10.1038/nm0102-68

[5] R. Kaundal, A. S. Kapoor, and G. P. S. Raghava, "Machine learning techniques in disease forecasting: a case study on rice blast prediction." *BMC Bioinformatics*, vol. 7, p. 485, 2006. [Online]. Available: http://dblp.uni-trier.de/db/journals/bmcbi/bmcbi7.html#KaundalKR06

[6] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, and J. Tohka, "Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects," *NeuroImage*, vol. 104, pp. 398 – 412, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811914008131

[7] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8 – 17, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2001037014000464

[8] 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. [Online]. Available: http://arxiv.org/abs/1606.05294

[9] D. Volkhonskiy, I. Nazarov, B. Borisenko, and E. Burnaev, "Steganographic generative adversarial networks," *CoRR*, vol. abs/1703.05502, 2017. [Online]. Available: http://arxiv.org/abs/1703.05502

[10] J. Hayes and G. Danezis, "Generating Steganographic Images via Adversarial Training," *ArXiv e-prints*, Mar. 2017.

[11] D. K. Bhattacharyya and J. K. Kalita, *Network Anomaly Detection: A Machine Learning Perspective*. Chapman & Hall/CRC, 2013.

[12] 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. [Online]. Available: https://doi.org/10.1007/978-3-319-65188-0

[13] F. Iglesias and T. Zseby, "Analysis of network traffic features for anomaly detection," *Machine Learning*, vol. 101, no. 1, pp. 59–84, Oct 2015. [Online]. Available: https://doi.org/10.1007/s10994-014-5473-9

[14] B. Cakir and E. Dogdu, "Malware classification using deep learning methods," in *Proceedings of the ACMSE 2018 Conference*, ser. ACMSE '18. New York, NY, USA: ACM, 2018, pp. 10:1–10:5. [Online]. Available: http://doi.acm.org/10.1145/3190645.3190692

[15] L. Liu, B.-s. Wang, B. Yu, and Q.-x. Zhong, "Automatic malware classification and new malware detection using machine learning," *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 9, pp. 1336–1347, Sep 2017. [Online]. Available: https://doi.org/10.1631/FITEE.1601325

[16] N. Peiravian and X. Zhu, "Machine learning for android malware detection using permission and api calls," in *Proceedings of the 2013 IEEE 25th International Conference on Tools with Artificial Intelligence*, ser. ICTAI '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 300–305. [Online]. Available: http://dx.doi.org/10.1109/ICTAI.2013.53

[17] A. Bansal and S. Mahapatra, "A comparative analysis of machine learning techniques for botnet detection," in *Proceedings of the 10th International Conference on Security of Information and Networks*, ser. SIN '17. New York, NY, USA: ACM, 2017, pp. 91–98. [Online]. Available: http://doi.acm.org/10.1145/3136825.3136874

[18] Z. Xu, S. Ray, P. Subramanyan, and S. Malik, "Malware detection using machine learning based analysis of virtual memory access patterns," in *Proceedings of the Conference on Design, Automation & Test in Europe*, ser. DATE '17. 3001 Leuven, Belgium, Belgium: European Design and Automation Association, 2017, pp. 169–174. [Online]. Available: http://dl.acm.org/citation.cfm?id=3130379.3130417

[19] 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*, 2016.

[20] B. Chernis and R. Verma, "Machine learning methods for software vulnerability detection," in *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, ser. IWSPA '18. New York, NY, USA: ACM, 2018, pp. 31–39. [Online]. Available: http://doi.acm.org/10.1145/3180445.3180453

[21] 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," *CoRR*, vol. abs/1801.01681, 2018. [Online]. Available: http://arxiv.org/abs/1801.01681

[22] 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.

[23] W. Brendel, J. Rauber, and M. Bethge, "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models," *ArXiv e-prints*, Dec. 2017.

[24] A. Shafahi, W. Ronny Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks," *ArXiv e-prints*, Apr. 2018.

[25] A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, X. Hu, C. Xie, J. Wang, Z. Zhang, Z. Ren, A. Yuille, S. Huang, Y. Zhao, Y. Zhao, Z. Han, J. Long, Y. Berdibekov, T. Akiba, S. Tokui, and M. Abe, "Adversarial Attacks and Defences Competition," *ArXiv e-prints*, Mar. 2018.

[26] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, ser. AISec '11. New York, NY, USA: ACM, 2011, pp. 43–58. [Online]. Available: http://doi.acm.org/10.1145/2046684.2046692

[27] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *ArXiv e-prints*, Dec. 2014.

[28] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. Berkay Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," *ArXiv e-prints*, Nov. 2015.

[29] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Berkay Celik, and A. Swami, "Practical Black-Box Attacks against Machine Learning," *ArXiv e-prints*, Feb. 2016.

[30] T. Miyato, S.-i. Maeda, M. Koyama, K. Nakae, and S. Ishii, "Distributional Smoothing with Virtual Adversarial Training," *ArXiv e-prints*, Jul. 2015.

[31] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks," *ArXiv e-prints*, Nov. 2015.

[32] J. Uesato, B. O'Donoghue, A. van den Oord, and P. Kohli, "Adversarial Risk and the Dangers of Evaluating Against Weak Attacks," *ArXiv e-prints*, Feb. 2018.

[33] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial Machine Learning at Scale," *ArXiv e-prints*, Nov. 2016.

[34] J. Su, D. Vasconcellos Vargas, and S. Kouichi, "One pixel attack for fooling deep neural networks," *ArXiv e-prints*, Oct. 2017.

[35] P. Cunningham, M. Cord, and S. J. Delany, *Supervised Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 21–49. [Online]. Available: https://doi.org/10.1007/978-3-540-75171-7_2

[36] O. Bousquet, U. von Luxburg, and G. Rätsch, "Advanced lectures on machine learning, ml summer schools 2003, canberra, australia, february 2-14, 2003, tübingen, germany, august 4-16, 2003, revised lectures," 01 2004.

[37] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[39] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR*, vol. abs/1406.1078, 2014. [Online]. Available: http://arxiv.org/abs/1406.1078

[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[41] S. Hitawala, "Comparative Study on Generative Adversarial Networks," *ArXiv e-prints*, Jan. 2018.

[42] S. Dourlens, "Applied neuro-cryptography," *Option Micro-informatique Micro-électronique*, 1995-1996, available online: https://tinyurl.com/y79tr94e.

[43] Misti Clark and Douglas Blank, "A neural network-based cryptographic system," *MAICS-98*, 1998, available online: https://www.aaai.org/Papers/MAICS/1998/MAICS98-012.pdf.

[44] I. Kanter, W. Kinzel, and E. Kanter, "Secure exchange of information by synchronization of neural networks," *EPL (Europhysics Letters)*, vol. 57, pp. 141–147, Jan. 2002.

[45] A. Klimov, A. Mityagin, and A. Shamir, "Analysis of neural cryptography," in *Advances in Cryptology — ASIACRYPT 2002*, Y. Zheng, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 288–298.

[46] M. Abadi and D. G. Andersen, "Learning to protect communications with adversarial neural cryptography," *CoRR*, vol. abs/1610.06918, 2016. [Online]. Available: http://arxiv.org/abs/1610.06918

[47] R. Ramamurthy, C. Bauckhage, K. Buza, and S. Wrobel, "Using echo state networks for cryptography," *CoRR*, vol. abs/1704.01046, 2017. [Online]. Available: http://arxiv.org/abs/1704.01046

[48] C. Gentry, "A fully homomorphic encryption scheme," Ph.D. dissertation, Stanford, CA, USA, 2009, aAI3382729.

[49] N. Dowlin, R. Gilad-Bachrach, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy," Tech. Rep., February 2016. [Online]. Available: https://tinyurl.com/y84dem5y

[50] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[51] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[52] H. Chabanne, A. de Wargny, J. Milgram, C. Morel, and E. Prouff, "Privacy-preserving classification on deep neural network," *IACR Cryptology ePrint Archive*, vol. 2017, p. 35, 2017.

[53] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *CoRR*, vol. abs/1603.04467, 2016. [Online]. Available: http://arxiv.org/abs/1603.04467

[54] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, Mar 1994.

[55] L. E. Bassham, III, A. L. Rukhin, J. Soto, J. R. Nechvatal, M. E. Smid, E. B. Barker, S. D. Leigh, M. Levenson, M. Vangel, D. L. Banks, N. A. Heckert, J. F. Dray, and S. Vo, "Sp 800-22 rev. 1a. a statistical test suite for random and pseudorandom number generators for cryptographic applications," Gaithersburg, MD, United States, Tech. Rep., 2010.

[56] R. G. Brown, "Dieharder: A Random Number Test Suite v3.31.1," Jun. 2011. [Online]. Available: http://www.phy.duke.edu/\~{}rgb/General/dieharder.php

Mejorando la Detección de Spam Social Utilizando la Subjetividad

Enaitz Ezpeleta, Mikel Iturbe, Iñaki Garitano, Iñaki Velez de Mendizabal, Urko Zurutuza
Mondragon Unibertsitatea
Loramendi 4, 20500, Arrasate-Mondragon
{eezpeleta, miturbe, igaritano, ivelez, uzurutuza}@mondragon.edu

***Resumen*—De la misma forma que el uso de las redes sociales sigue creciendo, las campañas no legítimas en los mismos están sufriendo un auge significativo. Esta es una de las principales razones por las que los usuarios de estas plataformas son victimas del spam social, comprometiendo su privacidad. Para resolver este problema y mejorar la detección de spam en las redes sociales, entre todas las técnicas de análisis de contenido, este estudio se centra en el uso de la subjetividad de los textos. Se demuestra que usando esta técnica es posible mejorar el filtrado de spam social. Primero, se identifican los mejores clasificadores utilizando para ello un conjunto de datos compuesto por comentarios de Youtube, entre ellos muchos comentarios spam. Posteriormente, se crea un nuevo conjunto añadiendo la característica de la subjetividad a cada uno de los textos, para después aplicar los mejores clasificadores al mismo. Los resultados comparativos muestran que esta característica puede ayudar a mejorar los resultados de filtrado de spam social: la precisión es mejorada en los 10 mejores resultados, y el número de falsos positivos es reducido en un 5,58 % de media.**

***Index Terms*—Subjetividad, spam social, redes sociales, Youtube, seguridad información**

I. Introducción

En los últimos años las Redes Sociales Online (RRSS) se han expandido en todo el mundo. La cantidad de usuarios de cada una de las plataformas demuestra la importancia de estos canales de comunicación en la sociedad actual: en diciembre del año 2017 Facebook alcanzó 1.4 billones de usuarios diarios activos [1]; Youtube contaba con más de un billón de usuarios en 2017 [2]; y Twitter tenía 330 millones de usuarios mensuales activos en Junio del 2017 [3].

El incremento exponencial de usuarios ha posibilitado a las organizaciones maliciosas llegar a una gran cantidad usuarios. Los autores en [1] demuestran que estos sitios web son actualmente uno de los sistemas de comunicación más utilizados para el envío de spam. Para ello analizaron el spam en varias RRSS y cuantificaron la cantidad de spam enviado a través de diferentes cuentas en las RRSS.

Con el objetivo de solucionar este problema, los autores en [2] demostraron que las técnicas de análisis de contenido pueden ayudar a mejorar el filtrado de spam. En el estudio, se analiza un conjunto de emails para enriquecer el conjunto con el sentimiento de cada mensaje, y así se consigue mejorar la precisión de filtrado. Siguiendo un procedimiento parecido, el estudio que se presenta en este trabajo selecciona otra técnica de análisis de contenido, la subjetividad, para mejorar el filtrado de spam social, partiendo de la hipótesis de que los mensajes automáticos (spam) son más objetivos.

Primero, se aplican varios clasificadores con distintas configuraciones al conjunto de comentarios de Youtube con el objetivo de identificar los mejores diez métodos de filtrado. Después, se aplica un analizador de subjetividad a cada comentario para crear nuevo conjunto de datos añadiendo esa característica al conjunto original. Una vez creado el conjunto enriquecido, los diez clasificadores previamente identificados son aplicados al conjunto con la subjetividad, para finalmente comparar y analizar los resultados obtenidos.

El resto del artículo está organizado de la siguiente manera. La Sección II analiza trabajos previos realizados en el área de las técnicas de filtrado de spam social, e introduce al lector en las técnicas de análisis de subjetividad. La sección III describe el proceso de clasificación de mensajes y de agregación de la subjetividad. En la Sección IV se comparan los resultados obtenidos a partir de modelos de clasificación clásicos con los modelos de clasificación con la subjetividad del mensaje como característica añadida. Finalmente, discutimos los resultados en la Sección V.

II. Trabajos relacionados

El spam social está convirtiéndose actualmente en uno de los segmentos más activos en la comunidad científica [3]. Stringhini et al. [4] demostraron que es posible identificar cuentas de spammers de forma automática en RRSS como Facebook, Twitter y MySpace, las cuales fueron bloqueadas posteriormente. Además, Wang et al. [5] proponían un sistema de detección de spam que es capaz de analizar las RRSS con el objetivo de buscar spam. De forma parecida, Egele et al. [6] presentaron COMPA, una herramienta capaz de detectar cuentas comprometidas en las RRSS basándose en comportamientos anómalos de los usuarios. En el caso de Gao et al. [7] emplearon métodos de clasificación y clustering para la detección de campañas de spam en las RRSS Twitter y Facebook. Ezpeleta et al. [8] mostraron que personalizar mensajes de spam utilizando información públicamente disponible en los perfiles de las RRSS aumenta significativamente la eficacia del spam.

En el área del spam en RRSS, sobretodo ha sido estudiado el spam en Twitter. Kwak et al. [9] identificaron las RRSS como herramientas de difusión de la información, además de confirmar que Twitter es una plataforma atractiva para realizar campañas de spam. Yang et al. [10] describen la dinámica que las cuentas criminales siguen en Twitter y cómo interactúan entre ellos y las cuentas no criminales. Adicionalmente, Song et al. [11] demostraron cómo es posible detectar spammers

[1] http://newsroom.fb.com/company-info/
[2] https://www.youtube.com/yt/about/press/
[3] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Tabla I
LOS 10 MEJORES RESULTADOS.

Nombre	TP	TN	FP	FN	Precisión
ComplementNaivestwv.go.ngtok	972	866	85	33	93,97
ComplementNaivestwv.go.ngtok.stemmer	974	862	89	31	93,87
NaiveBayesMultinomialUpdateable.c.stwv.go.ngtok	978	824	127	27	92,13
NaiveBayesMultinomialUpdateable.c.stwv.go.ngtok.stemmer	978	978	135	27	91,72
NaiveBayesMultinomialUpdateable.stwv.go.ngtok	975	975	113	30	92,69
NaiveBayesMultinomialUpdateable.stwv.go.ngtok.stemmer	975	975	119	30	92,38
NaiveBayesMultinomial.c.stwv.go.ngtok	978	978	127	27	92,13
NaiveBayesMultinomial.c.stwv.go.ngtok.stemmer	978	978	135	27	91,72
NaiveBayesMultinomial.stwv.go.ngtok	975	975	113	30	92,69
NaiveBayesMultinomial.stwv.go.ngtok.stemmer	975	975	119	30	92,38

basándose en características como la distancia y la conectividad entre emisor y receptor, las cuales son más difíciles de manipular por los spammers.

Aunque no en la misma cantidad que el email, Twitter o Facebook, pero el spam en Youtube también ha sido objeto de estudio. Chaudhary y Sureka [12] minaron las descripciones de los videos junto con las características de popularidad de los mismos para detectar videos spam en Youtube. O'Callaghan et al. [13] utilizaron información de la red para caracterizar usuarios con el objetivo de etiquetar e identificar cuentas propulsoras de campañas de spam.

Cabe destacar en este punto que aunque numerosas técnicas nuevas para la detección de spam han sido publicadas, el spam en las RRSS sigue siendo un problema sin resolver todavía [1].

En esa dirección, el análisis basado en contenido, donde se analiza el texto para extraer su significado u objetivo usando diferentes técnicas como el Análisis de Sentimiento (AS), se muestra como una estrategia prometedora para mejorar los resultados de detección de spam en RRSS [14], [2]. El objetivo del AS reside en la identificación de la naturaleza positiva o negativa de los documentos [15], y para ello es posible usar técnicas supervisadas de aprendizaje automático [16] o no supervisadas [17]. En cuanto a la efectividad de usar el AS para la detección de spam, se ha demostrado su valía en diferentes casos como el spam social [18], textos cortos informales [19], Twitter spam [14], email spam [2] y detección de fraude [20].

En estudios previos también se hace mención a la posibilidad de utilizar el AS junto a la subjetividad para la clasificación de documentos [21], y tomando esa investigación como base, este trabajo irá más allá del estado del arte para utilizar solamente la subjetividad para detectar spam en la plataforma social de videos Youtube.

III. DISEÑO E IMPLEMENTACIÓN DEL SISTEMA

Partiendo del conjunto de datos original, el proceso que se ha seguido en el estudio se divide en dos partes principalmente, tal y como se describe en la Figura 1.

Primero, varios clasificadores son aplicados al conjunto de datos compuesto por comentarios de social media (spam y ham) con el objetivo de identificar y seleccionar los mejores diez clasificadores de spam social. En el mismo paso, los diez mejores resultados son extraídos.

Segundo, la subjetividad de cada mensaje se añade al conjunto de datos original para crear un nuevo conjunto. Durante el análisis de subjetividad, se realiza un experimento descriptivo. En la siguiente fase, los diez mejores clasificadores identificados en la fase anterior se aplicarán al nuevo conjunto para comparar los resultados.

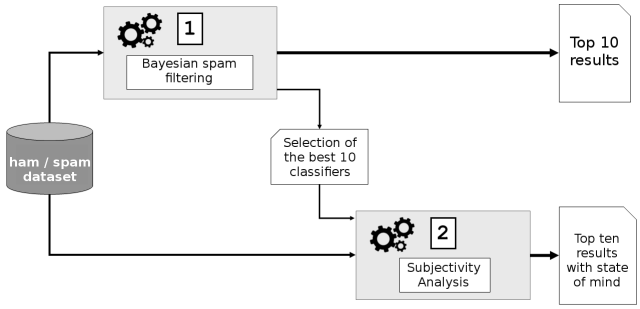


Figura 1. Proceso del experimento.

Para validar tanto los algoritmos como los resultados obtenidos, se ha utilizado la validación cruzada en 10 carpetas, y los resultados son analizados en términos de número de falsos positivos y de precisión, siendo este el porcentaje de comentarios correctamente clasificados, y los comentarios legítimos clasificados como spam son considerados falsos positivos.

III-A. Dataset

Durante el transcurso del estudio se ha utilizado un conjunto de datos públicamente disponible: *Youtube Comments Dataset* [4]: Presentado en [22]. Este conjunto contiene 6.431.471 comentarios de la red social Youtube[5]. Entre todos los comentarios, 481.334 mensajes están etiquetados como spam. Con el objetivo de utilizar un conjunto parecido al utilizado en [2], se ha creado un conjunto de 3.000 mensajes legítimos y 1.000 de spam. Estos mensajes han sido seleccionados de forma aleatoria y teniendo en cuenta solo los comentarios escritos en inglés.

III-B. Identificando el mejor clasificador de spam social

Con el objetivo de identificar los mejores detectores de spam, varios clasificadores con diferentes configuraciones y filtros han sido aplicados al conjunto de datos Youtube Comments dataset.

Siguiendo la estrategia presentada en [2], 7 clasificadores diferentes con 56 combinaciones de configuración por cada clasificador serán aplicados (un total de 392 combinaciones). Los mejores diez resultados se recogen en la Tabla I. Durante

[4]http://mlg.ucd.ie/yt/
[5]https://www.youtube.com

el transcurso de este estudio se han aplicado los siguientes clasificadores: (1) Large-scale Bayesian logistic regression for text categorization, (2) discriminative parameter learning for Bayesian networks, (3) Naive Bayes classifier, (4) complement class Naive Bayes classifier, (5) multi-nominal Naive Bayes classifier, (6) updateable Naive Bayes classifier, y (7) updateable multi-nominal Naive Bayes classifier.

Las nomenclaturas y los acrónimos utilizados en la Tabla I y durante el artículo se recogen en la Tabla II.

Tabla II
NOMENCLATURAS.

	Significado
CNB	Complement Naive Bayes
NBM	Naive Bayes Multinomial
NBMU	Naive Bayes Multinomial Updatable
.c	idft F, tft F, outwc T [5]
.i.c	idft T, tft F, outwc T [5]
.i.t.c	idft T, tft T, outwc T [5]
.stwv	String to Word Vector
.go	Opciones generales
.wtok	Word Tokenizer
.ngtok	NGram Tokenizer
.stemmer	Stemmer
.igain	Selección de atributos basados en ganancia de información

Una vez identificados los mejores clasificadores y los mejores resultados en el conjunto de datos Youtube Comments dataset, el análisis de la subjetividad de cada mensaje es realizada.

III-C. Análisis de la subjetividad

Para analizar la subjetividad de los comentarios de los youtubers, cada texto es analizado y la característica de subjetividad es añadida al conjunto de datos original. Así se crea un nuevo conjunto de datos al cual los mejores clasificadores identificados en la primera fase son aplicados tal y como se muestra en la Figura 2.

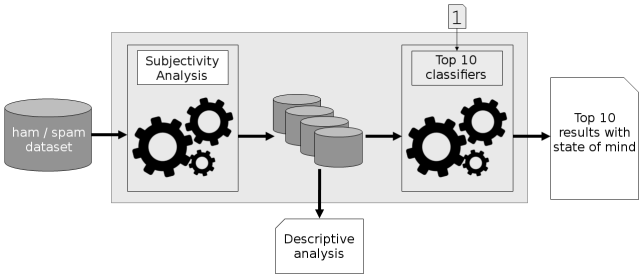


Figura 2. Metodología para el análisis de la subjetividad.

Como herramienta para el análisis de la subjetividad se utiliza una API llamada *TextBlob*. Esta API públicamente disponible fue desarrollada por Steven Loria orientada al procesamiento del lenguaje natural (PLN), para solventar las tareas típicas de PLN. En el caso de la subjetividad, recibiendo una frase (string) devuelve un valor real en el rango [0.0,1.0] representativo de la subjetividad de la frase, donde 0.0 es muy objetivo y 1.0 es muy subjetivo.

[5]idft es Inverse Document Frequency (IDF) Transformation; tft es Term Frequency score (TF) Transformation; outwc cuenta el número de veces que aparecen las palabras.

IV. RESULTADOS EXPERIMENTALES

Para validar la hipótesis planteada en este trabajo, se han desarrollado los experimentos anteriormente descritos de los cuales se han extraído las siguiente conclusiones.

IV-A. Análisis descriptivo

Con el objetivo de analizar el contenido del conjunto de datos y ver si las diferencias entre comentarios legítimos y no legítimos en cuanto a la subjetividad son significativas, se desarrollo un estudio descriptivo. De este estudio la conclusión más representativa que se extrae se muestra en la Figura 3. En ella, gracias a un diagrama de cajas, se puede apreciar la diferencia que existe entre los textos spam y no spam. Tal y como se puede ver, los mensajes legítimos tienden a ser más subjetivos mientras los comentarios spam se perfilan como más objetivos.

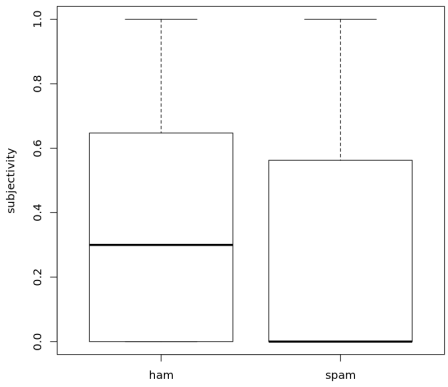


Figura 3. Comparativa de la subjetividad entre la spam y ham.

Este análisis ofrece mimbres para validar la hipótesis de que los textos no legítimos que en la mayoría de casos son comentarios automáticos, son más objetivos, y con ello se abre la puerta a que la utilización de esta característica pueda ser de ayuda en la mejora de los filtrados actuales.

Finalmente, durante esta fase, se crea el nuevo conjunto de datos que cuenta por un lado con la información original del texto y la etiqueta (ham/spam), y por otro la nueva característica de la subjetividad de cada mensaje. Esta información es añadida de forma numérica utilizando el resultado que la API *TextBlob* devuelve durante el análisis de cada comentario.

IV-B. Experimentos predictivos

En primer lugar, partiendo de los diez mejores resultados obtenidos en la fase previa, los cuales definen la base de resultados a mejorar, se aplican los diez mejores clasificadores al conjunto de datos resultante del análisis descriptivo. Una vez obtenidos los resultados de estos clasificadores aplicados sobre el conjunto enriquecido con información de la subjetividad, se comparan los nuevos resultados con los resultados mostrados en la Tabla I.

Esa comparativa en términos de precisión y número de falsos positivos se recogen en la Tabla III.

Tabla III
LOS 10 MEJORES RESULTADOS.

	Normal		Subjetividad		
Nombre	FP	Precisión	FP	Precisión	Mejoría en FP
CNB.stwv.go.ngtok	85	93,97	81	94,17	4,71
CNB.stwv.go.ngtok.stemmer	89	93,87	86	93,97	3,37
NBM.stwv.go.ngtok	113	92,69	105	93,00	7,08
NBMU.stwv.go.ngtok	113	92,69	108	92,84	4,42
NBM.stwv.go.ngtok.stemmer	119	92,38	111	92,79	6,72
NBMU.stwv.go.ngtok.stemmer	119	92,38	114	92,69	4,20
NBM.c.stwv.go.ngtok	127	92,13	116	92,59	8,66
NBMU.c.stwv.go.ngtok	127	92,13	120	92,38	5,51
NBM.c.stwv.go.ngtok.stemmer	135	91,72	127	92,13	5,93
NBMU.c.stwv.go.ngtok.stemmer	135	91,72	128	92,08	5,19
				avg	5,58

Como muestran los resultados de la Tabla III, enriquecer el conjunto de datos con la información del análisis de la subjetividad, ayuda a mejorar los mejores resultados. Por un lado, la máxima precisión obtenida durante los experimentos de la primera fase (sin subjetividad) es mejorada del 93,07 % al 94,17 %; a su vez, en los diez mejores algoritmos también se mejora la precisión; y también el número de falsos positivos es reducido en todos los casos en un 5,58 % de media.

Con el objetivo de validar los resultados obtenidos, en la Figura 4 se muestra la curva de ROC resultante de los dos casos, con subjetividad y sin subjetividad, donde se demuestra que el área utilizando el conjunto enriquecido con la subjetividad es ligeramente superior al área sin subjetividad.

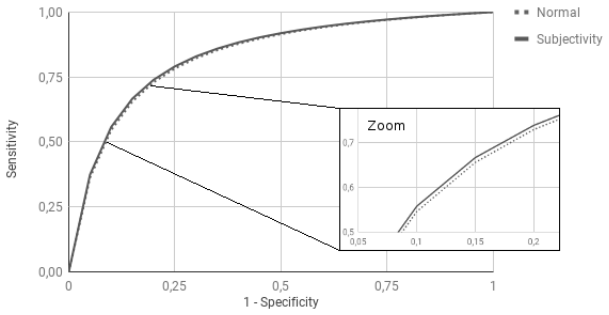


Figura 4. Curva de ROC de los mejores algoritmos con y sin subjetividad.

V. CONCLUSIONES

Este trabajo plantea una hipótesis que da pie a futuros trabajos para la mejora de los filtros de spam actuales. Aunque las herramientas actuales cuentan con un grado madurez en el mercado, los resultados dejan entrever que al añadir al conjunto de características variables relacionadas con el propio contenido o semántica del comentario, una mejoría de los resultados es posible. Así, el análisis de la subjetividad de los comentarios supone un avance hacia la detección de la intencionalidad de los mensajes, partiendo de la hipótesis de que los comentarios no legítimos son más objetivos dado su origen automático en muchos casos.

En el artículo se muestra cómo añadiendo la característica de la subjetividad en cada comentario, puede mejorar el filtrado de spam tanto en términos de precisión, así como en el número de falsos positivos. Muestra de ello son que la precisión máxima obtenida sin subjetividad ha sido mejorada del 93,07 % al 94,17 %, mejorando el resultado en los diez mejores clasificadores, y que el número de falsos positivos ha sido reducido de media en un 5,58 %.

AGRADECIMIENTOS

Este trabajo ha sido financiado parcialmente por MINECO/AEI/FEDER, a través del subproyecto TIN2017-84658-C2-2-R. Asimismo, Iñaki Garitano está parcialmente financiado por INCIBE bajo la beca "INCIBEC-2015-02495" correspondiente a "Ayudas para la Excelencia de los Equipos de Investigación avanzada en ciberseguridad".

REFERENCIAS

[1] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," in *Proceedings of the 17th ACM conference on Computer and communications security*, ser. CCS '10. New York, NY, USA: ACM, 2010, pp. 681–683. [Online]. Available: http://doi.acm.org/10.1145/1866307.1866396

[2] E. Ezpeleta, U. Zurutuza, and J. M. Gómez Hidalgo, "Does sentiment analysis help in bayesian spam filtering?" in *Hybrid Artificial Intelligent Systems: 11th International Conference, HAIS 2016, Sevilla, Spain, April 18-20, 2016*. Springer, 2016.

[3] A. Almaatouq, E. Shmueli, M. Nouh, A. Alabdulkareem, V. K. Singh, M. Alsaleh, A. Alarifi, A. Alfaris, and A. S. Pentland, "If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts," *International Journal of Information Security*, vol. 15, no. 5, pp. 475–491, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10207-016-0321-5

[4] 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: ACM, 2010, pp. 1–9. [Online]. Available: http://doi.acm.org/10.1145/1920261.1920263

[5] D. Wang, D. Irani, and C. Pu, "A social-spam detection framework," in *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*. ACM, 2011, pp. 46–54.

[6] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Compa: Detecting compromised accounts on social networks." in *NDSS*. The Internet Society, 2013. [Online]. Available: http://dblp.uni-trier.de/db/conf/ndss/ndss2013.html#EgeleSKV13

[7] H. Gao, Y. Chen, K. Lee, D. Palsetia, and A. N. Choudhary, "Towards online spam filtering in social networks." in *NDSS*. The Internet Society, 2012. [Online]. Available: http://dblp.uni-trier.de/db/conf/ndss/ndss2012.html#GaoCLPC12

[8] E. Ezpeleta, U. Zurutuza, and J. M. G. Hidalgo, "A study of the personalization of spam content using facebook public information," *Logic Journal of the IGPL*, vol. 25, no. 1, pp. 30–41, 2017. [Online]. Available: http://dx.doi.org/10.1093/jigpal/jzw040

[9] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 591–600.

[10] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 71–80.

[11] J. Song, S. Lee, and J. Kim, "Spam filtering in twitter using sender-receiver relationship," in *Recent Advances in Intrusion Detection*. Springer, 2011, pp. 301–317.

[12] V. Chaudhary and A. Sureka, "Contextual feature based one-class classifier approach for detecting video response spam on youtube," in *2013 Eleventh Annual Conference on Privacy, Security and Trust*, July 2013, pp. 195–204.

[13] D. O'Callaghan, M. Harrigan, J. Carthy, and P. Cunningham, "Network analysis of recurring youtube spam campaigns," *arXiv preprint arXiv:1201.3783*, 2012.

[14] N. Perveen, M. M. S. Missen, Q. Rasool, and N. Akhtar, "Sentiment based twitter spam detection," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 7, no. 7, pp. 568–573, 2016.

[15] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[16] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, ser. EMNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 79–86. [Online]. Available: http://dx.doi.org/10.3115/1118693.1118704

[17] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 417–424. [Online]. Available: http://dx.doi.org/10.3115/1073083.1073153

[18] X. Hu, J. Tang, H. Gao, and H. Liu, "Social spammer detection with sentiment information," in *Proceedings of the 2014 IEEE International Conference on Data Mining*, ser. ICDM '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 180–189. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2014.141

[19] M. H. Arif, J. Li, M. Iqbal, and K. Liu, "Sentiment analysis and spam detection in short informal text using learning classifier systems," *Soft Computing*, Jul 2017. [Online]. Available: https://doi.org/10.1007/s00500-017-2729-x

[20] S. Mahajan and V. Rana, "Spam detection on social network through sentiment analysis," *Advances in Computational Sciences and Technology*, vol. 10, no. 8, pp. 2225–2231, 2017.

[21] B. Liu, "Sentiment analysis and subjectivity." *Handbook of natural language processing*, vol. 2, pp. 627–666, 2010.

[22] D. O'Callaghan, M. Harrigan, J. Carthy, and P. Cunningham, "Network analysis of recurring youtube spam campaigns," *CoRR*, vol. abs/1201.3783, 2012. [Online]. Available: http://arxiv.org/abs/1201.3783

Constrained approximate bit-parallel search with application in cryptanalysis

Slobodan Petrović
Norwegian University of Science and Technology
P.O. box 191, N-2802 Gjøvik, Norway
slobodan.petrovic@ntnu.no

***Abstract*—The fastest known search algorithms are those that exploit the inherent bit-parallelism of computer systems. By using only simple bit operations, such as shifting and OR-ing, it is possible to parallelize the search algorithm operations and obtain a significant gain in efficiency compared to classical search algorithms. The gain in speed depends linearly on the length of a computer memory cell. If the search pattern is longer than a computer memory cell, it is possible to concatenate several memory cells and proceed along the same search procedure, but that operation slows down the computation. In addition, some applications require introduction of various constraints in search, which dictates significant modifications of the search algorithm. In this paper, a constrained approximate bit-parallel search algorithm with ordered runs of deletions and substitutions is defined. Some aspects of applications of such a search algorithm in cryptanalysis of stream ciphers employing irregularly clocked linear feedback shift registers are studied. These aspects include the appropriate constraints and the reconstruction of the clocking sequence. The challenges of implementation of such an algorithm in software are also discussed, as well as the preliminary experimental results regarding the time consumption of the cryptanalytic attack.**

***Index Terms*—Approximate search, Bit-parallelism, Cryptanalysis, Constraints**

I. Introduction

In the last 25 years, bit-parallel search algorithms and their applications in various fields have been extensively studied. Typical representatives of these algorithms are Shift-AND [1] (and its more efficient implementation Shift-OR), BNDM [2], BNDMq [3], etc. Their main advantage over the classical search algorithms such as the KMP (Knuth, Morris and Pratt) [4] or AC (Aho, Corasick) [5] is improved efficiency due to operating many comparators in parallel. The maximum number of these comparators is determined by the length of a computer memory cell, typically 32 or 64 bits nowadays. It is not conceptually difficult to implement these algorithms on hardware with longer memory cells if the search pattern is much longer than the length of a word used by commodity hardware. Alternatively, it is possible to implement these algorithms on any hardware by joining the memory cells at the cost of reduced efficiency (see, for example, [1]).

The applications of these algorithms have included computational biology, document retrieval, digital forensics, etc. In this paper, we study application of bit-parallel search algorithms in a new field - cryptanalysis of a class of stream ciphers employing so-called irregularly clocked Linear Feedback Shift Registers (LFSRs). This is a well known and widely used family of pseudorandom sequence generator schemes, whose popularity is based on a long period of the generated sequence, its large linear complexity, good statistical properties, and the speed of output sequence generation. However, it has been shown by many authors (see, for example, [6], [7]), that it is possible to launch a generalized correlation attack against the generators from this class. This attack uses the constrained edit distance, which is computed between the output sequence of the controlled LFSR without irregular clocking and the noised intercepted output sequence of the whole generator. If the constrained edit distance is smaller than a threshold given in advance, the initial state of the controlled LFSR that generated the tested output sequence is accepted as a candidate for the real initial state that also, when clock-controlled, produced the intercepted output sequence. The key part of this attack is computing the constrained edit distance, whose time complexity is quadratic in the length of the intercepted output sequence of the generator. To speed-up the computation of the constrained edit distance, parallelization is necessary.

Attempts to parallelize the dynamic programming-based computation of the edit distance have been made by various authors (see, for example, [8], [9]). But the produced algorithms parallelize the computation of the *unconstrained* edit distance. Modifying the existing procedures of this kind directly in order to take into account various constraints is a challenge. Namely, the basic assumptions regarding the differences between two neighboring columns of the edit distance matrix, which were exploited to parallelize the dynamic programming-based algorithms do not hold if the constraints are introduced.

Another way of parallelizing edit distance computation-related algorithms is the application of bit-parallelism search procedures mentioned above. However, design of bit-parallel search algorithms with constraints, especially those constraints that are needed in cryptanalysis, has not been in focus so far. Some constraints have been studied, such as those on the maximum number of certain edit operations [10], but not the ones needed in cryptanalysis.

In this paper, we propose application of a bit-parallel constrained search algorithm in cryptanalysis of stream ciphers employing irregular clocking. We modify the algorithm from [6] by replacing the dynamic programming-based constrained edit distance computation algorithm with a novel constrained bit-parallel search procedure. For this procedure, we define constraints on run lengths of edit operations and their ordering. We discuss the time and space complexities of the new attack and the possibility of reconstructing the clock control sequence after the attack.

The paper is organized as follows: in Section II, we give the fundamentals of approximate bit-parallel search and introduce

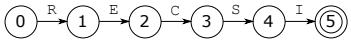


Fig. 1. The NFA assigned to the search pattern "RECSI".

the analyzed pseudorandom sequence generator employing irregular clocking. In Section III, we describe the modified attack against this generator using the constrained approximate search. In Section IV, we discuss the preliminary experimental results obtained with the classical and the new attack along with the theoretical gain in space and time by using the modified attack. Section V concludes the paper.

II. Bit-parallelism and cryptanalysis

A. Deterministic and non-deterministic finite automata

By search we mean finding every occurrence (possibly overlapping) of a *search pattern* in a *search string*, where both the search pattern and the search string consist of symbols from the same alphabet. Almost all search algorithms make use of Finite State Machines (FSM). These machines consist of special nodes called *states* that make transitions on input events. An input event in the context of search is appearance of a new symbol from the search string. In this paper, we study only very simple FSMs that have linear form.

An FSM can make one or several transitions on an input event. If only one transition is performed per input event (symbol) then such a machine is called Deterministic Finite Automaton (DFA). If more than one transition occurs per input symbol then such a machine is called Non-deterministic Finite Automaton (NFA). NFA is a theoretical concept, where infinite parallelism is assumed. In practice, we only simulate operation of an NFA. The bit-parallel search algorithms perform such simulations.

An NFA operates in the following way. At each input symbol, the existing copy of the machine tries to match it, i.e. to make a transition. If such a transition is possible then the machine remains *active*. Otherwise, the machine becomes inactive and cannot be re-activated at appearance of any future symbol. At the same time, a new copy of the same machine is created with an active initial state. This new copy of the machine also tries to match the same input symbol, which may result in remaining active or de-activating. This procedure is repeated for each new symbol of the search string.

If, after such copy making and performing transitions (if possible), some machine reaches the final state, then we have an *occurrence* of the search pattern. Infinite parallelism assumes infinite copy making and as such is impossible in practice. We simulate an NFA by limiting the number of existing copies of the machine at a time. Usually, we maintain at most m copies of the machine, where m is the length of the search pattern. The NFA is always based on the search pattern and does not depend on the search string. An example of an NFA that corresponds to the search pattern $w =$ "RECSI" is presented in Fig. 1. Each state makes a transition on the corresponding symbol except the final state (double circled). The fact that a new copy of the machine is always active in the state 0 is usually illustrated by adding a loop in that state.

An alternative representation of an NFA involves so-called ε-transitions, which are the transitions performed without any input. It can be shown that both representations are equivalent.

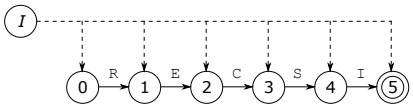


Fig. 2. The NFA with ε-transitions assigned to the search pattern "RECSI".

With ε-transitions, the machine starts its operation from a hypothetical initial state I, from which it can activate any other state of the machine that corresponds to the ending symbol of a *substring* of the search pattern. The NFA corresponding to the search pattern $w =$ "RECSI" employing ε-transitions is presented in Fig. 2. The appearance of the string "REC" makes the state 3 active, the appearance of the string "CSI" makes the state 5 active whereas after receiving the string "CIS" no state of the machine will be active since "CIS" is not a substring of the search pattern. The representation of the NFA with ε-transitions is convenient for presenting special machines used in *approximate* search.

B. Bit-parallelism and exact search

As we have explained above, we maintain at most m copies of the NFA assigned to the search pattern of length m at a time. Every newly created copy of the NFA at the appearance of the next symbol from the search string is active in the state 0 (the initial state). This means that after processing m symbols from the search string we have to remove the oldest copy of the NFA and add the newly created copy of the NFA. In addition, it is easy to notice that only the status (active or inactive) of a copy of the NFA is necessary to check in order to detect occurrence (if a copy of the NFA is still active after processing the last m consecutive symbols of the search string then we have found an occurrence). Thus, it is only necessary to memorize the status of each machine, which can be encoded with 1 (active) or 0 (inactive). We keep track of these status bits in a status word D, whose Most Significant Bit (MSB) corresponds to the oldest machine that we keep track of in the NFA simulation. Finally, it can be noticed that each active copy of the NFA always expects the same symbol (the oldest copy (MSB) always expects the last symbol of the search pattern, etc.) This can be exploited to create so-called bit masks that will be used in fast updating of the search status word D. In the bit mask assigned to a symbol, we set 1 at the position where this symbol is located in the *reversed* search pattern.

Having described all the elements of the NFA simulation, we can now summarize the procedure of search status word updating in the following way:

1. To remove the oldest copy of the NFA, shift the search status word one position to the left.
2. To create a new copy of the NFA active in the 0 state, OR the search status word D with $0^{m-1}1$.
3. To try to match the current symbol of the search string by all the copies of the NFA at the same time (bit-parallelism!), AND the search status word D with the bit mask corresponding to that symbol.
4. To check whether there is an occurrence, check if the MSB of the search status word D is equal to 1.

In the beginning, the value of the search status word is 0^m.

Thus, the search algorithm employing bit-parallelism is the procedure of updating the search status word D each time a new symbol of the search string appears. We formalize this by the following expression (the search status word update formula):

$$D_i = ((D_{i-1} << 1) \text{ OR } 0^{m-1}1) \text{ AND } B[S_i], \quad (1)$$

where D_i is the value of the search status word after processing the current symbol S_i of the search string and $B[S_i]$ is the value of the bit mask corresponding to that symbol.

The equation (1) defines so-called Shift-AND bit-parallel exact search algorithm [1], which was the first published algorithm of this kind. It is easy to see that, if we complement the search status word and the bit masks, we can save one bit operation on the search status word D, giving rise to the so-called Shift-OR algorithm, whose update formula is given by the equation (2).

$$D_i = (D_{i-1} << 1) \text{ OR } B[S_i], \quad (2)$$

To check whether we have found an occurrence, we check whether the MSB of the search status word is equal to 0.

In the rest of the paper, we shall only use the Shift-OR type algorithms since they are faster than Shift-AND. We illustrate the Shift-OR algorithm on an example. Suppose the search pattern is w =”RECSI” and the search string S =”laRECSI2018”. Here, $m = 5$ so the length of the search status word D is 5 bits. We first define the bit masks for the symbols present in the search pattern w (the bit masks for the other symbols are 1^m). We have $B['R'] = 11110$, $B['E'] = 11101$, $B['C'] = 11011$, $B['S'] = 10111$, and $B['I'] = 01111$. The initial value of the search status word is $D_0 = 11111$. We update the search status word by using the equation (2) for each symbol S_i of the search string, $i = 1, \ldots, n$, where n is the length of S. So we have

S_1 =’l’
$D_1 = (D_0 << 1) \text{ OR } B[S_1] =$
$(11111 << 1) \text{ OR } 11111 = 11111$

S_2 =’a’
$(11111 << 1) \text{ OR } 11111 = 11111$

S_3 =’R’
$(11111 << 1) \text{ OR } 11110 = 11110$

S_4 =’E’
$(11110 << 1) \text{ OR } 11101 = 11101$

S_5 =’C’
$(11101 << 1) \text{ OR } 11011 = 11011$

S_6 =’S’
$(11011 << 1) \text{ OR } 10111 = 10111$

S_7 =’I’
$(10111 << 1) \text{ OR } 01111 = 01111$
Since the MSB of the search status word D is equal to 0, we have found an occurrence of the search pattern w in the search string S at the position 7. The algorithm continues until all the search string symbols are processed since other occurrences may take place.

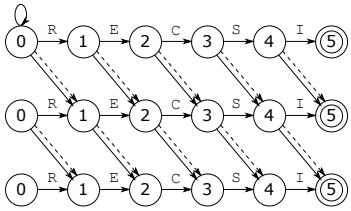


Fig. 3. NFA assigned to the search pattern ”RECSI” allowing 2 errors.

C. Approximate bit-parallel search

We can allow errors in either the search pattern (such as misspelling) or the search string. The errors that are relevant for our application in cryptanalysis are deletions of symbols and substitutions of symbols. For other applications, insertions and reversals can also be considered, (see, for example, [11]). To find the search pattern in a search string with errors, we extend the original Shift-OR algorithm by using $k + 1$ NFAs for exact search, where k is the allowed number of errors. Thus, we obtain an array of $k + 1$ NFAs. Possible transitions are: a match (a transition on an expected symbol), which is a horizontal transition; a substitution, which is a diagonal transition; a deletion, which is a diagonal ε-transition. An example of an NFA corresponding to the search pattern w =”RECSI” with $k = 2$ errors allowed is given in Fig. 3. A solid diagonal line represents a substitution and a dashed diagonal line represents a deletion. A loop in the state 0 of the 0-th row of the NFA array means that an arbitrary number of deletions is allowed before the first symbol of the search pattern w appears in the search string S.

All the transitions from the active states occur in parallel at each input symbol from the search string S. Regarding the relation between a match and a substitution, it is possible to allow a substitution transition even if a match transition in the previous row of the machine is possible (so-called Σ version of the machine, see [11]). The other alternative, which is more interesting for our particular application, is not to allow that (so-called $\bar{p}$ version of the machine in [11]). Instead of a search status word, we now have a search status array R consisting of $k + 1$ search status words of length m, where m is the length of the search pattern. The 0-th row of R corresponds to the case with no errors, the first row corresponds to the case with one error etc. For such a machine, the search status array update formula is the following (see, for example, [11], [12])

$$\begin{aligned} R_0^{'} &= (R_0 << 1) \text{ OR } B[S_j] \\ R_i^{'} &= ((R_i << 1) \text{ OR } B[S_j]) \text{ AND } \text{(match)} \\ &\quad ((R_{i-1} << 1) \text{ OR NOT } B[S_j]) \text{ AND } \text{(sub)} \\ &\quad (R_{i-1}^{'} << 1) \text{ (del)} \\ &\quad i = 1, \ldots, k \end{aligned} \quad (3)$$

The complement of the bit mask OR-ed with the shifted status word at a substitution ensures that no substitution transition occurs if a match occurred in the previous row (the $\bar{p}$ version of the machine).

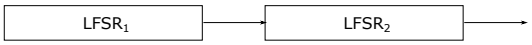


Fig. 4. The analyzed pseudorandom generator.

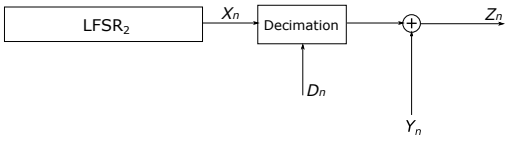


Fig. 5. The statistical model of the generator from Fig. 4.

D. The analyzed pseudorandom generator

We now define the pseudorandom generator scheme that is analyzed in this paper. It consists of two LFSRs, LFSR_1 and LFSR_2, where LFSR_1's output binary sequence clocks LFSR_2. This type of connection is called the Binary Rate Multiplier (BRM) [13]. The generator is presented in Fig. 4. It operates as follows:

1. LFSR_1 is clocked once.
2. If the output bit of LFSR_1 is 0, then LFSR_2 is clocked once and its output bit is sent to the output of the generator. Otherwise, LFSR_2 is clocked twice, but only the second generated bit from LFSR_2 is sent to the output of the generator.

Thus, if we consider only LFSR_2 without irregular clocking (i.e. as an autonomous device) then its output sequence is non-uniformly decimated by means of the clocking sequence from LFSR_1 in such a way that maximum 1 bit of the output sequence from the autonomous LFSR_2 is skipped at a time. This case is known as $0-1$ clocking, which is the simplest possible case of non-uniform decimation. But the importance of this case is high since this simple scheme produces the output bits at the fastest rate.

The statistical model of the generator from Fig. 4 is given in Fig. 5. This statistical model was used in [6] and in this paper we use the same model. The integer sequence D_n is binary in the case of $0-1$ clocking. The binary sequence Y_n is the noise sequence, in which the probability $P(Y_n = 1) < 0.5$.

III. Cryptanalysis of the generator

In this section, we first describe the original attack against the scheme presented in Fig. 4. Then we modify the attack by using bit-parallel search with special constraints instead of the original dynamic programming-based constrained edit distance computation used in [6].

A. The original attack

The original attack from [6] against the pseudorandom generator scheme presented in Fig. 4 proceeds as follows: Let E be the maximum number of output bits from LFSR_2 that can be skipped at a time (in this paper, we only consider the 0-1 clocking so $E = 1$). Let M be the length of the intercepted noised output sequence Z from the generator. Then, for each non-zero initial state of LFSR_2:

1. Generate the output sequence X from LFSR_2 without irregular clocking starting from this initial state.
2. Compute the constrained edit distance between the sequences X and Z.
3. If the constrained edit distance is below a threshold defined in advance, then accept this initial state of LFSR_2 as a candidate initial state that could produce the sequence Z.

The constraints used in the computation of the constrained edit distance are the following:

1. The maximum length of a run of deletions is equal to E (in our case, $E = 1$).
2. Deletions occur before substitutions. This reflects exactly the operation of the generator.

In [6], the constrained edit distance was computed using a dynamic programming formula of Oommen's type [14]. In this computation, it was allowed either to have an arbitrary length of a run of deletions at the beginning or at the end of the sequence X. Let W be the array of partial constrained edit distances (see [14]), where only deletions and substitutions are used. In that case, W is 2-dimensional with $N - M$ rows and M columns. Then the formula for the computation of the constrained edit distance with the constraints given above and with an arbitrary length of a run of deletions at the beginning is the following:

$$
\begin{aligned}
W[0,0] &= 0 \\
W[0,s] &= W[0,s-1] + d(X_s, Y_s), 1 \le s \le M \\
W[e,0] &= e, 1 \le e \le N - M \\
W[e,1] &= \min\{W[e-e_1,0] + e_1 d_e + d(X_{e+1}, Y_1)\} \\
&\quad 1 \le e \le N - M, 0 \le e_1 \le E \\
W[e,s] &= \min\{W[e-e_1, s-1] + e_1 d_e + d(X_{e+s}, Y_s)\}
\end{aligned}
$$

$$\max\{0, e - \min\{N - M, (s-1)E\}\} \le e_1 \le \min\{e, E\}$$

$$2 \le s \le M, 1 \le e \le \min\{N - M, N - s, sE\} \qquad (4)$$

B. The modified attack

In the modified attack, to speed up the phase using the computation of the constrained edit distance, we use the bit-parallel constrained approximate search. By means of search, we try to find the best approximate embedding of the intercepted noised output sequence of the generator from Fig. 4 into the output sequence of the autonomous LFSR_2. This procedure reduces the time complexity of this phase of the attack from quadratic to approximately linear in the length of the intercepted output sequence.

In order to use the bit-parallel search in this attack, we have to introduce the constraints defined above in the update formula. The modification of the update formula is necessary since if we would use the original unconstrained search procedure to find the best embedding, we would get many false positives since the attack would accept the impossible embeddings with more than E skipped bits from the output sequence of the autonomous LFSR_2. Because of that, we have to limit the lengths of runs of deletion transitions in the machine from Fig. 3 to E. In addition, to keep up with the ordering of the edit operations performed by the BRM, we have to eliminate the possibility of deletion at the end of the edit transformation. This is achieved by means of two

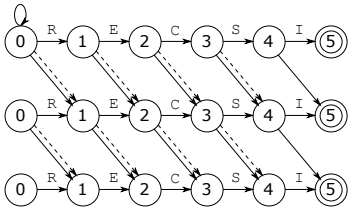


Fig. 6. The modified NFA from Fig. 3 (see text).

actions: in the search status array update formula we introduce an additional bit mask; we also use the previous status of the immediately preceding row in the search status array to compute the contribution of a deletion. The modified NFA is presented in Fig. 6. Note that before the last position in each row of the NFA there is no ε-transition.

The new search status update formula for the array R, which takes into account the constraints and the ordering of edit operations is given below (equation (5)). It is valid for $E = 1$, which is the case that we study in this paper. For a higher E, counters of deletions have to be assigned to the nodes of the NFA, which is more complicated to implement.

$$dm_i = 1^m, i = 1, \ldots, k$$

$$R_0^{'} = (R_0 << 1) \text{ OR } B[S_j]$$

$$del = 1^m$$

$$R_i^{'} = \text{match AND sub AND del}$$

$$\begin{aligned} \text{match} &= (R_i << 1) \text{ OR } B[S_j] \\ \text{sub} &= ((R_{i-1} << 1) \text{ OR NOT } B[S_j]) \\ \text{del} &= (R_{i-1} << 1) \text{ OR} \\ &\quad \text{NOT } ((\text{del} << 1) \text{ OR } 0^{m-1}1) \text{ OR} \\ &\quad \text{NOT } ((\text{dm}_{i-1} << 1) \text{ OR } 0^{m-1}1) \text{ OR} \\ &\quad (0^{m-1}1 << (m-1)) \end{aligned}$$

$$dm_i^{'} = \text{del}$$

$$i = 1, \ldots, k \qquad (5)$$

It is easy to notice that the part of the update formula (3) corresponding to the match and substitution was not changed in the update formula (5), since no constraints have been introduced on substitutions. Regarding the deletion, that part of the update formula (3) had to be extended with constraints on the contribution of the deletions from the previous row of both the current (the variable del) and the previous version (the array dm) of the search status array. In addition, to cancel the contribution of the deletions at the end, the MSB of the total contribution of the deletions is masked with $0^{m-1}1 << (m-1)$.

Since the alphabet used does not influence the operation of the NFA simulation, we could use the update formula (5) directly in the attack instead of computing the edit distance by means of dynamic programming that was used in [6]. The best embedding is obtained for the initial states of LFSR_2 for which the formula (5) gets the MSB of the lowest row of the search status array R equal to 0, which corresponds to the smallest constrained edit distance between the output sequence of the autonomous LFSR_2 for that initial state and the intercepted noised output sequence.

IV. Discussion and preliminary experimental results

In the original attack, to enable the second phase of the attack, in which the initial state of the control register LFSR_1 is reconstructed, it is necessary to maintain the matrix of the partial constrained edit distances W. Namely, the clock control sequence reconstruction procedure requires backtracking through that matrix. Thus, in the attack, we cannot use the memory-minimalistic variant of the constrained edit distance computation algorithm involving only two neighboring columns of the matrix W (see, for example, [12]).

In the case of the proposed modification of the attack, the backtracking through the NFA is possible and the memory consumption of the attack remains only $c(k+1)$ memory cells, where c is a small constant. The memory consumption increases linearly with the length of the intercepted noised sequence, since we have to make the used memory cells longer. On commodity hardware this would mean concatenation of ordinary memory cells (of length typically 32 or 64 bits).

Regarding time consumption, it is expected that the computation that involves only bit operations and execution of many of them in parallel computes the constrained edit distance faster than in the case of ordinary constrained edit distance computation. Preliminary experiments show small differences in execution time, but these experiments have been performed on short intercepted output sequences. With increased length of the intercepted output sequence, which is necessary for good discrimination between the statistical hypotheses (see [6]), we expect positive influence of parallelization from one hand, but negative influence of computer word concatenation on the other hand. This study is the topic of the future work.

V. Conclusions

In this paper, we have modified the generalized correlation attack on irregularly clocked LFSRs by replacing the constrained edit distance computation procedure with a bit-parallel constrained search algorithm that tries to find the best embedding of the intercepted noised output sequence of the generator in the output sequence of the controlled LFSR when it runs autonomously, i.e. without irregular clocking. The modified attack consumes less memory and the preliminary experimental results obtained on relatively short intercepted sequences show that the modified attack is not slower than the original one. Studying practical aspects of time consumption of the modified attack with longer intercepted output sequences is necessary in order to reveal its full potential. This will be the subject of the future work.

References

[1] R. Baeza-Yates and G. Gonnet: "A new approach to text searching", *Communications of the ACM*, Vol. 35, No. 10, pp. 74–82, 1992.

[2] G. Navarro and M. Raffinot: "Fast and flexible string matching by combining bit-parallelism and suffix automata", *ACM Journal of experimental algorithms*, Vol. 5, No. 4, pp. 1–36, 2000.

[3] B. Durian, J. Holub, H. Peltola and J. Tarhio: "Tuning BNDM with q-grams", *Proceedings of the Workshop on Algorithm Engineering and Experiments, ALENEX 2009*, pp. 29–37, 2009.

[4] D. Knuth, J. Morris and V. Pratt: "Fast pattern matching in strings", *SIAM J. Comput.*, Vol. 6, No. 2, pp. 323–350, 1977.

[5] A. Aho and M. Corasick: ”Efficient string matching: an aid to bibliographic search”, *Communications of the ACM*, Vol. 18, No. 6, pp. 333–340, 1975.
[6] J. Golić and M. Mihaljević: ”A generalized correlation attack on a class of stream ciphers based on the Levenshtein distance”, *J. Cryptology*, Vol. 3, No. 3, pp. 201–212, 1991.
[7] E. Zenner: ”On Cryptographic Properties of LFSR-based Pseudorandom Generators”, PhD thesis, University of Mannheim, Germany, 2004.
[8] H. Hyyrö: ”Explaining and extending the bit-parallel approximate string matching algorithm of Myers”, *Technical report, University of Tampere*, Finland, 2001.
[9] G. Myers: ”A fast bit-vector algorithm for approximate string matching based on dynamic programming”, *J. ACM*, Vol. 46, No. 3, pp 495–515, 1999.
[10] A. Chitrakar and S. Petrović: ”Constrained row-based bit-parallel search in intrusion detection”, *Proceedings of the Norwegian Information Security Conference (NISK 2016)*, pp. 68–79, Bergen, 2016.
[11] J. Holub: ”Simulation of Nondeterministic Finite Automata in Pattern Matching”, PhD thesis, Czech Technical University, Prague, Czech Republic, 2000.
[12] G. Navarro and M. Raffinot, ”Flexible Pattern Matching in Strings”, Cambridge University Press, 2002.
[13] W. Chambers and S. Jennings: ”Linear equivalence of certain BRM shift register sequences”, *Electronics Letters*, Vol. 20, No. 24, pp. 1018-1019, 1984.
[14] B. Oommen: ”Constrained string editing”, *Information Sciences*, Vol. 40, No. 3, pp. 267–284, 1986.

Hacia una Arquitectura de Referencia Segura para Big Data

Julio Moreno
Grupo de investigación GSyA
Universidad de Castilla-La Mancha, España
Julio.Moreno@uclm.es

Manuel A. Serrano
Grupo de investigación Alarcos
Universidad de Castilla-La Mancha, España
Manuel.Serrano@uclm.es

Eduardo Fernández-Medina
Grupo de investigación GSyA
Universidad de Castilla-La Mancha, España
Eduardo.FdezMedina@uclm.es

Eduardo B. Fernandez
Departamento de computación e Ingeniería Eléctrica y Ciencias de la computación
Florida Atlantic University, EEUU
Fernande@fau.edu

Abstract- **Las compañías son conscientes de la importancia de Big Data y los datos como forma de realizar sus actividades diarias, pero existen problemas de seguridad y privacidad en Big Data. La seguridad es crucial en Big Data, pero desafortunadamente, Big Data no fue concebido como un entorno seguro. Además, esta tarea es difícil por las diferentes configuraciones que Big Data puede tener. Para resolver este problema se debe tener una perspectiva global, utilizar una Arquitectura de Referencia (RA) es una forma de implementar sistemas complejos. Existen diferentes propuestas de RA para Big Data, pero ninguna se centra en la seguridad. Centrarse en seguridad añadiendo elementos para abordar las amenazas es una buena forma de resolver amenazas, convirtiendo las RA en Arquitecturas de Referencia Segura (SRA). Este artículo define una SRA para Big Data utilizando modelos UML intentado facilitar la implementación segura de Big Data y permitiendo aplicar patrones de seguridad.**

***Index Terms*- Big Data, Arquitectura de Referencia Segura, Patrones de seguridad**

I. INTRODUCCIÓN

Las compañías son cada vez más conscientes de la importancia de Big Data [1]. Para ellos, los datos son esenciales para llevar a cabo sus actividades diarias y para ayudar a la alta gerencia a alcanzar los objetivos de negocio, y en consecuencia, a tomar mejores decisiones basadas en la información extraída de dichos datos [2]. Big Data implica un cambio respecto a las técnicas tradicionales de análisis de datos en tres características principalmente: la cantidad de datos (volumen), el ratio de generación y transmisión de los datos (velocidad) y la heterogeneidad de los tipos de datos estructurados y no estructurados que pueden ser gestionados (variedad) [3]. Estas propiedades son conocidas como las tres Vs de Big Data.

Normalmente las nuevas tecnologías suponen la aparición de nuevos problemas, y Big Data no es una excepción. Estos problemas se encuentran relacionados no solo con las tres Vs de Big Data, sino que también surgen en los ámbitos de privacidad y seguridad [4]. Estos problemas de seguridad suceden por el hecho de que Big Data no fue inicialmente concebida como un entorno seguro [5], y por ello, los principales problemas de seguridad se relacionan con la arquitectura propia de Big Data que hace que sea más complicado proteger la privacidad de los datos usados [6].

Obtener un nivel adecuado de seguridad puede influenciar en su implementación en una institución debido a la pérdida de reputación que pueden sufrir en caso de recibir una multa administrativa, debido a la legislación vigente, en el caso de producirse un filtrado de datos; de hecho, si no se garantiza la seguridad, Big Data no puede alcanzar un nivel apropiado de aceptación [7]. Por consiguiente, es importante contar con guías, metodologías y mecanismos que permitan implementar de forma apropiada tanto el sistema Big Data como su seguridad. Los entornos Big Data suelen ser muy complejos, por ello, para abordar su seguridad es necesario empezar desde una perspectiva global. La seguridad debe ser abordada a través de políticas de alto nivel que puedan ser mapeadas a los niveles inferiores [8]. Diferentes autores [9], [10] destacan que una Arquitectura de Referencia (RA) ha demostrado ser una destacable herramienta para guiar la seguridad en diferentes entornos, por ejemplo, en Cloud Computing [8] o en Internet de las Cosas [11].

Una RA es una arquitectura abstracta de software que se basa en uno o más dominios y que no tiene detalles de implementación [9]. Además, una RA debe ser expresada con un alto nivel de abstracción, con el objetivo de ser reutilizable, extensible y configurable. Este tipo de arquitectura puede estar compuesta por diferentes patrones para facilitar la implementación del sistema y mejorar la inclusión de requisitos no funcionales [12]. Si se añaden patrones de seguridad para controlar amenazas identificadas, se convierte una RA en una Arquitectura de Referencia Segura (SRA). Una SRA es una arquitectura de alto nivel que incorpora una serie de elementos para facilitar la definición de requisitos de seguridad y permite un mejor entendimiento de conceptos como las políticas de seguridad, las amenazas, las vulnerabilidades, etc. y que puede ser usada para describir un modelo conceptual de seguridad para sistemas Big Data [13].

Así, como primer paso para mejorar la seguridad y confianza en entornos Big Data hemos creado una SRA específica para este tipo de tecnologías. Para ello, hemos destacado la importancia de los patrones de seguridad para facilitar la implementación de mecanismos de seguridad en un

ecosistema Big Data. Para ello, hemos realizado una modificación de la RA propuesta por el Instituto Nacional de Estándares y Tecnologías (NIST) para Big Data, con el objetivo de crear una arquitectura más detallada, en la cual, las relaciones entre las diferentes partes del Big Data sean claramente expuestas. Además pondrá el foco en asegurar el entorno Big Data. Esta SRA mejorada permitirá un mejor entendimiento del ecosistema Big Data. Para conseguirlo, nuestra SRA se encuentra especificada usando diagramas UML [14]. Finalmente, junto a nuestra SRA hemos creado un ejemplo parcial que muestra cómo aplicar nuestra arquitectura; para ello, hemos considerado diferentes amenazas que pueden afectar a un sistema Big Data, y cómo los diferentes componentes que se involucran en su prevención son instanciados; por ejemplo, los patrones de seguridad pueden ayudar en la solución de dichos problemas.

Nuestro artículo se organiza de la siguiente forma: primero, mostramos una sección que explica las principales características de la propuesta de RA del NIST. Después, se presentan los componentes y estructura de nuestra SRA junto con un ejemplo de cómo usar los patrones de seguridad para afrontar amenazas de un entorno Big Data en concreto. A continuación, presentamos una comparación de nuestra propuesta con otras propuestas de RA para Big Data. Finalmente, incluimos una sección de conclusiones y trabajo futuro.

II. Modelo de referencia: Arquitectura de Referencia del NIST para Big Data

En los últimos años, la organización NIST ha definido una RA para Big Data que ha recibido el consenso general por parte de la industria y la comunidad científica [15]. Esta arquitectura recopila diferentes ideas y características para crear un ecosistema Big Data. Este conjunto de características fueron extraídas de las diferentes propuestas de arquitectura para Big Data hechas por las principales compañías del sector, como Oracle o IBM. Como resultado, el NIST creó la RA que se puede observar en la Figura 1. Su propuesta se divide en cinco componentes que pueden interactuar entre ellos y cuentan con diferentes objetivos. Estos componentes son:

- System Orchestrator (SO): Es uno de los componentes más importantes de un ecosistema Big Data, ya que, se encuentra a cargo de definir e integrar las actividades y los datos requeridos en el ecosistema. Su propósito principal es configurar y gestionar los otros componentes de la arquitectura Big Data, para ello, supervisa los requisitos y restricciones del sistema; por ejemplo, políticas, arquitectura o requisitos de negocio.
- Data Provider (DP): Este componente supervisa la introducción de nuevos datos al ecosistema Big Data. Para conseguir este objetivo, este componente cuenta con una serie de interfaces, o servicios, entre el sistema y las fuentes de datos. Este conjunto de interfaces actúa como puertas entre el mundo exterior y el sistema Big Data.
- Big Data Application Provider (BDAP): El componente BDAP prove de una serie de servicios específicos a lo largo del ciclo de vida de los datos para cumplir con los requisitos establecidos por el SO. Es importante destacar que su propósito principal es encapsular la lógica de negocio y su funcionalidad para ser ejecutados por la arquitectura. En un escenario normal de Big Data, existen varias aplicaciones usando al mismo tiempo los mismos datos. Mientras que los datos se propagan a través del ecosistema, estos van siendo transformados y procesados de diferentes formas para obtener información valiosa. Para alcanzar dicho objetivo, el BDAP se compone de diferentes servicios o actividades que se pueden considerar la capa SaaS de un sistema Big Data. Estas actividades son: colección, preparación, análisis, visualización y acceso. Estas actividades pueden ser implementadas como funciones independientes y desplegadas como servicios autónomos. Además estas actividades pueden interactuar con la capa inferior de Big Data Framework Provider, con el DP, con el Data Consumer o incluso entre ellas.
- Big Data Framework Provider (BDFP): Este componente se considera como la plataforma de implementación de la lógica del sistema Big Data. Soporta las actividades definidas por el BDAP. En general, las implementaciones de Big Data son híbridos que combinan múltiples tecnologías. Tiene tres sub-componentes principales: infraestructura (virtual o física), plataforma (cómo los datos se distribuyen y organizan) y procesamiento (cómo los datos son procesados para soportar las aplicaciones de Big Data). El BDFP también provee de soporte a servicios auxiliares para el sistema como las comunicaciones o la gestión de recursos.
- Data Consumer (DC): Es similar al componente DP. Normalmente el actor que interactúa con este componente es un usuario final de otro sistema. Como el DP, se encuentra compuesto de un conjunto de interfaces entre el usuario final y la información.

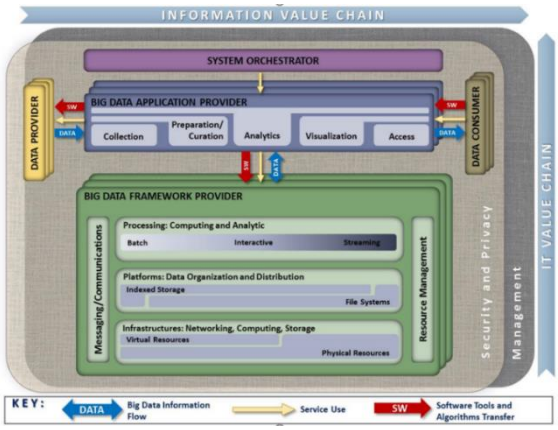


Fig. 1. Propuesta del NIST para una arquitectura Big Data [15].

La propuesta del NIST no puede ser considerada una SRA, sin embargo, reconoce la importancia de la seguridad y la privacidad en un entorno Big Data. Para afrontar los problemas de seguridad, esta arquitectura dispone de un componente específico para tratar la seguridad y privacidad.

Desde nuestro punto de vista, esta representación basada en bloques no es lo suficientemente expresiva. Este tipo de especificación es demasiado abstracta y hace poco énfasis en los detalles de los subcomponentes y su relación entre ellos. Este enfoque puede dificultar el diseño e implementación de

un ecosistema Big Data. Siguiendo la misma aproximación, la organización ISO/IEC se encuentra trabajando en la creación de una RA para Big Data bajo el estándar 20547-3. Aunque se trata de un trabajo en progreso, se espera que tenga una perspectiva similar a la propuesta del NIST.

III. Arquitectura de Referencia Segura (SRA) para Big Data

En esta sección, describiremos nuestra propuesta de SRA, la cual, se estructura siguiendo como base el mismo esquema y componentes de la propuesta del NIST. Consideramos que si nuestra SRA se encuentra alineada con dicha RA, será más fácil de implementar. Además esta arquitectura destaca la importancia de implementar soluciones de seguridad basadas en conceptos de la SRA. La Figura 2 muestra un diagrama de clases con todos los componentes que forman nuestra arquitectura y la relación entre ellos.

A. System Orchestrator (SO)

El objetivo de este componente es garantizar la aplicación de los diferentes requisitos que el sistema Big Data debe cumplir. También organiza cómo se conectan los requisitos con todos los componentes de la arquitectura; en esta sección, nos centraremos en los requisitos de seguridad y su relación entre ellos y los diferentes componentes. Debido a las características de este componente, sus actividades de seguridad relacionadas, en general, se centran en los requisitos: cómo implementarlos y monitorizarlos. Estos requisitos deben cumplir con los objetivos de Big Data y deben estar alineados con los diferentes objetivos de negocio y políticas de la compañía. En este aspecto, el rol del administrador de seguridad es crucial para asegurar la observación de los requisitos de seguridad. Estos requisitos de seguridad deben cumplir con las regulaciones que afectan al contexto de cada ecosistema Big Data. De hecho, existen muchos tipos de requisitos diferentes para cumplir las necesidades de un ecosistema Big Data; por ejemplo, requisitos de arquitectura, de calidad, o de gobierno.

Existen múltiples ejemplos de requisitos de seguridad que deben ser abordados en un contexto Big Data. Temas como la privacidad de los datos y cómo asegurar la propia arquitectura Big Data son los más investigados por la comunidad científica [16]. Estos problemas pueden ser atacados siguiendo mecanismos generales como la autorización y autenticación de usuarios, la detección de fraude, el control de riesgo, la auditoría o las técnicas de encriptación [17]. Estos son mecanismos generales de seguridad pero deben ser modificados para ser aplicados a los distintos tipos específicos de sistema y a sus posibles amenazas.

Tal y como se observa en la Figura 2, estos requisitos de seguridad pueden ser satisfechos por medio de diferentes soluciones de seguridad que siguen políticas de seguridad de la compañía y tienen como principal objetivo abordar las amenazas para controlar las vulnerabilidades. Un ejemplo de una política de seguridad en una compañía puede ser la obligación de usar comunicaciones seguras, esta política puede derivar en la creación de un requisito de seguridad en el entorno Big Data que especifique que los datos transferidos entre componentes deben estar asegurados. Una forma de cumplir requisitos de seguridad es mediante soluciones de seguridad, las cuales, deben ser implementadas en los componentes en un nivel inferior de la arquitectura, en los componentes BDAP y BDFP. Sin embargo, estas soluciones de seguridad no son fáciles de implementar; por ello, nuestro modelo contempla la posibilidad de usar patrones de seguridad como guía. Un patrón de seguridad es una solución a un problema recurrente que indica cómo defenderse contra una amenaza, o un conjunto de amenazas, de una forma concisa y reusable [18]. Los patrones son soluciones abstractas que deben adaptarse al lugar donde se apliquen. Además, es posible usar patrones de mal uso [19] como forma de entender cada ataque y guiar la aplicación de diferentes patrones de seguridad para defenderse de una amenaza. Por otro lado, los metadatos de seguridad pueden ser definidos como una forma de facilitar la coordinación y realización de los requisitos de seguridad. Otro tema que se encuentra cubierto por nuestra arquitectura es el contexto en el que se encuentra el activo; por ejemplo, las consideraciones de seguridad que debe tener un registro médico son totalmente distintas a aquellas que debe tener un archivo de log. Es importante evaluar el nivel de seguridad que debe tener cada recurso.

B. Data Provider (DP)

El componente DP crea una abstracción de las fuentes de datos considerando sus metadatos de seguridad, si existen. Estos metadatos permiten al DP identificar los tipos de acceso y análisis permitidos por las fuentes de datos y sus requisitos de seguridad. Tal y como se explicó en la sección 2, el DP cuenta con una serie de interfaces. Dichas interfaces deben considerar las restricciones de cada una de las fuentes de datos y también las diferentes políticas de seguridad y requisitos expresados por el SO. En este elemento, pueden existir conflictos entre los requisitos de seguridad de la fuente de datos y los del sistema Big Data. Estos conflictos deben ser gestionados de forma que satisfagan a ambos bandos. Los problemas de seguridad y privacidad de este componente están normalmente relacionados con cómo identificar y validar las entradas de las fuentes. Las interfaces del DP deben evaluar el origen de las fuentes de datos. Es un desafío crítico en la colección de datos el conocer cómo validar que una fuente de datos no es maliciosa y filtrar aquellas que sí lo son [6].

En nuestra SRA, las interfaces se encuentran conectadas con el servicio “Colector” del componente BDAP que será descrito en la siguiente subsección. En general, los elementos que componen una fuente de datos incluyen: los datos que pueden ser estructurados, no estructurados o semiestructurados; los requisitos de seguridad de la fuente de datos; y los metadatos de seguridad de la fuente de datos. Dichos elementos no se representan en el diagrama de la Figura 2, debido a que se consideramos las fuentes de datos como un agente externo al sistema Big Data. Aun así, es importante conocerlos para aplicar sus restricciones.

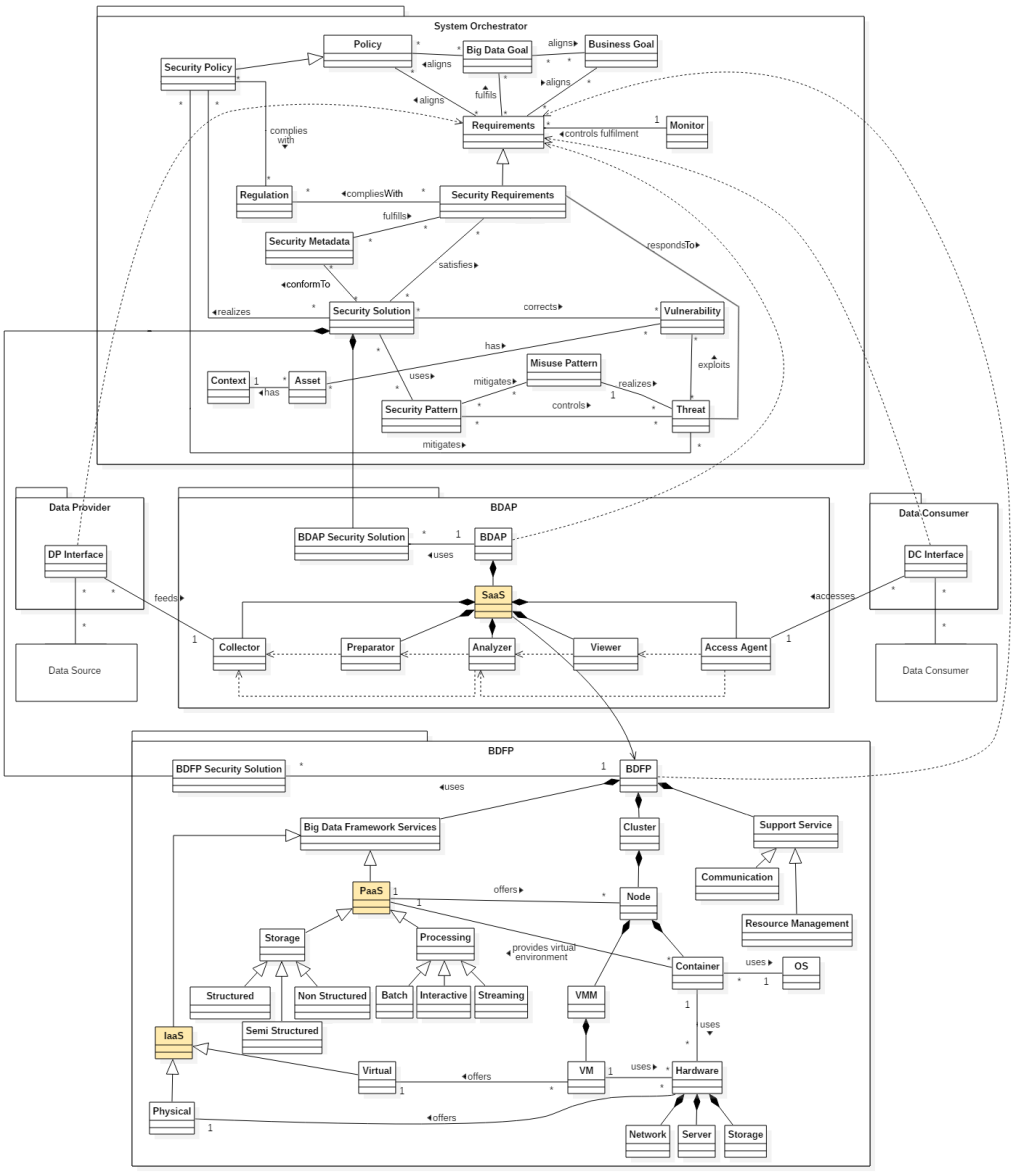


Fig. 2. Diagrama de SRA para Big Data.

C. Big Data Application Provider (BDAP)

El componente BDAP tiene el objetivo de cumplir los requisitos establecidos por el SO, incluyendo los requisitos de seguridad y privacidad. Para conseguirlo, el BDAP se compone de diferentes servicios o actividades que pueden ser considerados como la capa SaaS (Software como servicio) del ecosistema Big Data; en nuestro caso, asumimos que, en general, los sistemas Big Data se implementan en una plataforma en la nube, lo cual, afectará a cómo definimos la SRA en el componente BDFP. En la Figura 2, se pueden observar los diferentes servicios que constituyen este componente, y también la solución de seguridad para BDAP, la cual, se trata de una instanciación más específica de la definida en el SO; por ejemplo, un mecanismo de autorización puede controlar quién puede realizar qué operaciones para hacer un análisis de los datos.

Tal y como se representa en el diagrama, no todas las actividades pueden comunicarse entre sí, existe un orden secuencial de ejecución. Esto significa que algunas de estas actividades no son obligatorias en un ecosistema Big Data. La fase de preparación tiene el propósito de validar, limpiar y almacenar los datos, sin embargo, en un escenario en tiempo real donde los datos deben ser analizados al momento que entran al sistema, esta actividad puede no realizarse Algo similar ocurre con la fase de visualización, ya que, si el consumidor de los datos no es un humano sino otro sistema, como un almacén de datos, esta actividad podría no ser necesaria.

Por otro lado, las otras tres actividades son básicas en un ecosistema Big Data: la actividad de colección actúa como un proceso ETL (Extracción, Transformación y Carga) y combina conjuntos de datos de estructura similar con el objetivo de unificarlos; la fase de análisis incluye un conjunto de técnicas para obtener conocimiento valioso de los datos; por ejemplo, los algoritmos MapReduce, y finalmente, la actividad de acceso tiene el propósito de comunicarse con el DC, actuando como una interfaz entre el DC y las actividades de visualización y análisis. La relación mediante estas actividades se observa en la Figura 2 mediante líneas discontinuas, puesto que es una relación de uso temporal.

D. Big Data Framework Provider (BDFP)

En general, el componente BDFP se encuentra compuesto por un conjunto de *clusters,* que a su vez, se componen de nodos. Estos nodos pueden ser desplegados por medio de máquinas virtuales o contenedores, los cuales, interactúan con el hardware y el sistema operativo.

NIST define este componente de forma muy abstracta, falto de detalles en los subcomponentes necesarios para llevar a cabo los procesos. Por ello, nuestra propuesta hace más énfasis en los diferentes elementos y en cómo se conectan entre sí. Nuestra SRA hace hincapié en la idea de un ecosistema Big Data con la posibilidad de implementar el sistema mediante un entorno de Cloud y técnicas de visualización. Respecto a los problemas de seguridad, en este componente las actividades pueden centrarse en la encriptación y la gestión de los datos, el aislamiento y uso de contenedores del proceso de ejecución, la autorización, la autenticación, los logs de auditoría y cómo asegurar el almacenamiento y la red. Esos problemas de seguridad deben ser resueltos mediante soluciones de seguridad definidas en el nivel del SO, que pueden ser implementadas en este nivel como soluciones de seguridad del BDFP.

E. Data Consumer (DC)

El componente DC está, al igual que el DP, compuesto por un conjunto de interfaces. La interacción puede incluir visualización interactiva, creación de informes, o análisis de los datos usando técnicas de inteligencia de negocio. Es importante destacar que estas interfaces deben cumplir con las funciones de autorización y autenticación para alcanzar el objetivo de que el DC junte los metadatos relacionados con el usuario final y los requisitos de seguridad y políticas de la información.

F. Ejemplos de aplicación de patrones de seguridad

Como forma de mostrar la utilidad de nuestra SRA, explicaremos un ejemplo de cómo emplear los patrones de seguridad utilizando nuestra arquitectura. Hemos creado un ejemplo que identifica algunas de las amenazas que pueden ser encontradas en las diferentes actividades del componente BDAP. Dichas amenazas pueden ser afrontadas mediante el uso de patrones de seguridad, los cuales en algunos casos, pueden ser modificados desde patrones de seguridad generales para cumplir con las características inherentes de Big Data. La modificación de dichos patrones y la creación de nuevos patrones, si es necesario, se encuentra fuera del propósito de este artículo y es considerado trabajo futuro. La Tabla I resume algunas de las amenazas de cada actividad y los patrones generales que pueden ser aplicados para resolverlos. Dichos patrones se encuentran definidos en [18]. Como una forma de entender mejor cómo integrar los diferentes componentes de nuestra SRA y los patrones de seguridad, definiremos cómo la amenaza TC1 puede ser afrontada mediante el uso de patrones de seguridad. Para explicarlo, usaremos un diagrama de objetos que se puede ver en la Figura 3. En este escenario, tenemos como activo principal a proteger los datos almacenados, dicho activo tiene una vulnerabilidad: no tiene protección, esta vulnerabilidad puede ser explotada mediante una amenaza como TC1. Para prevenir dicha situación es necesario implementar una solución de seguridad. Para facilitar la implementación de esta solución, se pueden utilizar dos patrones de seguridad: Control de acceso basado en roles y Autenticación. Sin embargo, esta solución de seguridad seguirá teniendo un alto nivel de abstracción pues que se define en el componente SO. Por ello, una implementación de bajo nivel de la solución de seguridad debe ser realizada en el nivel de BDAP, en este caso, la amenaza TC1 puede afectar diferentes servicios proporcionados por el BDAP. Esta es la razón por la cual la solución de seguridad debe ser implementada aquí y no en otro componente.

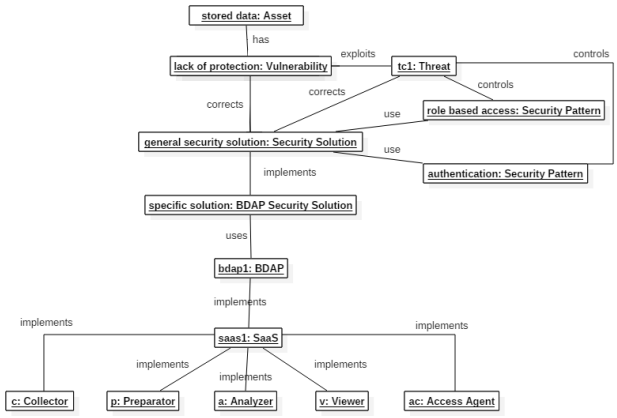


Fig. 3. Uso de patrones de seguridad para abordar una amenaza

IV. COMPARACIÓN CON OTRAS PROPUESTAS

No existen demasiadas arquitecturas de referencia para sistemas Big Data; si nos centramos en la seguridad, hay incluso menos. Sin embargo, diferentes autores y organizaciones han propuesto diferentes arquitecturas de referencia para Big Data. En esta sección, definimos algunas de las más relevantes. Demchenko et al. [20] proponen una arquitectura que establece el ciclo de vida de los datos en ecosistemas Big Data. Como en la propuesta del NIST, utilizan una representación basada en bloques; pero con un mayor detalle en las relaciones entre los diferentes componentes de la arquitectura. Sin embargo, ellos abordan la seguridad de una forma muy escasa y como si fuera algo aislado, no conectado con el resto de componentes. En [21], los autores proponen una arquitectura completa en cuanto a relaciones entre diferentes componentes; sin embargo, encontramos una falta de consideración en cuanto a aspectos de seguridad y privacidad. Klein et al. proponen en [22] un arquitectura de referencia específica para Big Data en el dominio de la seguridad nacional. Su arquitectura es bastante similar a la propuesta por el NIST. Nuestro objetivo es obtener una mejor abstracción de la arquitectura, pero sigue siendo interesante cómo aborda algunos problemas utilizando patrones de solución. Ellos destacan la importancia de tener un dominio específico para los requisitos. En nuestro caso, los requisitos son la parte principal del componente SO. Además de estas propuestas, diferentes compañías han realizado propuestas como Oracle [23], IBM [24] o Microsoft [25].

V. CONCLUSIONES Y TRABAJO FUTURO

En este artículo presentamos nuestra SRA que considera las diferentes propuestas de RA publicadas, incluyendo las

propuestas del NIST, Oracle, y diferentes investigadores. Para definirla, hemos utilizado diagramas UML como forma de intentar facilitar la implementación segura de Big Data. Este tipo de diagramas ha sido utilizado debido a que algunas propuestas no presentan de forma explícita la relación entre los diferentes componentes. Además, gracias a estos diagramas es posible aplicar diferentes patrones de seguridad, puesto que suelen ser definidos siguiendo modelos UML. Como forma de explicar la utilidad de nuestra propuesta, hemos creado un ejemplo, en el cual se listan una serie de amenazas que pueden afectar a un ecosistema Big Data; sin embargo, es necesario un entendimiento más profundo de estas amenazas. Para solucionar este problema, tenemos la intención de crear diferentes casos de uso y escenarios para identificar sus amenazas siguiendo el método de [19]. Una vez tengamos las amenazas identificadas, buscaremos, adaptaremos o crearemos diferentes patrones de seguridad para afrontarlas. Estos serán los siguientes pasos para completar nuestra SRA.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el proyecto SEQUOIA (Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional FEDER, TIN2015-63502-C3-1-R).

REFERENCIAS

[1] J. Akoka, I. Comyn-Wattiau, y N. Laoufi, «Research on Big Data – A systematic mapping study», *SI New Model. Big Data*, vol. 54, n.º Part 2, pp. 105-115, nov. 2017.

[2] V. Mayer-Schönberger y K. Cukier, *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt, 2013.

[3] M. Chen, S. Mao, y Y. Liu, «Big data: A survey», *Mob. Netw. Appl.*, vol. 19, n.º 2, pp. 171–209, 2014.

[4] H. Wang, X. Jiang, y G. Kambourakis, «Special issue on Security, Privacy and Trust in network-based Big Data», *Inf. Sci. Int. J.*, vol. 318, n.º C, pp. 48–50, 2015.

[5] P. P. Sharma y C. P. Navdeti, «Securing big data hadoop: a review of security issues, threats and solution», *Int J Comput Sci Inf Technol*, vol. 5, 2014.

[6] B. D. W. G. Cloud Security Alliance (CSA), «Expanded Top Ten Big Data Security and Privacy», abr-2013. [En línea]. Disponible en: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf. [Accedido: 11-nov-2017].

[7] B. Thuraisingham, «Big data security and privacy», en *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*, 2015, pp. 279–280.

[8] E. B. Fernandez, R. Monge, y K. Hashizume, «Building a security reference architecture for cloud systems», *Requir. Eng.*, vol. 21, n.º 2, pp. 225-249, jun. 2016.

[9] P. Avgeriou, «Describing, Instantiating and Evaluating a Reference Architecture: A Case Study», *Default J.*, 2003.

[10] N. Medvidovic y R. N. Taylor, «Software architecture: foundations, theory, and practice», en *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 2*, 2010, pp. 471-472.

[11] S. Krco, B. Pokric, y F. Carrez, «Designing IoT architecture (s): A European perspective», en *Internet of Things (WF-IoT), 2014 IEEE World Forum on*, 2014, pp. 79-84.

[12] E. B. Fernandez, N. Yoshioka, H. Washizaki, y M. H. Syed, «Modeling and Security in Cloud Ecosystems», *Future Internet*, vol. 8, n.º 2, p. 13, abr. 2016.

[13] F. Liu *et al.*, «NIST cloud computing reference architecture», *NIST Spec. Publ.*, vol. 500, n.º 2011, p. 292, 2011.

[14] J. Rumbaugh, I. Jacobson, y G. Booch, *Unified modeling language reference manual, the*. Pearson Higher Education, 2004.

[15] NBD-WG, NIST, «NIST Big Data Reference Architecture», 2015. [En línea]. Disponible en: https://bigdatawg.nist.gov/_uploadfiles/M0639_v1_9796711131.docx. [Accedido: 18-oct-2017].

[16] Julio Moreno y Manuel A. Serrano; Eduardo Fernández-Medina, «Main Issues in Big Data Security», *Collab. Technol. Syst. CTS 2013 Int. Conf. On*, pp. 42–47, may 2013.

[17] E. Bertino, «Big Data - Security and Privacy», en *2015 IEEE International Congress on Big Data*, 2015, pp. 757-761.

[18] E. B. Fernandez, *Security patterns in practice: designing secure architectures using software patterns*. John Wiley & Sons, 2013.

[19] E. B. Fernandez, N. Yoshioka, y H. Washizaki, «Modeling misuse patterns», en *Availability, Reliability and Security, 2009. ARES'09. International Conference on*, 2009, pp. 566-571.

[20] Y. Demchenko, C. De Laat, y P. Membrey, «Defining architecture components of the Big Data Ecosystem», en *Collaboration Technologies and Systems (CTS), 2014 International Conference on*, 2014, pp. 104–112.

[21] P. Pääkkönen y D. Pakkala, «Reference architecture and classification of technologies, products and services for big data systems», *Big Data Res.*, vol. 2, n.º 4, pp. 166-186, 2015.

[22] J. Klein, R. Buglak, D. Blockow, T. Wuttke, y B. Cooper, «A reference architecture for big data systems in the national security domain», en *Proceedings of the 2nd International Workshop on BIG Data Software Engineering*, Austin, Texas, 2016, pp. 51-57.

[23] D. Cackett, «Information Management And Big Data A Reference Architecture», *Oracle Febr.*, 2013.

[24] IBM Corporation, «IBM Big Data & Analytics RA», 2014.

[25] Microsoft, «Microsoft Big Data Solution Brief». .

.

Tabla I. Amenazas identificadas y patrones de seguridad para las diferentes actividades

ID	Actividad	Amenaza	Patrones de seguridad
TC1	Común a todas las actividades	Datos modificados	Autenticación, Control de acceso basado en roles
TC2	Común a todas las actividades	Datos destruidos	Autenticación, Control de acceso basado en roles
TC3	Común a todas las actividades	Data leídos ilegalmente	Encriptación, Autenticación, Control de acceso basado en roles
TCo1	Colección	Fuente de datos maliciosa	Autenticación
TP1	Preparación	Filtro malicioso	Logger y Auditor, Control de acceso a la sesión, Autenticación, Control de acceso basado en roles
TA1	Análisis	Inferir PII desde datos anonimizados	Encriptación, Logger y Auditor, Seguridad multinivel, Autenticación, Control de acceso basado en roles
TA2	Análisis	Algoritmos de análisis maliciosos	Logger y Auditor, Control de acceso a la sesión, Autenticación, Control de acceso basado en roles
TV1	Visualización	PII* expuestos debido a la alta granularidad gráfica	Seguridad multinivel, Autenticación, Control de acceso basado en roles
TAc1	Acceso	Múltiples accesos maliciosos	Autenticación, Control de acceso basado en roles

*PII – Información Personal Identificable

Sistema Basado en Lógica Difusa para la Detección del Robo de Identidad en Redes Sociales

José Á. Concepción-Sánchez, Pino Caballero-Gil, Iván Santos-González
Departamento de Ingeniería y Sistemas
Universidad de La Laguna, Tenerife, España
{jaconcep, pcaballe, jsantosg}@ull.edu.es

***Resumen*—Las redes sociales son muy populares en el mundo actual y poseen una gran importancia en la interconexión entre las personas. Es por ello que el robo de las credenciales de acceso para acceder a las cuentas de los usuarios está a la orden del día. Muchos usuarios son atacados por ciberdelincuentes de diversas formas para intentar apoderarse de sus usuarios y contraseñas y así poder obtener información de éstos, reclamar rescates por sus cuentas o para realizar técnicas como el spam o el phishing entre sus contactos. Con el fin de detectar el robo de identidad, en este documento se propone el uso de un sistema basado principalmente en lógica difusa para decidir si una persona está siendo víctima de este problema. Para ello, se tiene en cuenta el lenguaje natural de cada usuario, considerando valores como la frecuencia total de aparición de palabras y expresiones no frecuentemente usadas por ellos. Otros factores como la ubicación desde la cual los usuarios inician sesión o el número de mensajes enviados en un corto espacio de tiempo también son tenidos en cuenta para tomar la decisión final.**

***Index Terms*—Robo de Identidad, Seguridad, Lógica Difusa, Redes Sociales**

I. Introducción

La ciberdelincuencia es una nueva amenaza que tiene su raíz en el auge de las tecnologías en los últimos años. Una de sus formas se conoce como robo de identidades. El robo de identidad se puede definir como el uso deliberado de la identidad de otra persona para acometer acciones en su nombre y que, en muchas ocasiones, es en perjuicio o detrimento de la víctima [1] [2]. Entre las finalidades más comunes de los robos de identidad se encuentran los relacionados con el dinero, como los fraudes de tarjetas de crédito o cuentas bancarias, o también los relacionados con las redes sociales, un medio que permite llegar a muchas personas.

El robo de identidad está a la orden del día en todo el mundo. Por ejemplo, en el Reino Unido se estima que más de 100.000 personas se ven afectadas por este problema cada año [3]. Este otro estudio [4] también revela que 15,4 millones de personas fueron afectadas en todo el mundo en 2016, siendo el récord histórico hasta la fecha. En el caso de las redes sociales, donde el 86 % de la población española entre los 16 y 65 años [5] se encuentra registrada, cobra especial importancia debido a que es un medio que permite extender virus, malware o captar nuevas víctimas aprovechándose de la identidad de las cuentas ya robadas. Es por ello que se considera necesario el desarrollo de un sistema que permita detectar en estas plataformas cuando un usuario ha sido víctima de este problema para poder solventarlo lo más rápidamente posible.

Por ello, en este documento se propone el uso de un sistema compuesto por mecanismos de minería de datos y lógica difusa diseñado para la toma de una decisión final sobre si la cuenta de un usuario puede estar siendo comprometida, lo que permitirá mejorar la calidad del servicio y la seguridad en las redes sociales. Para ello, el sistema tiene en cuenta el vocabulario que cada usuario utiliza normalmente así como también otros factores como la ubicación desde la que los usuarios están conectados o la creación de eventos masivos como publicaciones o mensajes en la red social en un corto periodo de tiempo.

La estructura de este documento es la siguiente. En primer lugar, en la Sección 2 se mencionan algunos trabajos relacionados con esta propuesta. En segundo lugar, en la Sección 3 se describe el funcionamiento del sistema propuesto basado en lógica difusa para la detección del robo de identidad. Y, en tercer lugar, la Sección 4 cierra este documento con las conclusiones y algunos trabajos futuros.

II. Trabajos Relacionados

Debido a la importancia que tiene el proteger las cuentas de usuario se han realizado diversos estudios, de los que se siguen varios consejos que permiten aumentar la seguridad de dichas cuentas. Por ejemplo, organizaciones como OWASP [6] dan una serie de indicaciones para que las cuentas de usuarios sean seguras. Por otro lado, en [7] también se describen una serie de reglas a nivel de usuario y de desarrolladores que permiten evitar el engaño en las redes sociales. A modo de ejemplo y entre otras acciones, se recomienda el uso de técnicas de análisis de textos como el procesamiento de lenguaje natural, técnica de la que hace uso este trabajo por medio de la minería de datos.

En cuanto a sistemas para la detección del robo de identidades en redes sociales, se han encontrado distintos trabajos [8] [9]. En particular, el documento [10] propone un sistema que pre-procesa texto en bruto y extrae información semiestructurada automáticamente, utilizando minería de datos. Otras investigaciones como [11] [12] también proponen el uso de técnicas de minería de datos para la identificación del comportamiento de los usuarios y para la obtención de información de la web respectivamente.

Por otro lado, se han encontrado muy pocas investigaciones que hagan uso de lógica difusa. Este trabajo [13] defiende que los factores de autenticación tienen sus vulnerabilidades y propone un protocolo de autenticación difusa utilizando interacciones pasadas y políticas lingüísticas junto con una evaluación experimental extensiva. Otros trabajos como [14] describen un sistema de minería de datos borroso para la detección de phishing en páginas web de e-banking. Sin embargo, no se ha encontrado ninguna propuesta que haga

uso de lógica difusa o una técnica similar para predecir si la identidad de un usuario pudo haber sido robada.

Por ello, a diferencia de otros trabajos, esta propuesta destaca por el uso de un sistema de lógica difusa que a su vez se apoya en el procesamiento de texto mediante minería de datos para detectar casos de robo de identidad en redes sociales. Gracias a esta metodología, el sistema se puede basar en un conjunto de valores de entrada para deducir finalmente la probabilidad de que la identidad en una cuenta de un usuario esté siendo suplantada. Además, este sistema se puede definir como global, ya que las entradas en el sistema de lógica difusa no están diseñadas para una red social en concreto, sino que son valores que la mayoría de las plataformas podrían obtener fácilmente como son, por ejemplo, las publicaciones, mensajes, geolocalización, etc.

III. Sistema propuesto

El creciente uso de las redes sociales en la sociedad es una realidad. Prueba de ello son los millones de usuarios que tienen creada una cuenta en alguna red social como Facebook, Twitter o Linkedin entre otras muchas. Es por eso que muchos ciberdelincuentes han visto a estas plataformas como un camino fácil de extorsionar a las personas, robando sus identidades, aprovechándose de sus datos, e incluso haciéndose pasar por las propias víctimas para compartir campañas con algún tipo de malware. De esta forma, en este documento se propone el uso de un sistema que puede ser implementado en cualquiera de estas plataformas, permitiendo detectar el robo de identidad. Para ello, el sistema está compuesto por un sistema de lógica difusa apoyado en minería de datos para generar un conjunto de entradas que permiten deducir si realmente se está produciendo un caso de robo de identidad. La Figura 1 muestra un esquema de cómo está compuesto el sistema propuesto:

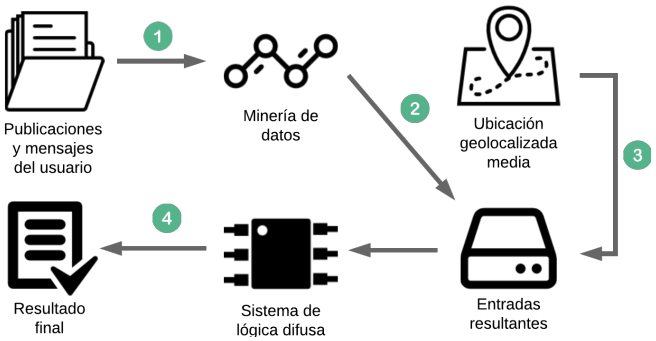


Figura 1. Esquema del sistema propuesto.

Cada uno de estos pasos representa:

1. Uso de técnicas de minería de datos para la obtención de palabras y expresiones a partir de las publicaciones y mensajes del usuario.
2. Análisis de las palabras y expresiones resultantes en busca de coincidencias con el vocabulario normalmente usado con el usuario.
3. Cálculo de la posición geolocalizada media del usuario.
4. Uso del sistema de lógica difusa para la comprobación de la probabilidad de que exista un robo de identidad en la cuenta analizada.

A continuación, cada uno de los procesos son detallados de manera más profunda.

III-A. Minería de datos

La minería de datos [16] se refiere al proceso de recoger información no trivial de conjuntos de textos en busca de patrones que den algún tipo de información. Las tareas más comunes en minería de datos incluyen la categorización y agrupación de texto, la extracción de conceptos/entidades y el resumen de documentos. En este trabajo, la minería de datos es usada para analizar los últimos mensajes y publicaciones creados por el usuario en busca de algún patrón.

Para ello, primero se preprocesan los datos para luego buscar coincidencias entre el conjunto de tokens resultantes y las dos bases de datos usadas.

III-A1. Preprocesamiento: El preprocesamiento de los datos es usado para eliminar factores dependientes de cada idioma en el textos. Ejemplos pueden ser los artículos, preposiciones u otras palabras que generalmente se utilizan pero no tienen más sentido que el de conectar frases o un conjunto de palabras. Para ello, la tokenización [17] es una de las técnicas más usadas para el procesamiento de textos. Este método consiste en dividir un texto de entrada, como una frase, en un conjunto de palabras y expresiones. Una vez dividido el texto en tokens, todas las palabras innecesaria son eliminadas.

III-A2. Búsqueda de coincidencias: Usando el conjunto de tokens resultantes, el sistema propuesto hará una búsqueda de posibles coincidencias entre los tokens y las dos bases de datos (véase Figura 2).

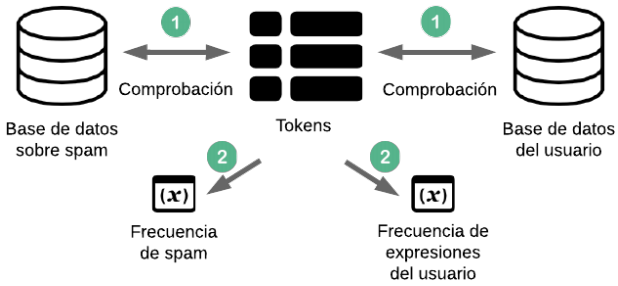


Figura 2. Método para la obtención de frecuencias de aparición.

Estas bases de datos se corresponden con:

- Una base de datos general que es usada independientemente del usuario que está siendo analizado y que contiene un conjunto de palabras y expresiones que suelen ser usadas frecuentemente en campañas de spam.
- Una base de datos específica de cada usuario que contiene las palabras y expresiones más usadas por cada uno de ellos. Esta base de datos se alimenta con los tokens creados en el paso previo. Además, para evitar que se estén guardando todos los tokens desde el registro del usuario en la plataforma, en la base de datos se asigna una fecha de caducidad para cada token, mediante la cual será eliminado si no ha vuelto a ser usado después de un tiempo determinado.

De este paso se obtienen dos de las variables lingüísticas que son usadas por el sistema de lógica difusa, correspondientes a la frecuencia de aparición en las bases de datos.

Por último, en cuanto al inicio en frío, en inglés *cold start* [18], es un problema que afecta principalmente cuando no existe una cantidad suficiente de datos, lo que repercute negativamente en los resultados finales. Debido a esto, en la base de datos que contiene las palabras y expresiones relacionadas con cada usuario, se ha establecido un mínimo de manera que hasta que no contenga un conjunto de datos razonable, esta variable no sea utilizada por el sistema de lógica difusa. De este modo, se consigue evitar que esta variable pueda provocar falsos positivos cuando aún no existen suficientes datos.

III-B. Ubicación geolocalizada media del usuario

La ubicación geográfica es otro mecanismo útil para la detección del robo de identidad. Ejemplo de ello puede ser cuando se inicia sesión en una cuenta de usuario desde un lugar poco frecuente, alejado de la zona donde suele encontrarse el usuario. Por ello, otra de las variables de las que hace uso el sistema de lógica difusa para la detección del robo de identidad es la distancia actual respecto a la ubicación media en la que se encuentra el usuario.

Para su cálculo se realiza un promedio de las latitudes/longitudes, lo que equivale a encontrar el punto medio en un mapa plano de proyección rectangular. Este método es usado debido a su sencillez y a que en distancias de hasta 400km dan unos resultados muy parecidos a otros métodos existentes.

Una vez obtenida la ubicación media del usuario, se procede finalmente al sistema de lógica difusa, quién será el encargado de tomar la decisión final de si realmente se está produciendo un caso de robo de identidad para la cuenta de usuario que está siendo analizada.

III-C. Lógica difusa

La lógica difusa [19] es una metodología que proporciona una forma sencilla y elegante de obtener una conclusión a partir de información de entrada vaga, ambigua o incompleta. En este trabajo, este mecanismo, basado en el método de inferencia difusa de Mamdani [20], es el responsable de tomar la decisión final sobre si realmente se está produciendo un robo de identidad en el caso que está siendo analizado. Los pasos que se llevan a cabo en esta metodología son los siguientes:

En primer lugar, la fuzzificación, que se encarga de transformar las entradas del sistema (valores nítidos) en conjuntos difusos. Para la detección del robo de identidad, este sistema hace uso de un total de seis variables lingüísticas:

- *Total Posts* (TP): Número total de publicaciones que se han creado en la cuenta del usuario analizado en la última hora. Un número muy alto de publicaciones en un corto periodo de tiempo podría ser una forma de hacer spam a los contactos del usuario.
- *Total Messages* (TM): Número total de contactos a los que se les ha enviado algún mensaje en la última hora. Al igual que la variable lingüística anterior, un número alto de mensajes a distintas personas en un corto periodo de tiempo podría significar una acción de spam.
- *Link Detection* (LD): Generalmente las campañas de spam suelen incluir algún enlace que redirecciona a sus víctimas a páginas web maliciosas o ficheros infectados. De esta manera, esta variable indica la existencia de algún token en forma de url.
- *Distance* (DIST): Indica la distancia que existe desde la última ubicación conocida del usuario y la media de sus ubicaciones calculada en el paso previo.
- *User Frequency* (UF): Frecuencia de aparición de palabras y expresiones utilizadas por el usuario en las publicaciones y mensajes analizados.
- *Spam Frequency* (SF): Frecuencia de aparición de palabras y expresiones comúnmente utilizadas en mensajes de spam.

Cada una de estas variables lingüísticas tienen asociados a su vez una serie de términos lingüísticos que representan la probabilidad de que pueda existir un caso de robo de identidad (HIGH, MEDIUM y LOW), exceptuando la variable lingüística LD, cuyos términos (YES y NO) indican la existencia o no de urls en el texto analizado. En la Figura 3 se muestra un ejemplo de la gráfica donde para la variable lingüística TP son representados los valores de las funciones de pertenencia de sus términos lingüísticos.

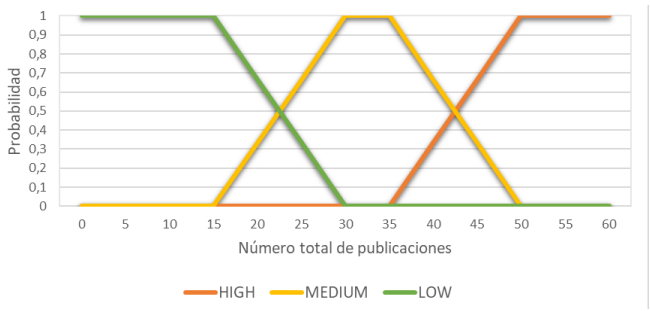


Figura 3. Representación de los términos lingüísticos de la variable TP.

Una vez fuzzificados los valores de entrada, el siguiente paso corresponde al motor de inferencia. Este paso simula el razonamiento humano haciendo inferencia sobre las entradas y un conjunto de reglas. Estas reglas pertenecen a una base de conocimiento y son de estructura IF-THEN. A continuación, un ejemplo de algunas reglas de la base del conocimiento son mostradas, donde RESULT es la variable lingüística de salida y cuyos términos lingüísticos son YES y NO.

Algorithm 1: Ejemplo de un conjunto de reglas.

Input: TP, TM, LD, DIST, UF y SF (fuzzificados)
Output: YES y NO (fuzzificados)

if $((TM == HIGH)$ *or* $(TP == HIGH))$ *and* $(SF == HIGH)$ *and* $(LD == YES)$*;*
then
  $RESULT = YES$;
end
if $(UF == LOW)$ *and* $(SF == LOW)$ *and* $(TP == LOW)$ *and* $(TM == LOW)$*;*
then
  $RESULT = NO$;
end

Según las reglas de ejemplo descritas, si el valor de TP o TM fue fuzzificado como HIGH, fue detectado algún enlace

(LD fuzzificada como YES) y la frecuencia de spam fue alta (SF fuzziciado como HIGH), el sistema creará un nuevo conjunto borroso de salida con resultado YES. En caso de que se cumpla más de una regla, se combinarán todos los conjuntos borrosos en un único conjunto utilizando la función de agregación AND.

Finalmente, la defuzzificación es el último paso en el sistema de lógica difusa. Este paso realiza el proceso de adaptación de los valores difusos generados en valores precisos. Una vez defuzzificado, el valor final indicará la probabilidad de que el caso analizado pueda estar relacionado con un robo de identidad.

Existen numerosos métodos que se pueden utilizar para realizar este paso, aunque finalmente se decidió utilizar la técnica del Centroide. Esta técnica calcula el valor absoluto encontrando el valor del Centro de Gravedad (COG), una categorización numérica del área total del gráfico y siendo el resultado final del sistema de lógica difusa.

Una vez obtenido el valor final, la aplicación o plataforma que haga uso de este sistema se encargaría de interpretar los resultados y, si es necesario, tomaría las decisiones oportunas para solventar el problema. En el caso contrario, si no se detecta un caso de robo de identidad, el análisis serviría de forma preventiva.

IV. Conclusiones

El robo de identidad es un problema que afecta globalmente a la sociedad. Muestra de ello son las estadísticas que indican cómo cada año el número de casos aumenta. Uno de los lugares más comunes donde se realizan estos robos es en las redes sociales, debido a que es un medio que permite a los ciberdelincuentes extorsionar a sus víctimas con su información, o compartir campañas de malware y virus entre los contactos de la misma. Sin embargo, los trabajos relacionados con este problema están enfocados a detectarlo en entornos muy específicos así como en el estudio de los efectos que provoca en sus víctimas.

Así, a diferencia de otras propuestas, este documento propone el uso de un sistema basado en lógica difusa para la detección del robo de identidad en las redes sociales. Para ello, hace uso previamente de técnicas de minería de datos para buscar patrones en los textos analizados de cada individuo y tiene en cuenta datos la localización media del usuario y el comportamiento del usuario usando el número total de mensajes y publicaciones creados en un espacio de tiempo determinado. Entre las ventajas que ofrece el uso de lógica difusa para la detección del robo de identidad se encuentra la velocidad de cómputo debido a que los conceptos matemáticos en los que se basa son simples, la flexibilidad para mejorar la precisión añadiendo nuevas reglas o variables sin tener que eliminar el diseño previo y la posibilidad de uso de datos imprecisos manejando conjuntos difusos.

Finalmente, como trabajos futuros se estudia el uso de nuevas metodologías que permitan mejorar la precisión en la toma de decisiones. Para ello, un ejemplo pueden ser las redes neuronales, que podrían permitir que para cada caso a analizar, el sistema se pudiera adaptar a las particularidades del individuo. Además, también se pretende trabajar en una aplicación móvil que haciendo uso de un sistema de detección de robo similar al propuesto en este documento, pueda detectar cuando alguien se apropia de su dispositivo móvil.

Agradecimientos

Investigación apoyada por el Ministerio de Economía y Competitividad, TESIS2015010106, el Fondo Europeo FEDER y la Fundación CajaCanarias, en el marco de los proyectos TEC2014-54110-R, RTC-2014-1648-8, MTM2015-69138-REDT y DIG02-INSITU.

Referencias

[1] Synthetic ID Theft, https://web.archive.org/web/20151009122632/http://www.unc.edu/~dubal/idtheft/synthetic.htm\#top, (Accessed on 19 February 2018).
[2] C. J. Hoofnagle (2007). Identity theft: Making the known unknowns known. Harv. JL & Tech., 21, 97.
[3] Report on identity theft/fraud prevention expert group. http://ec.europa.eu/internal_market/fpeg/docs/id-theft-report_en.pdf, (Accessed on 16 April 2018).
[4] C. Warren (October 13, 2017). How Common Is Identity Theft? The Latest Stats. https://www.lifelock.com/education/how-common-is-identity-theft/, (Accessed on 16 April 2018).
[5] R. Sotelo (April 19, 2017). VIII Estudio anual de redes sociales en España 2017. https://marketing4ecommerce.net/viii-estudio-anual-de-redes-sociales-en-espana-2017/, (Accessed on 16 April 2018).
[6] OWASP, https://www.owasp.org/index.php/Main_Page, (Accessed on 20 February 2018).
[7] M. Tsikerdekis and S. Zeadally (May 22, 2015). Detecting and Preventing Online Identity Deception in Social Networking Services in Internet Computing, vol. 19, no. 3, pp. 41-49.
[8] K. Veena and K. Meena (2017). Determination of performance to verify the synthetic identity theft by training the neural networks, International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Chennai, pp. 246-250.
[9] B. Z. He, C. M. Chen, Y. P. Su and H. M. Sun (2014). A defence scheme against identity theft attack based on multiple social networks. Expert Systems with Applications, 41(5), 2345-2352.
[10] R. N. Zaeem, M. Manoharan, Y. Yang and K. S. Barber (2017). Modeling and analysis of identity threat behaviors through text mining of identity theft stories. Computers & Security, 65, 50-63.
[11] J. Peng, K. K. R. Choo, and H. Ashman (2016). Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles. Journal of Network and Computer Applications, 70, 171-182.
[12] Z. Xu, H. Zhang, C. Hu, L. Mei, J. Xuan, K. K. R. Choo and Y. Zhu (2016). Building knowledge base of urban emergency events based on crowdsourcing of social media. Concurrency and Computation: Practice and Experience, 28(15), 4038-4052.
[13] R. Khan, and R. Hasan (2015, June). Fuzzy authentication using interaction provenance in service oriented computing. In Services Computing (SCC), 2015 International Conference on (pp. 170-177).
[14] M. Aburrous, M. A. Hossain, K. Dahal and F. Thabtah (2010). Intelligent phishing detection system for e-banking using fuzzy data mining. Expert systems with applications, 37(12), 7913-7921.
[15] M. Aburrous, M. A. Hossain, F. Thabatah and K. Dahal (2008, April). Intelligent phishing website detection system using fuzzy techniques. In Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on (pp. 1-6).
[16] M. A. Hearst (1997, July). Text data mining: Issues, techniques, and the relationship to information access. In Presentation notes for UW/MS workshop on data mining (pp. 112-117).
[17] Text tokenization & multiwords, https://www.meaningcloud.com/developer/resources/doc/models/models/text-tokenization-multiwords, (Accessed on 17 April 2018).
[18] H. Gaspar (July 14, 2015). The Cold Start Problem for Recommender Systems. http://www.yusp.com/blog/cold-start-problem-recommender-systems/, (Accessed on 17 April 2018).
[19] L. A. Zadeh (1988). Fuzzy logic. Computer, 21(4), 83-93.
[20] E. H. Mamdani and S. Assilian (1975). An experiment in linguistic synthesis with a fuzzy logic controller. International journal of man-machine studies, 7(1), 1-13.

Marisma-CPS: Una Aproximación al Análisis y Gestión de Riesgos para Sistemas Ciberfísicos

David G. Rosado
Instituto de Tecnologías y Sistemas de la Información (ITSI)
Universidad de Castilla-La Mancha
Ciudad Real, España
David.GRosado@uclm.es

Julio Moreno
Grupo de Investigación GSyA
Universidad de Castilla-La Mancha
Ciudad Real, España
Julio.Moreno@uclm.es

Antonio Santos-Olmo
Departamento de I+D+i
Sicaman Nuevas Tecnologías S.L.
Tomelloso, España
asolmo@sicaman-nt.com

Luis E. Sánchez
Departamento de I+D+i
Marisma Shield S.L.
Tomelloso, España
lesanchez@emarisma.com

Manuel A. Serrano
Grupo de Investigación Alarcos
Universidad de Castilla-La Mancha
Ciudad Real, España
Manuel.Serrano@uclm.es

Eduardo Fernández-Medina
Grupo de Investigación GSyA
Universidad de Castilla-La Mancha
Ciudad Real, España
Eduardo.FdezMedina@uclm.es

Abstract- **Los sistemas ciberfísicos (Cyber-physical systems - CPS) son sistemas inteligentes que afectan a casi todos los aspectos de nuestra vida cotidiana. Los CPS son altamente dinámicos, heterogéneos, ampliamente interconectados y con necesidad de respuesta en tiempo real, y aunque la evaluación de riesgos de sistemas TI tradicionales se puede considerar bastante madura, surge la necesidad de un método de evaluación de riesgos específico para las características de los CPS, dando cobertura a los crecientes problemas de seguridad debidos a las diferencias entre sistemas TI tradicionales y CPS. En este artículo se propone una técnica de análisis y gestión de riesgos para CPS, relacionada con una metodología de gestión de seguridad (MARISMA), apoyada por un entorno tecnológico en la nube (eMARISMA) que ya utilizan docenas de clientes. Esta metodología y su entorno están diseñados para adaptarse fácilmente a contextos dinámicos y heterogéneos como CPS. Nuestra propuesta, denominada MARISMA-CPS, se basa en los marcos principales, tales como NIST o ENISA.**

***Index Terms*- Análisis y gestión de riesgos, Sistemas Ciberfísicos, MARISMA.**

I. Introducción

Los sistemas ciberfísicos son sistemas inteligentes que incluyen capacidades de computación, almacenamiento y comunicación con capacidades de seguimiento y/o control de objetos en el mundo físico [1] y proveen un amplio rango de aplicaciones y servicios innovadores a ciudadanos y empresas [2]. Los CPS cubren desde comunicaciones Maquina-a-Maquina (M2M) e Internet de las Cosas (IoT), integración de datos heterogéneos de múltiples fuentes, hasta su integración dentro de plataformas Cloud Computing y Big Data [3]. La investigación acerca de sistemas ciberfísicos (CPS) está llamando la atención de universidad, industria y gobierno dado su alto impacto social, económico y ambiental [4-8].

El surgimiento y avances de nuevos sistemas integrados inteligentes, sensores y dispositivos embebidos con conexión constante a Internet (lo que generalmente se conoce como Internet de las Cosas (IoT)), junto con las capacidades de detección y actuación permiten escenarios sin precedentes de interacción entre los mundos real y virtual. Esto creará una inundación de información del mundo real, enriquecerán considerablemente las aplicaciones, haciéndolas más conscientes de lo que sucede en el mundo real, en tiempo real, en todas partes.

Mientras que el actual trabajo de investigación se centra en el logro de objetivos como la estabilidad, rendimiento, robustez y eficiencia de los sistemas físicos [7], la ciberseguridad en CPS es ignorada [4], a pesar de que muchos CPS se integran en infraestructuras críticas, y las consecuencias de algún fallo de seguridad puede tener consecuencias catastróficas.

La ciberseguridad es una característica necesaria de la arquitectura CPS para asegurar que las capacidades de los CPS no se ven comprometidas por agentes maliciosos, y que la información utilizada, procesada, almacenada y transferida conserva su integridad y confidencial donde sea necesario [5].

Aunque los CPS y los sistemas de TI están íntimamente relacionados son todavía diferentes en diversos aspectos, especialmente en el dinamismo de los CPS, así como en otros requisitos tales como el tiempo real, el cambio continuo y la heterogeneidad de los actores. Los métodos estáticos de evaluación sólo pueden estimar el riesgo de forma aproximada en un período de tiempo, pero no pueden evaluar el riesgo con precisión en puntos de tiempo concretos. Para ser eficaz, una evaluación del riesgo debería ser capaz de prever y considerar los posibles sistemas que puedan surgir antes de la siguiente evaluación periódica. Esto supone un desafío importante, y los enfoques actuales normalmente no lo tienen en cuenta [9].

De esta forma, el método tradicional de evaluación de riesgos de sistemas TI no se puede aplicar directamente a los CPS, por lo que estos sistemas se enfrentan a un riesgo de seguridad enorme. Aunque hay una evidente madurez en la evaluación tradicional de riesgos sobre sistemas de TI, se necesita un método diferente de evaluación de riesgos para CPS de forma que se cubran los crecientes problemas de seguridad debidos a las grandes diferencias entre sistemas TI y CPS.

Aunque existen numerosas propuestas para abordar de

forma sistemática el análisis y gestión de riesgos de TI, tales como MAGERIT [10], OCTAVE [11], CRAMM [12], estas propuestas presentan frecuentemente dificultades para su implantación práctica en sistemas altamente dinámicos como CPS. O bien no cuentan con herramientas adecuadas para su tratamiento, o bien las que usan no son demasiados usables, ya que están diseñadas para aplicarse en grandes compañías y no son sensibles al contexto (sin permitir capacidad de adaptación a entornos especiales, que requieren un tratamiento especial de sus riesgos). Para cubrir estas carencias se desarrolló la metodología MARISMA (Methodology for the Analysis of Risks on Information System, using Meta-Pattern and Adaptability) [13], así como un entorno tecnológico llamado eMARISMA que la soporta (www.emarisma.com). MARISMA es una metodología basada en la reutilización del conocimiento para los procesos de análisis y gestión de riesgos mediante el uso de estructuras llamadas patrones que permiten soportar distintos tipos de casos, ayudando a reducir considerablemente el esfuerzo del proceso. El entorno tecnológico eMARISMA es una plataforma en la nube que implementa los procesos de MARISMA, permitiendo automatizar el proceso de análisis y gestión de riesgos, apoyar la reutilización y permitiendo la actualización en tiempo real de indicadores de riesgo.

De esta forma, nuestra propuesta se basa en la creación de un patrón específico para MARISMA (MARISMA-CPS), que permita gestionar y controlar los riesgos en CPS, y que se ha creado en base a las recomendaciones dadas por la norma ISO/IEC 27000, a las recomendaciones de seguridad para IoT propuestas por la Agencia ENISA [14] y en el marco para CPS del NIST [5], que considera las necesidades inherentes a este tipos de sistemas. Cabe destacar que este modelo de riesgos no pretende cubrir todos los riesgos organizativos que pueden afectar indirectamente a los CPS.

El resto de este trabajo está organizado de la siguiente manera: Primero, se realiza una introducción a los antecedentes relacionados la temática, seguido de una sección que presenta las principales características de la metodología, así como el entorno tecnológico que la sustenta. A continuación, se explica MARISMA-CPS, incluido el proceso para crearlo y los diferentes estándares y marcos de referencia que toma en consideración. Finalmente, se incluye una sección de conclusiones y trabajo futuro.

II. ANTECEDENTES

En esta sección se mencionarán los diferentes marcos, metodologías y estándares relacionados con análisis y gestión de riesgos. Por otra parte, también se explicarán las diferentes propuestas relacionadas con CPS en este campo.

Entre las principales propuestas de análisis y gestión de riesgos se pueden resaltar MAGERIT, OCTAVE, CRAMM, así como los estándares ISO/IEC, COBIT o NIST. MAGERIT implementa el proceso de gestión de riesgos dentro de un marco para que los órganos de gobierno tomen decisiones, teniendo en cuenta los riesgos derivados del uso de tecnologías de la información. Por otra parte, OCTAVE es una técnica de planificación estratégica y consultoría en seguridad basada en riesgos. CRAMM es una metodología de gestión de riesgos que consiste en tres fases: identificación, análisis y gestión de riesgos.

Con respecto a los principales estándares de gestión de la seguridad, muchos de ellos han intentado incorporar el análisis y gestión de riesgos en sus procesos: el grupo de normas ISO/IEC 27000, y específicamente la norma ISO/IEC 27005 [15]. ITIL [16] ofrece un elemento para la correcta gestión de riesgos: el conocimiento actualizado y detallado de todos los activos de la organización y de las relaciones, pesos y dependencias entre ellos. COBIT [17] es una metodología para el control adecuado de proyectos de tecnología, flujos de información y los riesgos derivados de la falta de controles adecuados. Incluye un proceso destinado a evaluación de riesgos que se centra principalmente en los criterios de confidencialidad, integridad y disponibilidad, y secundariamente en criterios de eficacia, eficiencia, cumplimiento y confiabilidad. Finalmente, la organización NIST [18] ha propuesto un marco de gestión de riesgos genérico aplicable a cualquier sistema de información.

Respecto al análisis y gestión de riesgos de CPS, cabe destacar la falta de propuestas para sistemas reales. La mayor parte de las propuestas son preliminares y están basadas sólo en la evaluación de riesgos, sin entrar en la gestión de los mismos. Entre estas propuestas preliminares cabe destacar el marco RECYPHR [19], que está basado en la norma ISO 31000 pero carece de implementación o herramientas de apoyo del marco. Además, podemos encontrar varias propuestas genéricas como la taxonomía de riesgos de ciberseguridad operacional compilados por Cebula and Young [20] en base a una propuesta del NIST, pero que es sólo una clasificación de los riesgos y no cuenta con metodologías o herramientas para análisis/gestión de riesgos. Destacando la importancia en este campo podemos encontrar una revisión sistemática sobre seguridad de sistemas ciberfísicos [21], centrada en el análisis y modelado de riesgos.

Existen otras propuestas de evaluación de riesgos en CPS como las de in [22-24], trabajos interesantes y prometedores pero que no están soportados por ninguna norma y son sólo trabajos en fase temprana, y por tanto no adecuados para la gestión y análisis de riesgos reales. A lo largo de estos trabajos preliminares también podemos encontrar algunas propuestas interesantes centradas en entornos diferentes, como las propuestas de evaluación de riesgos en IoT (Internet de las Cosas) [25], sistemas SCADA y fabricación manufacturing [26-28] o Smart Grids (Redes Inteligentes) [29]. La mayoría de los trabajos son soluciones preliminares o parciales, generalmente no fundamentados por una norma, y carecen de metodologías y herramientas para apoyar la gestión y análisis de riesgos en el contexto de CPS.

Finalmente, como forma de resolver la necesidad de llevar a cabo el análisis y gestión de riesgos en CPS, decidimos utilizar la metodología MARISMA, que fue desarrollada tratando de mejorar las propuestas anteriores, y puesto que tiene los patrones como un mecanismo para ser extendido y adaptado a nuevos contextos dinámicos, creemos que es adecuada para nuestro propósito.

III. MARISMA

A lo largo de los últimos años hemos identificado una serie de carencias asociadas a los Análisis de Riesgos que imposibilitaban la eficacia de los mismos, y que estos pudieran generar valor en las compañías. Para solucionar dichos problemas, se desarrolló una metodología denominada

MARISMA, y una herramienta que la soporta denominada eMARISMA (www.emarisma.com).

Entre las principales problemáticas que se identificaron, se pueden destacar: i) Elevado coste y complejidad a la hora de hacer un análisis de riesgos; ii) Su falta de orientación a las PYMES; iii) El resultado no ayudaba a que el siguiente análisis de riesgos fuera más sencillo. Es decir, no se reutilizaba el conocimiento adquirido y iv) Los análisis de riesgos eran estáticos.

MARISMA asocia el análisis y la gestión del riesgo a los controles necesarios para la gestión de la seguridad y consta de tres procesos muy importantes que interactúan entre sí para dar solución a las problemáticas detectadas. El proceso GPRA se encarga de generar y contener los patrones (estructuras de conocimiento que relacionan los elementos necesarios para elaborar un análisis de riesgos), los cuales son instanciados por un cliente a la hora de realizar un análisis de riesgos (proceso GARM). Una vez realizado un análisis de riesgos, este sufrirá eventos de seguridad, que serán gestionados mediante el proceso DRM. Estos eventos de seguridad generaran conocimiento para adaptar los niveles asociados a los elementos del análisis de riesgos, haciendo que se recalcule el riesgo (dinamismo), y permitirá adaptar también los elementos asociados al patrón seleccionado (evolución). Finalmente, la propia evolución del patrón permitirá sugerir mejoras a todos los clientes que usaron ese patrón para analizar sus riesgos.

A. *Patrón en MARISMA*

Uno de los grandes problemas que surgió al estudiar los análisis de riesgos, fue como se podía reutilizar el conocimiento y experiencia que adquirían los consultores al realizar el proceso.

Para solucionar este problema, se requería ver que elementos componían los diferentes análisis de riesgos, con independencia de las normativas que se utilizarán y que relaciones debían establecerse entre ellos, con el objetivo de crear una meta-estructura que pudiera contenerlos. De este análisis se identificó que todos los análisis de riesgos tenían controles, activos de información, amenazas y que estos elementos estaban relacionados entre sí.

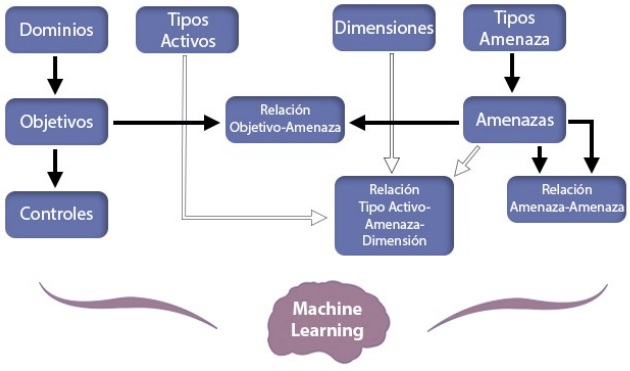


Fig. 2. Esquema general de un patrón de MARISMA

Cada conjunto de elementos que se crea recibe el nombre de patrón (Ej: El patrón de Gestión de la Seguridad, está formado por elementos de la ISO27001 y MAGERIT). La estructura que contendrá todos estos patrones es un meta-patrón que se denominó C.A.T (Control – Assets – Theats), al ser estos los tres elementos base de un análisis de riesgos. Dicho meta-patrón será la estructura común a todos los patrones. En la Fig. 2, podemos ver los elementos básicos que contienen el meta-patrón CAT de MARISMA.

Una vez definidos estos elementos se crea un patrón para realizar un análisis de riesgos dentro de una compañía.

B. *Herramienta eMARISMA*

Para dar soporte a la metodología MARISMA, se desarrolló en modelo SaaS y con tecnología Java una herramienta denominada eMARISMA (www.emarisma.com), que permite soportar todos los procesos que componen la metodología. Esta herramienta ha permitido desarrollar ya patrones para gestión de la seguridad o infraestructuras críticas, que han sido validados con más de 20 compañías de España y Colombia. En la Fig. 3, se puede ver una captura de pantalla de la herramienta.

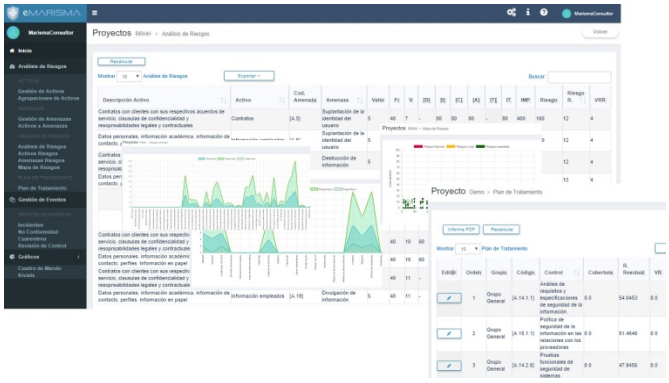


Fig. 3. Herramienta eMARISMA. Generador de Análisis de Riesgos. Creación del Análisis y el Plan de Tratamiento del Riesgo.

De esta forma, se ha desarrollado un proceso de muy bajo coste de mantenimiento para la compañía y que le permite tener un sistema de análisis de riesgos completamente dinámico y con la capacidad de informar en todo momento de la situación real de la compañía, reducir el nivel de subjetividad mediante la adaptación de los parámetros del riesgo e incluir la capacidad de no solo tener en cuenta el riesgo interno, sino también el riesgo externo y temporal al que están sometidas las compañías.

La metodología MARISMA permite desarrollar un marco de gestión de la seguridad propuesto por el NIST y especializado en el campo de los CPS. Para ello se definiría un patrón especializado, utilizando el proceso GPRA y él meta-patrón CAT. Se trata de un marco genérico aplicable a cualquier sistema de información.

IV. MARISMA CPS

En esta sección se presentará, en primer lugar, una breve descripción de las normas y marcos de referencia relacionados con CPS. A continuación, se presentarán los diferentes componentes que son parte del patrón específico para la gestión y control de riesgos en CPS, MARISMA-CPS.

A. *Normas y marcos de referencia para CPS*

NIST está estableciendo el PWG CPS (Cyber-Physical Systems Public Working Group) para ayudar a definir y conformar los aspectos clave de los CPS con el fin de acelerar su desarrollo y aplicación en múltiples sectores de nuestra

economía.

CPS son sistemas construidos a partir de la integración transparente de componentes físicos y computacionales. Estos sistemas son el núcleo de nuestras infraestructuras críticas y conforman las bases de nuestros futuros servicios inteligentes. Estos prometen mayor eficiencia y una interacción entre las redes de computadores y el mundo físico, permitiendo avances que redundan en la mejora de la calidad de vida. Los actores claves en CPS han identificado la necesidad de desarrollar un vocabulario de consenso y una arquitectura de referencia para facilitar la interoperabilidad entre sistemas y elementos, así como promover la comunicación entre los diferentes actores de los CPS [5]. Este marco de referencia define como aspecto principal la confiabilidad, que es la probabilidad demostrable de que el sistema funciona según comportamiento definido bajo cualquier conjunto de condiciones, evidenciado por características que incluyen, pero no se limitan a, seguridad operacional (safety), seguridad, privacidad, confiabilidad y resiliencia. Este será el aspecto a tener en cuenta en nuestro enfoque de MARISMA-CPS, que presentamos en este trabajo, y que debe soportar el aspecto de la confiabilidad.

Otra referencia importante a tener en cuenta es el informe de ENISA (European Union Agency for Network and Information Security) titulado "Recomendaciones de seguridad para IoT en el contexto de infraestructuras de información críticas" [14]. El objetivo principal del estudio es ofrecer consejos de seguridad sobre IoT para organizaciones en Europa teniendo en cuenta factores como la complejidad de los activos críticos, ciberamenazas existentes y soluciones para la protección de sistemas IoT y ecosistemas relacionados como son CPS o infraestructuras críticas. IoT es un concepto emergente que describe un gran ecosistema donde los dispositivos interconectados y los servicios recogen, intercambian y procesan datos con el fin de adaptarse dinámicamente a un contexto. IoT está íntimamente ligado a CPS, y en este sentido es un habilitador de Infraestructuras Inteligentes por el hecho de mejorar su calidad de suministro del servicio. Las Infraestructuras inteligentes ofrecen numerosas ventajas, consiguiendo importantes ahorros de coste y mejoras en la eficiencia. Este tipo de entornos basados en datos, alimentados por los dispositivos conectados y la conectividad de redes, se convierten en una nueva vía de ataque por las ciberamenazas. Por lo tanto, ENISA desarrolla directrices para asegurar IoT y CPS ante ciberamenazas, destacando las buenas prácticas de seguridad y proponiendo recomendaciones a los operadores, fabricantes y a las personas que toman decisiones. La identificación de ciberamenazas se propone en una taxonomía [30] donde se define una clasificación de los tipos de amenazas y amenazas a varios niveles. Además de las ciberamenazas, esta taxonomía también contiene amenazas físicas que pueden afectar a los activos de TI. Nuestro enfoque MARISMA-CPS está basada en una particularización del conjunto de amenazas para CPS.

B. Patrón MARISMA-CPS

- Dominios, objetivos y controles: Para definir los diferentes dominios y objetivos de control de nuestro patrón se ha partido de las recomendaciones para IoT de ENISA [14]. Los dominios, objetivos y controles resultantes se muestran (solo un fragmento por limitaciones de espacio) en la Tabla I.

- Tipos de activos: Para especificar los diferentes tipos de activos se ha partido de las Recomendaciones de seguridad para IoT en el contexto de infraestructuras de información críticas realizadas por la agencia ENISA [14]. Este documento muestra un conjunto de tipos de activos típicos en sistemas de IoT, que son un componente esencial para CPS, y que puede cambiar dependiendo de dominios o sectores empresariales. Los principales tipos de activos son: i) Dispositivos: incluyen hardware, software, actuadores y sensores; ii) Dispositivos de Ecosistema: incluyen dispositivos para interactuar con cosas, dispositivos para gestionar Things, sistemas embebidos; iii) Comunicaciones: activos de redes y protocolos; iv) Infraestructura: engloba activos como routers, pasarelas, suministro de energía y elementos de seguridad; v) Plataforma y Backend: incluye activos como servicios basados en web o servicios e infraestructuras en la Nube; vi) Toma de decisiones: engloban algoritmos de minería de datos, así como de procesamiento de datos y computación; vii) Aplicaciones y Servicios: activos relacionados con análisis y visualización de datos, gestión de redes y dispositivos y utilización de dispositivos; y viii) Información/Datos: incluye información almacenada en una base de datos (en reposo), información enviada o intercambiada por una red (en tránsito) e información utilizada por aplicaciones, servicios o elementos IoT (en uso).

- Dimensiones: Para las dimensiones de seguridad del patrón MARISMA-CPS se han considerado los aspectos de confiabilidad, explicado anteriormente, definidos en el marco para CPS publicado por NIST [5].

- Tipos de amenazas y amenazas (ver Tabla II): Para definir los diferentes tipos de amenazas que pueden afectar a los CPS se ha seguido la taxonomía desarrollada por la agencia ENISA, que en [30] clasifica en varios niveles de detalle los tipos de amenazas y amenazas. ENISA identifica los siguientes tipos de amenazas: i) Ataque físico (deliberado/intencionado): Los ataques físicos son acciones ofensivas intencionadas destinadas a destruir, exponer, modificar, desactivar, robar o lograr acceso no autorizado a los activos físicos; ii) Daño involuntario/Pérdida de información o activos de TI: Se refiere a la destrucción, daño o lesión de bienes o personas, y causa una falta o reducción de utilidad por accidente, o centrada en los activos de TI, e implica intención; iii) Desastre: Un desastre es una interrupción severa del funcionamiento de una sociedad; iv) Fallos/Mal funcionamiento: Se define como tal la condición de funcionamiento insuficiente o nulo de un activo; v) Interrupciones: Son cortes inesperados en el servicio, o disminución de la calidad del mismo por debajo de un nivel requerido. Incluye toda clase de activos, incluso recursos humanos; vi) Escuchas/Interceptación/Secuestro: Implica toda clase de acciones con el objetivo de escuchas, interrumpir o

tomar el control de una comunicación de terceros sin consentimiento; vii) Actividad nefasta/Abuso: Se trata de acciones dolosas dirigidas a sistemas TIC, a través de actos maliciosos destinados a un objetivo determinado; y viii) Legal: Pueden ser acciones legales previstas, planeadas o en curso por terceras partes (contratantes o no) basadas en la legislación vigente.

TABLA I

FRAGMENTO DE LA TABLA DE DOMINIOS, OBJETIVOS DE CONTROL Y CONTROLES DEL PATRÓN MARISMA-CPS

Dominio	Objetivos	Controles[1]
Gobierno de la seguridad y gestión de riesgos de los Sistemas de Información	Diseño de la privacidad	GP-PS-08, 09
	Identificación y Evaluación de Riesgos y Amenazas	GP-PS-11, 12
	Capacitación y conciencia sobre seguridad de los recursos humanos	GP-OP-09,10,11
	Relaciones con terceros	GP-OP-12, 13, 14
	Protección de datos y cumplimiento legal	GP-TM-10,11,12,13,14
	Fuerte seguridad y privacidad por defecto	GP-TM-08
	Control de acceso - Seguridad física y ambiental	GP-TM-29, 30
	Interfaces seguras y servicios de red	GP-TM-47
	Monitoreo y auditoría	GP-TM-57
Gestión del ecosistema	Diseño de la Seguridad	GP-PS-01, 02
	Gestión de Confianza e Integridad	GP-TM-07
	Comunicaciones seguras y confiables	GP-TM-42,43
	Interfaces seguras y servicios de red	GP-TM-48,49

GP-PS: Buenas Prácticas - Políticas
GP-OP: Buenas Prácticas - Medidas organizativas, de personas y de procesos
GP-TM: Buenas Prácticas – Medidas Técnicas

[1] Controles definidos por ENISA. https://www.enisa.europa.eu/publications/baseline-security-recommendations-for-iot/at_download/fullReport

Tabla II
TIPOS DE AMENAZAS Y AMENAZAS PARA CPS

Tipos de Amenazas	Amenazas
Ataque físico (AF)	Modificación del dispositivo
	Destrucción del dispositivo (sabotaje)
Daño Involuntario /Pérdida de Info (DI)	Datos / fuga de información sensible
Desastre (De)	Desastre natural
	Desastre ambiental
Fallos/Mal Funcionamiento (F/M)	Vulnerabilidades de software
	Fallo de terceros
Interrupciones (Int)	Fallo de los dispositivos
	Fallo del sistema
	Pérdida de servicios de soporte
	Caída de la red
Escuchas/Interceptación/Secuestro (E/I/S)	Secuestro del protocolo de comunicación
	Reconocimiento de red
	Intercepción de información
	Secuestro de sesión
	Recopilación de información
	Reproducción de mensajes
	Hombre en el medio
Actividades Mal Intencionadas/Abuso (AMI)	Malware
	Kit de Explotación
	Ataques dirigidos
	DDoS
	Falsificación por dispositivos maliciosos
	Ataques a la privacidad
	Modificación de la información
Legal (L)	Violación de reglas y regulaciones /
	Violación de la legislación
	Incumplimiento de los requisitos contractuales
	Abuso de datos personales

Cualquier sistema de información puede utilizar esta categoría de amenazas al tener una visión general. Las diferentes amenazas pertenecientes a estos tipos de amenazas se usan para describir las características específicas de un CPS, donde en necesario indicar sólo los conjuntos de amenazas que pueden afectar a este tipo de sistemas, descartando al resto de amenazas. Las matrices que el patrón MARISMA-CPS define en la Fig.2 están mostradas en las Tabla III y IV donde sólo se muestran ejemplos de fragmentos de los elementos del patrón CPS, debido a las limitaciones de espacio.

Tabla III
FRAGMENTO DE LA MATRIZ DE OBJETIVOS-AMENAZAS DE CPS

Objetivo	Dominio	Tipos de Amenazas							
		AF	DI	De	F/M	Int	E/I/S	AMI	L
Diseño de Seguridad	Gestión del ecosistema							X	
	Arquitectura de seguridad de TI							X	
	Mantenimiento de seguridad de TI							X	
	Seguridad física y ambiental							X	
Diseño de Privacidad	Gobierno de Seguridad y Gestión de Riesgos de Sistemas de Información		X					X	X
Gestión de activos	Arquitectura de seguridad de TI		X				X	X	

Tabla IV
FRAGMENTO DE LA MATRIZ DE TIPOS DE ACTIVOS – AMENAZAS – DIMENSIONES DE MARISMA-CPS

Tipo de amenazas	Amenazas	Activos							
		Dispositivos	Disp. ecosistema	Comunicaciones	Infraestructura	Platafor / Backend	Toma de Decisiones	Aplicac y servicios	Información / Datos
Ataque físico	Modificación del dispositivo	X		X					
	Destrucción del dispositivo	X	X		X	X			
Daño involuntario /Pérdida de información	Datos / fuga de información sensible	X	X			X			X
Desastre	Desastre natural	X	X		X	X			
	Desastre ambiental		X		X	X			

V. CONCLUSIONES Y TRABAJO FUTURO

Los métodos de gestión de riegos tradicionales tienen madurez, pero no son apropiados para CPS debido a las características particulares de estos, siendo sistemas altamente dinámicos y sistemas heterogéneos en los que multitud de

actores (dispositivos, infraestructuras, servicios, etc.) pueden formar parte del sistema en cualquier momento, haciendo que éste cambie muy rápidamente, o sistemas con requisitos de tiempo real que hacen que los métodos de gestión de riesgos tradicionales no sean suficientes para CPS.

En este artículo mostramos cómo se usa la metodología MARISMA, así como la herramienta eMARISMA que la soporta, para generar un patrón de análisis y gestión de la seguridad enfocado a CPS, permitiendo la gestión dinámica del riesgo asociado con los CPS en una compañía.

Como resultado, nuestro trabajo futuro se enfocará a la aplicación en un caso real en algún dominio de negocio como los hospitales inteligentes. MARISMA-CPS se redefinirá y validará con la experiencia de este caso real, y el uso de este patrón nos ayudará a crear sub-patrones de forma automática gracias a la herramienta eMARISMA. Estos nuevos patrones compartirán la misma estructura de MARISMA-CPS, pero estarán especializados en sectores concretos relacionados con CPS como salud, fabricación energía, urbanismo, etc.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el proyecto SEQUOIA (Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional FEDER, TIN2015-63502-C3-1-R), y el proyecto GENESIS (Consejería de Educación, Cultura y Deportes de la Dirección General de Universidades, Investigación e Innovación de la JCCM, SBPLY/17/180501/000202). Agradecemos la ayuda de las compañías Sicaman Nuevas Tecnologías SL (www.sicaman-nt.com) y Marisma Shield SL (www.emarisma.com), que han facilitado el uso de la herramienta eMARISMA.

REFERENCIAS

[1] Mike Burmestera, Emmanouil Magkos, and V. Chrissikopoulos, "Modeling security in cyber–physical systems," International Journal of Critical Infrastructure Protection, vol. 5, no. 3-4, 2012.

[2] European Commission, "Cyber-Physical Systems: Uplifting Europe's Innovation Capacity," Brussels2013.

[3] A. J. Jara, D. Genoud, and Y. Bocchi, "Big data for cyber physical systems: an analysis of challenges, solutions and opportunities," in Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2014 Eighth International Conference on, 2014, pp. 376-380: IEEE.

[4] K.-D. Kim and P. R. Kumar, "Cyber–physical systems: A perspective at the centennial," Proceedings of the IEEE, vol. 100, no. Special Centennial Issue, pp. 1287-1308, 2012.

[5] NIST, "Framework for Cyber-Physical Systems Release 1.0," 2016.

[6] R. Baheti and H. Gill, "Cyber-physical systems," The impact of control technology, vol. 12, pp. 161-166, 2011.

[7] R. R. Rajkumar, I. Lee, L. Sha, and J. Stankovic, "Cyber-physical systems: the next computing revolution," in Proceedings of the 47th Design Automation Conference, 2010, pp. 731-736: ACM.

[8] E. A. Lee, "Cyber physical systems: Design challenges," in Object Oriented Real-Time Distributed Computing (ISORC), 2008 11th IEEE International Symposium on, 2008, pp. 363-369: IEEE.

[9] S. Taubenberger, J. Jürjens, Y. Yu, and B. Nuseibeh, "Problem analysis of traditional it-security risk assessment methods–an experience report from the insurance and auditing domain," in IFIP International Information Security Conference, 2011, pp. 259-270: Springer.

[10] Magerit v3, "Methodology for Information Systems Risk Analysis and Management (MAGERIT version 3)," Ministerio de Administraciones Públicas, Spain2012.

[11] C. J. Alberts and A. Dorofee, "Managing information security risks: the OCTAVE approach." Addison-Wesley Longman Publishing Co., Inc., 2002, p.^pp. Pages.

[12] CRAMM v5.0, CCTA Risk Analysis and Management Method., 2003.

[13] L. E. Sánchez, A. Santos-Olmo, D. G. Rosado, and M. Piattini, "Managing Security and its Maturity in Small and Medium-sized Enterprises," J. UCS, vol. 15, no. 15, pp. 3038-3058, 2009.

[14] ENISA, "Baseline Security Recommendations for IoT in the context of Critical Information Infrastructures," European Union Agency For Network And Information Security2017, Available: https://www.enisa.europa.eu/publications/baseline-security-recommendations-for-iot/at_download/fullReport.

[15] ISO/IEC 27005:2011, Information Technology - Security Techniques - Information Security Risk Management Standard, 2011.

[16] Office of Government Commerce, "ITIL - Service Operation. TSO," London2007.

[17] ISACA, "Cobit 5 A Business Framework for the Governance and Management of Enterprise," 2012.

[18] L. A. J. Ronald S. Ross, "Guide for Assessing the Security Controls in Federal Information Systems: Building Effective Security Assessment Plans " NSIT2010.

[19] A. G. Hessami, H. Jahankhani, and M. Nkhoma, "Responsive Cyber-Physical Risk Management (RECYPHR)," in International Conference on Global Security, Safety, and Sustainability, 2015, pp. 263-274: Springer.

[20] J. L. Cebula and L. R. Young, "A taxonomy of operational cyber security risks," CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST2010.

[21] P. Dong, Y. Han, X. Guo, and F. Xie, "A systematic review of studies on cyber physical system security," Int. J. Secur. Appl, vol. 9, no. 1, pp. 155-164, 2015.

[22] D. M. Nicol, "Risk Assessment of Cyber Access to Physical Infrastructure in Cyber-Physical Systems," in Proceedings of the 2nd ACM International Workshop on Cyber-Physical System Security, 2016, pp. 1-2: ACM.

[23] W. Wu, R. Kang, and Z. Li, "Risk assessment method for cyber security of cyber physical systems," in Reliability Systems Engineering (ICRSE), 2015 First International Conference on, 2015, pp. 1-5: IEEE.

[24] Y. Peng, T. Lu, J. Liu, Y. Gao, X. Guo, and F. Xie, "Cyber-physical system risk assessment," in Intelligent Information Hiding and Multimedia Signal Processing, 2013 Ninth International Conference on, 2013, pp. 442-447: IEEE.

[25] J. R. Nurse, S. Creese, and D. De Roure, "Security risk assessment in Internet of Things systems," IT Professional, vol. 19, no. 5, pp. 20-26, 2017.

[26] P. A. Ralston, J. H. Graham, and J. L. Hieb, "Cyber security risk assessment for SCADA and DCS networks," ISA transactions, vol. 46, no. 4, pp. 583-594, 2007.

[27] Y. Cherdantseva et al., "A review of cyber security risk assessment methods for SCADA systems," Computers & security, vol. 56, pp. 1-27, 2016.

[28] Z. DeSmit, A. E. Elhabashy, L. J. Wells, and J. A. Camelio, "An approach to cyber-physical vulnerability assessment for intelligent manufacturing systems," Journal of Manufacturing Systems, vol. 43, pp. 339-351, 2017.

[29] T. Hecht, L. Langer, and P. Smith, "Cybersecurity risk assessment in Smart Grids," Tagungsband ComForEn 2014, p. 39, 2014.

[30] ENISA, "ENISA Threat Taxonomy. A tool for structuring threat information," European Union Agency For Network And Information Security2016,

Aplicación de técnicas de compresión de información a la identificación de anomalías en fuentes de datos heterogéneas: análisis y limitaciones

Gonzalo de la Torre-Abaitua
Instituto Nacional de Ciberseguridad
INCIBE
gonzalo.torre@incibe.es

Luis F. Lago-Fernández
Escuela Politécnica Superior
Departamento de Ingeniería Informática
Universidad Autónoma de Madrid
luis.lago@uam.es

David Arroyo
Escuela Politécnica Superior
Departamento de Ingeniería Informática
Universidad Autónoma de Madrid
david.arroyo@uam.es

***Resumen*—Hoy en día la seguridad en los sistemas de información ha de ser interpretada desde un prisma multidisciplinar que integre fuentes heterogéneas de datos, y que permita tener una visión completa de los eventos de seguridad que afectan a una corporación o a un individuo. Por tanto, el ámbito de la seguridad ya no solo queda limitado a analizar eventos estructurados de *logs*, sino que se deben observar otras fuentes de datos como pueden ser las redes sociales o la utilización de listas blancas/negras. El nexo común de todas estas fuentes de datos es que pueden ser interpretadas como texto codificado mediante lenguaje natural o estructurado. Tradicionalmente, la forma de abarcar estos problemas ha sido mediante la aplicación de algoritmos y mecanismos específicos para cada uno de ellos. Recientemente hemos propuesto una metodología para la detección de anomalías en peticiones HTTP. En este trabajo analizamos la extensibilidad de esta metodología para su aplicación a la detección de anomalías en texto. En concreto analizamos su rendimiento en la detección de *spam* y de dominios DGA. Los resultados obtenidos muestran un buen rendimiento en la detección de *spam* y un rendimiento aceptable en la detección de DGA al compararse con mecanismos clásicos, de forma que se enfatiza el carácter agnóstico del marco de trabajo. Además, el método empleado no requiere realizar ingeniería de características y precisa un número reducido de parámetros. Por último, se ha observado que se alcanzan resultados positivos con menos datos de entrenamiento que los necesarios para otras aproximaciones al problema.**

***Index Terms*—NCD, spam, DGA, IDS, detección de anomalías, fuentes heterogéneas de datos**

I. Introducción

Los sistemas actuales y las múltiples ramificaciones que tiene la ciberseguridad hoy en día deparan un escenario complejo [1], [2]. En este escenario, la naturaleza de las incidencias de seguridad va más allá de la caracterización *clásica* de eventos mediante *logs* de diferentes aplicaciones o listas negras/blancas de activos de información. En efecto, también es posible encontrar información sobre una compañía, vulnerabilidades, o *zero-days* en las redes sociales [3], [4]. A todo ello hay que añadir el sesgo interpretativo de los usuarios y operadores de los sistemas de información, de forma que hay que tener muy en cuenta el papel que aquellos juegan en la interpretación de eventos e incidencias de seguridad [5]. Todos estos elementos pueden interpretados como texto codificado de acuerdo con algún tipo de lenguaje natural o, al menos, estructurado. A la hora de explotar dicha información, es importante utilizar un mecanismo común para comparar eventos. Es en este escenario donde la utilización de técnicas de compresión de información se revela como un mecanismo capaz de aglutinar fuentes de información correspondientes a sistemas de codificación diversos. Esta idea ha sido aplicada con éxito en diferentes tareas como la clasificación o *clusterización* de información en el ámbito de la genómica [6], la identificación de diferentes lenguas euro-asiáticas o la detección de plagio en trabajos académicos [7], [8]. En el dominio específico de la ciberseguridad, recientemente hemos utilizado técnicas de compresión de información para el descubrimiento de tráfico HTTP anómalo [9].

Normalmente, los análisis de cualquiera de las técnicas comentadas anteriormente se llevan a cabo desde un punto de vista aislado, analizando cada problema, extrayendo las características relevantes en cada caso y definiendo un modelo específico [10], [11], [12].

Nuestro trabajo está enfocado a analizar las posibilidades de afrontar de forma holística el problema, mediante la propuesta y análisis de un marco integral aplicable a cualquiera de ellos. Esto hace que se deba encontrar un mecanismo capaz de abstraerse de cada problema individual y sin realizar modificaciones en el marco sea capaz de identificar las características intrínsecas a cada problema. Esta aproximación es similar al concepto de aprendizaje agnóstico aplicado sobre aprendizaje correcto probablemente aproximado (PAC) [13]. Para ello, en esta investigación hemos extendido un trabajo previo en el que mediante el uso de una variación de la distancia de compresión normalizada (NCD) [7] se establece un procedimiento para detectar anomalías en peticiones web [9].

Los resultados obtenidos muestran que la metodología propuesta proporciona rendimientos similares o ligeramente inferiores a los de otros mecanismos tradicionales cuando se aplica problemas de distintos ámbitos. No obstante, esta diferencia queda compensada por otras bondades mostradas por la metodología, ya que alcanza rendimientos aceptables con un menor número de datos para entrenamiento y es agnóstica respecto al problema concreto.

El resto del artículo se organiza de la siguiente forma: en la sección II se introducen los problemas abordados, las

estrategias de análisis habituales, y la aproximación seguida en este trabajo; en la sección III se describe la metodología propuesta y se describen los resultados obtenidos; finalmente, en la sección IV se presentan las conclusiones y el trabajo futuro.

II. Detección de spam y DGAs mediante técnicas de análisis de texto

Con el objetivo de analizar la extensibilidad del marco desarrollado, en este trabajo hemos estudiado su aplicación a dos problemas concretos y relacionados con el análisis de texto en ciberseguridad, como son la detección de *spam* y la detección de dominios generados mediante algoritmos de generación de dominios (DGA, por sus siglas en inglés -Domain Generation Algorithms-).

Tradicionalmente, la detección de *spam* se realiza mediante la extracción y el uso de distintas características para posteriormente aplicar algoritmos de aprendizaje automático. Estos algoritmos engloban desde la aplicación de *Naïve Bayes* o *regresión logística* hasta redes neuronales [14]. Así mismo, también se ha confiado en técnicas de análisis de texto o de clustering basándose en frecuencias de palabras (TF -Term Frequency-) y frecuencia inversa (TFIDF -Term Frequency Inverse Document Frequency-) [15], en el uso de distancias edit [16] o en la aplicación de técnicas de compresión [17] [18]. Estas últimas técnicas obtienen buenos resultados y ventajas adicionales en relación con otros métodos de procesamiento de lenguaje natural clásicos, ya que son robustas frente a la introducción de ruido en los mensajes que pudiera provocar una clasificación errónea de los mismos [19]. Aprovechando el buen rendimiento descrito y las ventajas adicionales de estas técnicas, hemos explorado la aplicación de la NCD como distancia de compresión utilizada para detectar de forma eficiente mensajes de *spam*.

Por otro lado, hemos tratado el problema de la detección de dominios DGA. Los dominios DGA son dominios generados de manera aleatoria por un *bot* u ordenador zombie y son utilizados para esconder los paneles de comando y control (C&C). Este es uno de los mecanismos utilizados por las redes de ordenadores zombies o *botnets*, que consisten en múltiples ordenadores, servidores o dispositivos IoT que han sido infectados con un *malware* que hace posible su control de forma remota por el dueño de la botnet[1]. Para ello, si el *malware* hace uso de la técnica de DGA, cuando un ordenador es infectado y quiere comunicarse con el panel C&C, genera múltiples nombres de dominio e intenta conectar con ellos. Solo uno de estos dominios, que ha sido previamente reservado por el dueño de la *botnet*, va a responder lo que espera el *bot* e iniciar la comunicación entre ellos. De esta forma la *botnet* es más resiliente ya que en ningún momento la IP o el nombre de dominio real de los paneles C&C aparece en el código del *malware*[2]. La detección de estos dominios es, por tanto, crucial ya que se podría detener el funcionamiento de una botnet si se conocen y son bloqueados mediante cortafuegos o sistemas IPS (Intrusion Prevention Systems).

[1]https://www.trendmicro.com/vinfo/us/security/definition/botnet. Último acceso 01/05/2018

[2]https://blog.trendmicro.com/domain-generating-algorithms-dgas/. Último acceso 01/05/2018

Las principales técnicas estudiadas para la detección de dominios DGA implican el análisis de tráfico DNS. Se basan en la frecuencia y el número de dominios que no se resuelven (esto es, aquellos cuya resolución de nombre devuelve NXDomain) y que se observan en el tráfico de una red específica. En base a esto se han realizado estudios para *clusterizar* e identificar la familia DGA responsable del tráfico [20], [21].

Otra posibilidad es tratar de predecir si un dominio ha sido generado mediante DGA basándose únicamente en el nombre del mismo. La aproximación más simple a este problema es mediante la definición de patrones una vez que se ha aplicado ingeniería inversa y se ha logrado extraer del binario el algoritmo que se usa para la generación de los dominios. Esta acción consume mucho tiempo y no siempre es posible, por lo que la utilización de otras aproximaciones se hace necesaria. En este punto se han aplicado múltiples técnicas que implican tanto la extracción de características que posteriormente son usadas para clasificar, como la utilización de redes neuronales recurrentes [22], [23]. Así mismo, se han utilizado técnicas basadas en compresión que a su vez requieren la extracción de múltiples características para la clasificación de los dominios, tales como el nivel del dominio o procedimientos basados en el análisis de n-gramas [24]. En este escenario, el método propuesto en este trabajo no requiere la extracción de características previas. Al ser un modelo basado en compresión, es capaz de identificar las secuencias de símbolos que forman un dominio legítimo frente a aquellas que dan lugar a un dominio DGA. Así mismo, al estar basado en NCD trabaja correctamente sobre cadenas cortas como pueden ser los nombres de dominio [25].

III. Detección de spam y DGAs mediante NCD: metodología y resultados

La metodología a aplicar está basada en la propuesta que recientemente hemos desarrollado para la detección de anomalías en peticiones web [9]. De forma sucinta, este método parte de la interpretación de cada muestra (petición HTTP, mensaje SMS o una URL) como una cadena de caracteres, de forma que independientemente del tipo del problema todos ellos son tratados como texto. Dicha cadena de texto es codificada mediante un vector de características que mide las distancias de la cadena a un conjunto de aglomeraciones de cadenas (*generadores de características*). El *dataset* original se divide en dos grupos disjuntos. El primero se utiliza para construir los generadores de características, k ficheros que aglutinan, cada uno, un conjunto de cadenas, todas de la misma clase. El segundo contiene las cadenas que se utilizarán para entrenar el clasificador. Cada una de ellas será descrita por un vector de k componentes que representan, respectivamente, la distancia de la cadena a cada uno de los k generadores de características. La distancia entre la cadena b y el generador a viene dada por [8]:

$$D(a,b) = \frac{C(b|a)}{C(b)} = \frac{C(ab) - C(a)}{C(b)} \quad (1)$$

donde ab representa la concatenación de a y b y $C(x)$ es el tamaño en bytes de x tras aplicar el algoritmo de compresión gzip.

III-A. Preparación de los datos

Para analizar la extensibilidad del mecanismo anterior al aplicarse a cualquier tipo de texto, hemos utilizado dos *datasets* diferentes con eventos relevantes para la ciberseguridad. Por un lado, la detección de dominios DGA basándose en el nombre de dominio y, por otro, la detección de mensajes de *spam* en SMS basándose exclusivamente en el texto del SMS. En ambos casos hemos utilizado *datasets* que son accesibles públicamente.

III-A1. DGA: Para la realización de esta prueba se han utilizado como dominios normales el top un millón de Alexa[3] y opendns, y como maliciosos aquellos generados por 11 familias de *malware*, como zeus, cryptolocker, pushdo o conficker (la lista completa se puede consultar en github[4]). Los datos en bruto contienen un dominio por línea. En total contamos con 800000 dominios DGA y 1000000 dominios normales por lo que al estar utilizando un *dataset* balanceado, hemos extraído de forma aleatoria 800000 dominios normales del conjunto de dominios normales inicial. Una vez balanceados los datos, se tiene un total de 1600000 dominios, los cuales se dividen en los dos grupos disjuntos previamente explicados. El grupo utilizado para entrenar el clasificador consta de 1600 dominios balanceados. El resto de dominios de ambas clases, 1568000, forman los generadores de características. Estos a su vez son divididos en k grupos como puede verse en la tabla I.

III-A2. Spam SMS: El *dataset* utilizado es accesible tanto desde la página web de los autores[5] como desde Kaggle[6]. Contiene un total de 5574 mensajes clasificados como *spam* (747 mensajes) o neutrales (*ham*, 4827 mensajes). Antes de aplicar la metodología hemos balanceado los datos para obtener el mismo número de ejemplos de cada clase, tomando todos los mensajes *spam* y 747 mensajes *ham* elegidos al azar. En total utilizamos 1494 mensajes, de los que hemos obtenido de manera aleatoria 200 mensajes *ham* y 200 *spam* para formar el grupo de entrenamiento, utilizando para los generadores de características el resto de mensajes, 547 de cada clase.

III-B. Resultados

Una vez obtenidos los vectores de características para cada uno de los problemas, se utiliza como clasificador una máquina de soporte vectorial (SVM) con kernel rbf, usando la implementación de la librería scikit-learn[7]. La selección de parámetros se realiza mediante la estrategia de validación cruzada descrita en [9].

III-B1. DGA: Se han realizado experimentos con distinto número de características, $k \in \{8, 16, 80, 160\}$ cuyos resultados se pueden ver en la tabla I. El mejor resultado se obtiene para $k = 80$, con un área bajo la curva ROC (acrónimo de Receiver Operating Characteristic, o Característica Operativa del Receptor) de 0,88. En adelante notaremos esta curva usando las siglas en inglés AUC (Area Under Curve). El anterior resultado es inferior al obtenido en otros trabajos en los que se indica un *accuracy* que varía entre 0,9 y 0,99 en una clasificación en n clases por familia de malware [24] y un *accuracy* de 0,91 con AUC de 0,97 para una clasificación binaria utilizando métodos tradicionales [26]. No obstante, las particularidades del marco de referencia hacen que el número de elementos disponibles para entrenamiento sea menor al utilizado en el resto de investigaciones, ya que el *dataset* completo se divide en dos grupos como se explica en III. Teniendo esto en cuenta, los resultados son muy prometedores. También se debe tener en cuenta, que los experimentos llevados a cabo han sido sobre una tarea binaria en la que múltiples dominios generados por distintas familias DGA engloban una clase y dominios normales engloban la otra. Hasta hace poco, gran parte del esfuerzo investigador se centraba en clasificar de manera individual o en grupos pequeños cada una de las familias DGA pero últimamente se han incrementado las investigaciones basadas en redes neuronales con el objetivo de realizar una clasificación binaria [22][26]. El marco de referencia analizado logra resultados aceptables sin la necesidad de realizar una obtención o manipulación previa de características (ingeniería de características) ni la necesidad de entrenar múltiples parámetros, estableciendo un mecanismo de rápida y sencilla aplicación capaz de identificar la variación de entropía existente en las múltiples variaciones que se dan en los nombres de dominio generados por las 11 familias DGA analizadas. El mecanismo sigue siendo capaz de detectar de manera automática las características que identifican cada clase pese a la variabilidad en los nombres de dominio.

Tabla I
RESULTADOS PARA EL PROBLEMA DGA. NOMBRES DE COLUMNAS: k, NÚMERO DE ATRIBUTOS; n_q, NÚMERO DE DOMINIOS EN CADA FICHERO GENERADOR; C, PARÁMETRO DE COMPLEJIDAD; γ, PARÁMETRO DEL KERNEL; ACC., *accuracy* MEDIO AL APLICAR VALIDACIÓN 10-FOLD; AUC, ÁREA MEDIA BAJO LA CURVA ROC.

k	n_q	C	γ	Acc.	AUC
8	196000	50	25	0,717 +/-0,005	0,785
16	98000	0,5	115	0,77+/-0,008	0,84
80	19600	1	25	0,82 +/-0,006	0,88
160	9800	1,5	25	0,795 +/-0,005	0,869

III-B2. Spam SMS: Hemos realizado experimentos para $k \in \{8, 16, 32, 80, 160\}$, donde los mejores resultados se obtienen para $k = 160$, con un AUC de $0,96$ y *accuracy* de $0,9$ (tabla II). Estos resultados se asemejan a los obtenidos en la literatura, donde análisis sobre el mismo dataset aplicando diferentes modelos, reportan un *accuracy* entre 0,85 y 0,97 dependiendo del modelo utilizado [27]. Si bien, con el modelo estudiado se reduce y simplifica todo el trabajo previo de manipulación y selección de características así como el entrenamiento de múltiples parámetros e hiperparámetros. Además se tiene la ventaja añadida de que se alcanza este rendimiento utilizando un número menor de elementos para entrenar el modelo como se explica en III-A2

IV. CONCLUSIONES

En este trabajo hemos explorado la aplicación de un marco de trabajo para analizar de modo integral múltiples fuentes de datos en el ámbito de la ciberseguridad. El procedimiento

[3]https://www.alexa.com/topsites, Último acceso 01/05/2018

[4]https://github.com/andrewaeva/DGA. Último acceso 01/05/2018

[5]http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/. Último acceso 01/05/2018

[6]https://www.kaggle.com/uciml/sms-spam-collection-dataset/data. Último acceso 01/05/2018

[7]http://scikit-learn.org/stable/. Último acceso 01/05/2018

Tabla II
RESULTADOS PARA EL PROBLEMA *spam* SMS. NOMBRES DE COLUMNAS: k, NÚMERO DE ATRIBUTOS; n_q, NÚMERO DE LÍNEAS EN CADA FICHERO GENERADOR; C, PARÁMETRO DE COMPLEJIDAD; γ, PARÁMETRO DEL KERNEL; ACC., *accuracy* MEDIO AL APLICAR VALIDACIÓN 10-FOLD; AUC, ÁREA MEDIA BAJO LA CURVA ROC.

k	n_q	C	γ	Acc.	AUC
8	136	115	25	0,783 +/-0,095	0,85
16	68	1	100	0,827 +/-0,015	0,89
32	34	1,5	50	0,86 +/-0,005	0,92
80	13	1,5	25	0,888 +/-0,013	0,95
160	6	5	25	0,9 +/-0,009	0,96

utiliza la NCD para construir el conjunto de atributos que habilita la clasificación de los datos mediante máquinas de soporte vectorial. El marco de trabajo desarrollado en este artículo combina elementos de la teoría de la información y herramientas de aprendizaje. Este enfoque es, pues, de carácter multidisciplinar y habilita la incorporación por parte de los analistas de ciberseguridad de soluciones propuestas en el campo de la caracterización de lenguajes naturales y/o estructurados. De modo más específico, la metodología que se ha analizado parte de nuestros resultados previos en el ámbito de la detección de ataques en peticiones HTTP [9], y se ha evaluado tomando en consideración los problemas de la detección de *spam* y de dominios DGA.

El rendimiento alcanzado en la detección de *spam* es muy positivo y establece la base para ampliar el uso del mecanismo estudiado a otras tareas relevantes dentro del análisis de lenguaje natural y/o estructurado. Esto es importante ya que, aparte de la detección de *spam*, la metodología puede servir de base en caracterización fuentes abiertas de datos [3], en la clasificación de formatos de codificación [25] y en la atribución de incidencias de seguridad [28], [29]. Por otro lado, el rendimiento alcanzando en la detección de dominios DGA, sin ser un rendimiento malo, se muestra inferior al propuesto en la literatura. No obstante, este hecho se ve compensando gracias a que se simplifica el trabajo previo de ingeniería de características, siendo una fase que con este mecanismo no es necesaria, así como la simplicidad en el número de parámetros e hiperparámetros que hay que entrenar. Así mismo, estos rendimientos se alcanzan utilizando un menor número de elementos del *dataset* para entrenar el modelo.

Con la popularización de las redes neuronales profundas y su aplicación a distintos ámbitos, se ha observado un gran avance en la detección de dominios DGA basándose exclusivamente en el nombre de dominio [22]. Esto hace que los resultados obtenidos en este análisis aparentemente queden desvirtuados. No obstante, hay que hacer notar que el mecanismo propuesto alcanza su rendimiento utilizando solo 1600000 dominios frente a los 6900000 dominios utilizados para entrenar la red neuronal [26]. Así mismo, la decisión sobre la arquitectura del modelo y los distintos parámetros e hiperparámetros que se necesitan se simplifica, ya que en el marco de referencia estudiado estos son idénticos independientemente del problema tratado.

Por otro lado, el mecanismo propuesto no requiere de la extracción de parámetros previos, siendo únicamente necesaria la fijación de los parámetros del algoritmo de clasificación elegido, en nuestro caso SVM. Adicionalmente, revela un rendimiento óptimo utilizando un número menor de datos para entrenar. También cabe destacar, que el mecanismo propuesto es exactamente el mismo para los tres problemas tratados, lo que da una ventaja sobre el resto de procedimientos aplicados en la literatura ya que estos se centran en un solo problema.

Los resultados obtenidos muestran que la metodología propuesta es potencialmente extensible a otras tareas de clasificación binaria. Dichas tareas abarcan la clasificación de textos con diferentes longitudes y características, texto estructurado de diferente longitud como son los *logs* de peticiones HTTP o los nombres de dominio y texto de lenguaje natural como son los mensajes SMS. Esto es indicador de versatilidad y polivalencia de nuestro método, por lo que se revela como una herramienta útil para analistas de seguridad o administradores de sistemas, proporcionando un mecanismo que permite reducir la superficie de análisis y haría factible centrarse únicamente en los eventos interesantes. No obstante, todavía queda espacio para la mejora del rendimiento en la detección de dominios DGA y posteriores experimentos se realizarán para mejorar los resultados con la premisa de seguir manteniendo el enfoque de amplio espectro de la metodología.

Dicho esto, pese a las ventajas explicadas, la precisión de nuestro método es superada por propuestas actuales basadas en redes neuronales profundas. Nuestro trabajo futuro estará orientado a realizar una comparación exhaustiva de nuestra metodología con aquellos esquemas basados en redes recurrentes, lo que requerirá que estos sean entrenados y validados con un conjunto de datos de similar tamaño al que nosotros hemos empleado en la evaluación de nuestro procedimiento. Este estudio se realizará de modo solidario a la consideración de nuestros contextos de aplicación para la ciberseguridad, como son las tareas de análisis de sentimiento, la detección de familias de malware DGA de forma individual, y la concreción de mecanismos de visualización de datos que ayuden a los analistas a discernir de forma rápida incidencias de seguridad de interés.

AGRADECIMIENTOS

Este trabajo ha sido financiado por la Comunidad de Madrid (España) dentro del proyecto S2013/ICE-3095-CM (CIBERDINE) y por el Gobierno Español dentro de los proyectos MINECO/FEDER DPI2015-65833-P y TIN2014-54580-R.

REFERENCIAS

[1] Harel, Yaniv y Gal, Irad Ben y Elovici, Yuval: "Cyber Security and the Role of Intelligent Systems in Addressing Its Challenges", en *ACM Trans. Intell. Syst. Technol.*, vol. 8, n. 4, pp. 49:1-49:12, 2017.

[2] Zuech, Richard y Khoshgoftaar, Taghi M. y Wald, Randall: "Intrusion detection and Big Heterogeneous Data: a Survey", en *Journal of Big Data*, vol. 2, n. 1, 2015.

[3] Sabottke, Carl y Suciu, Octavian y Dumitraş, Tudor: "Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-world Exploits", en *Proceedings of the 24th USENIX Conference on Security Symposium*, pp. 1041-1056, 2015.

[4] Okutan, Ahmet y Yang, Shanchieh Jay y McConky, Katie: "Forecasting Cyber Attacks with Imbalanced Data Sets and Different Time Granularities", en *arXiv preprint arXiv:1803.09560*, 2018.

[5] Ovelgönne, Michael y Dumitras, Tudor y Prakash, B Aditya y Subrahmanian, VS y Wang, Benjamin: "Understanding the Relationship between Human Behavior and Susceptibility to Cyber Attacks: A Data-Driven Approach", en *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, n. 4, 2017.

[6] Ferragina, Paolo y Giancarlo, Raffaele y Greco, Valentina y Manzini, Giovanni y Valiente, Gabriel: "Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment", en *BMC bioinformatics*, vol. 8, n. 1, 2007.

[7] Cilibrasi, Rudi y Vitanyi, Paul: "Automatic Extraction of Meaning from the Web" en *2006 IEEE International Symposium on Information Theory*, pp. 2309-2313, 2006.

[8] Cilibrasi, Rudi y Vitányi, Paul M B: "Clustering by compression", en *IEEE Transactions on Information Theory*, vol. 51, n. 4, pp. 1523-1545, 2005.

[9] de la Torre-Abaitua, Gonzalo y Lago-Fernández, Luis F. y Arroyo, David: "A Parameter-Free Method for the Detection of Web Attacks" en *International Joint Conference SOCO'17-CISIS'17-ICEUTE'17*, pp. 661-671, 2017.

[10] Tingmin Wu y Sheng Wen y Yang Xiang y Wanlei Zhou: "Twitter spam detection: Survey of new approaches and comparative study", en *Computers & Security*, 2017.

[11] Shirin Nilizadeh y Francois Labreche y Alireza Sedighian y Ali Zand y José M. Fernandez y Christopher Kruegel y Gianluca Stringhini y Giovanni Vigna: "POISED: Spotting Twitter Spam Off the Beaten Paths", en *CoRR*, vol. abs/1708.09058, 2017.

[12] Zhang, Han y Gharaibeh, Manaf y Thanasoulas, Spiros y Papadopoulos, Christos: "BotDigger: Detecting DGA Bots in a Single Network", en *Proceedings of the IEEE International Workshop on Traffic Monitoring and Analaysis*, pp. 16-21, 2016.

[13] Kearns, Michael J. y Schapire, Robert E. y Sellie, Linda M.: "Toward Efficient Agnostic Learning", en *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 341-352, 1992.

[14] Abu-Nimeh, Saeed y Nappa, Dario y Wang, Xinlei y Nair, Suku: "A Comparison of Machine Learning Techniques for Phishing Detection", en *Proceedings of the Anti-phishing Working Groups 2Nd Annual eCrime Researchers Summit*, pp 60-69, 2007.

[15] Mallikarjunappa, Basavaraju y R. Prabhakar, Dr: "A Novel Method of Spam Mail Detection using Text Based Clustering Approach", en *International Journal of Computer Applications*, vol. 5, 2010.

[16] Tee, Huat Heng: "FPGA Unsolicited Commercial Email Inline Filter Design Using Levenshtein Distance Algorithm and Longest Common Subsequence Algorithm", en *Jabatan Sistem dan Teknologi Komputer, Fakulti Sains Komputer dan Teknologi Maklumat, Universiti Malaya*, 2010.

[17] Delany, Sarah Jane y Bridge, Derek: "Catching the Drift: Using Feature-Free Case-Based Reasoning for Spam Filtering", en *Case-Based Reasoning Research and Development*, pp. 314-328, 2007.

[18] Michal Prilepok y Petr Berek y Jan Platos y Vaclav Snasel: "spam detection using data compression signatures", en *Cybernetics and Systems*, vol. 44, pp. 533-549, 2013.

[19] Bratko, Andrej y Filipič, Bogdan y Cormack, Gordon V. y Lynam, Thomas R. y Zupan, Blaž: "Spam Filtering Using Statistical Data Compression Models", en *J. Mach. Learn. Res.*, vol. 7, pp. 2673-2698, 2006.

[20] Thomas, Matthew y Mohaisen, Aziz: "Kindred Domains: Detecting and Clustering Botnet Domains Using DNS Traffic", en *Proceedings of the 23rd International Conference on World Wide Web*, pp. 707-712, 2014.

[21] Antonakakis, Manos y Perdisci, Roberto y Nadji, Yacin y Vasiloglou, Nikolaos y Abu-Nimeh, Saeed y Lee, Wenke y Dagon, David: "From Throw-away Traffic to Bots: Detecting the Rise of DGA-based Malware" en *Proceedings of the 21st USENIX Conference on Security Symposium*, pp. 24, 2012.

[22] Jonathan Woodbridge y Hyrum S. Anderson y Anjum Ahuja y Daniel Grant: "Predicting Domain Generation Algorithms with Long Short-Term Memory Networks", en *CoRR*, vol. abs/1611.00791, 2016.

[23] Ahluwalia, Aashna y Traore, Issa y Ganame, Karim y Agarwal, Nainesh: "Detecting Broad Length Algorithmically Generated Domains", en *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pp. 19-34, 2017.

[24] Tong, Van y Nguyen, Giang: "A Method for Detecting DGA Botnet Based on Semantic and Cluster Analysis", en *Proceedings of the Seventh Symposium on Information and Communication Technology*, pp. 272-277, 2016.

[25] Axelsson, Stefan: "The Normalised Compression Distance as a file fragment classifier", en *Digital Investigation*, vol. 7, 2010.

[26] Pierre Lison y Vasileios Mavroeidis: "Automatic Detection of Malware-Generated Domains with Recurrent Neural Models", en *CoRR*, vol. abs/1709.07102, 2017.

[27] Almeida, Tiago A. e Hidalgo, José María G. y Yamakami, Akebo: "Contributions to the Study of SMS Spam Filtering: New Collection and Results", en *Proceedings of the 11th ACM Symposium on Document Engineering*, pp. 259-262, 2011.

[28] Stamatatos, Efstathios: "A survey of modern authorship attribution methods", in: *Journal of the Association for Information Science and Technology*, vol. 60, 2009.

[29] Ding, Steven HH y Fung, Benjamin CM y Iqbal, Farkhund y Cheung, William K: "Learning Stylometric Representations for Authorship Analysis", in *IEEE Transactions on Cybernetics*, 2017.

Mapeo de Dependencias para el Impacto de Ciberataques en Misiones: Una Visión Global

R. Barragán Gil[1], J. López Hernández-Ardieta[2], P. García Teodoro[1]
[1]Universidad de Granada - *Network Engineering & Security Group*
rbg@ugr.es, pgteodor@ugr.es
[2]Minsait by Indra
jlhardieta@minsait.com

***Resumen*—Este trabajo se enmarca en el contexto de una tesis industrial recientemente iniciada entre Indra Sistemas y la Universidad de Granada. En ella se busca proponer soluciones al tema del impacto de misión debido a ciber-amenazas o fallos en la infraestructura CIS. En este contexto, la tesis se enfoca en la realización del mapeo de dependencias entre la infraestructura CIS y las tareas de la misión. Se busca que este mapeo sea capaz de absorber el dinamismo de la infraestructura CIS y esto se lleve a cabo de la forma más automática posible. Además, se pretende buscar nuevos modelos de representación que permitan el empleo de razonamiento automático, de igual forma que se propondrán y evaluarán nuevas métricas para identificar elementos críticos en misión.**

***Index Terms*—Mapeo de dependencias, impacto de misión, activos críticos de misión**

I. Introducción

Los sistemas de comunicación e información (CIS) están concebidos para permitir procesos de negocio, o misiones. Aunque los ciberataques se desarrollan en el ciberespacio y, como tal, tienen un impacto directo sobre los CIS, su efecto último son los procesos de negocio que soportan y ofrecen los CIS.

Atendiendo a lo anterior, la pregunta clave no es a qué parte del CIS afecta un cierto ciberataque, sino cuál es su impacto a nivel de proceso de misión. Esto recibe el nombre de *mission impact assesment* (MIA), y constituye un área de investigación que está ganando una elevada atención en los últimos años. MIA forma parte de un área de estudio más amplia conocida como *cyber situational awareness* (CSA). CSA persigue posibilitar a los responsables de la toma de decisiones de un cierto entorno observar, comprender y evaluar las ciber-amenazas a fin de facilitar la labor de planificación de riesgos y toma de decisiones en las operaciones de negocio.

MIA contribuye a CSA traduciendo los eventos de la red en un valor de impacto en el negocio. Esto es, entendiendo las dependencias que a nivel de negocio o misión tiene la infraestructura CIS que sustenta las actividades, procesos y personas que realizan dicha misión. Ello resulta fundamental para garantizar la continuidad de negocio y, por tanto, que la organización logre sus objetivos estratégicos. Un mapeo completo, coherente y dinámico de la dependencia de la ciber-misión permitiría diseñar y desplegar actividades adecuadas de mitigación de riesgos y respuestas a ataque prioritarias para mantener el impacto sobre la misión dentro de unos umbrales aceptables. Este mapeo de dependencias resulta una tarea crítica durante la evaluación del impacto de la misión y conlleva importantes retos debido no sólo a los cambios constantes en la infraestructura CIS, sino también en los objetivos mismos de la misión, sus prioridades e interdependencias.

II. Contexto

A día de hoy existe en el sector privado un gran interés en CSA y, por tanto, en MIA. Especialmente grandes organizaciones, aceptan el hecho de que las infraestructuras CIS afectan a sus operaciones tanto como cualquier otra capacidad y, en consecuencia, se precisa una adecuada gestión para asegurar la correcta operación. De esta forma, la ejecución de la misión depende en muchos casos del correcto funcionamiento de la red, *hosts* y servicios subyacentes La complejidad y singularidad de los procesos de negocio, en particular los que soportan grandes empresas y operaciones militares, hacen que abordar CSA y MIA de forma efectiva sea un problema difícil que precisa respuestas innovadoras. La OTAN, por ejemplo, está invirtiendo grandes esfuerzos en el desarrollo de capacidad CSA, realizando en 2015 un estudio exhaustivo de las soluciones comerciales que pueden brindar una capacidad parcial o completa en este sentido [1]. Uno de los requisitos de la OTAN es la visualización de riesgos en diferentes niveles de abstracción, incluidas misiones, facilitando a los mandos tomar decisiones teniendo en cuenta el equilibro costo/beneficio y donde la ejecución pudiese ser completamente automatizada.

En un informe reciente, el Grupo Asesor Industrial de la OTAN (NIAG) realizó una encuesta actualizada de soluciones y tecnologías sobre CSA [2]. El estudio cubrió varias áreas que contribuyen a CSA: tecnologías de gestión de datos, tecnologías analíticas de datos y tecnologías de apoyo a la toma de decisiones. La conclusión principal del estudio es que ninguna solución única cumple con todos los requisitos operativos de la OTAN. Atendiendo al *gap* evidenciado entre capacidades ofertadas y requeridas, el NIAG sugiere algunas áreas de mayor investigación, entre las que se encuentra MIA. La OTAN también destaca la importancia del filtrado avanzado de la información para ajustar la visualización de cada usuario dependiendo de sus requerimientos, reduciendo la sobrecarga de información y aumentando la eficiencia operacional.

Por su parte, la Agencia de Defensa Europea está llevando un proyecto de fortalecimiento de conciencia cibernética. La primera fase del proyecto se completó en 2015, y describía las necesidades operativas del usuario y definía una arquitectura conceptual [3]. La segunda fase, iniciada en 2017, pretende desarrollar los requisitos del sistema y la arquitectura de destino que precede a la construcción de un prototipo de

investigación [4]. Dos son los aspectos más relevantes de la solución prevista: una herramienta de apoyo a la toma de decisiones, y la capacidad de gestión dinámica de riesgos.

Otras organizaciones militares, como el ejército EE.UU., ha lanzado iniciativas para desarrollar capacidades de evaluación de impacto en misiones (MIA) y mapeo de misión [5]. En trabajos como el llevado a cabo por las Fuerzas Aéreas de EE.UU. [6], se analizan diferentes tecnologías que pueden ayudar durante la identificación de dependencias en ciber-misiones. Los autores comienzan definiendo un modelo de mapeo de misión que incluye abstracciones de la información (misiones, procesos, capacidades de red, *logs*), las aplicaciones soportadas por el mapa de misión y el trabajo de usuarios y defensores. Se permite así, por ejemplo, realizar un particionado de la red para reducir la superficie de ataque sobre activos críticos.

Las tecnologías de mapeo de misión analizadas en [6] se clasifican como *técnicas de análisis impulsadas por procesos* o *técnicas de análisis basadas en artefactos*. Las primeras siguen un enfoque centrado en el ser humano, donde el mapeo de la misión es construido manualmente por expertos en la materia (SME -*Subject Matter Expert*). Esto presenta una serie de limitaciones: (1) es altamente subjetivo; (2) consume mucho tiempo; y (3) el mapeo de la misión no se puede actualizar en tiempo real. Frente a estas, las técnicas de análisis basadas en artefactos utilizan registros y datos de los *hosts* y la red para inferir los activos (y por lo tanto las dependencias) utilizados por los procesos durante las misiones. Los principales desafíos están relacionados con garantizar la integridad del mapeo de la misión debido a la falta de fuentes de datos apropiadas, así como con producir datos fáciles de interpretar por los humanos. En comparación con los enfoques basados en procesos, estas técnicas son mucho más rápidas, admiten el descubrimiento de activos en tiempo real y pueden identificar dependencias no triviales. Los autores sugieren un enfoque compuesto, semi-automatizado, donde las técnicas basadas en artefactos se utilizan para descubrir activos de red y construir una correspondencia de dependencia oportuna entre los procesos y los activos (de abajo hacia arriba), pero donde la correspondencia de dependencia entre misiones y procesos es creada por un analista humano (de arriba hacia abajo).

En [7] los autores presentan *Orion*, sistema que realiza un mapeo automático de relaciones entre servicios. Para realizar dicho mapeo se basan en tráfico de red, usando como entrada del sistema solo cabeceras de dicho tráfico; concretamente: (1) dirección de destino, (2) puerto destino y (3) protocolo. Con esta información y los *timestamp* se realiza un análisis del retardo entre flujos, que da como resultado un mapa relacional entre los diferentes servicios. Este descubrimiento automático de servicios, así como las relaciones de dependencia entre ellos, cuenta con una buena tasa de éxito y un bajo número de falsos positivos.

Partiendo del análisis de las tecnologías existentes, los autores de [6] concluyen que ninguna propuesta proporciona un conjunto completo de las funcionalidades requeridas. Los autores identifican, así, algunas áreas que necesitan más investigación. Estas incluyen recopilación de datos activos para construir asignaciones de misión más confiables, identificación de dependencias de misión en misiones no recurrentes, actualización oportuna del mapeo de misiones a medida que evolucionan las necesidades de la misión e implicaciones de seguridad para la misión si un atacante compromete el mapeo de la misión. Los autores también sugieren indicaciones para futuras investigaciones:

1. Perturbación de activos para sondear y medir la sensibilidad de la misión a cada activo de red.
2. Arquitecturas compatibles con misiones, que dependen en gran medida de redes definidas por software (SDN -*Software Defined Network*) y técnicas de virtualización para implementar arquitecturas de sistemas que ya reflejan las dependencias de la misión.
3. Enfoques de representación de la información en YAML (*Yet Another Markup Language*) basados en el aprendizaje automático, así como técnicas de Big Data de misiones pasadas para ayudar al analista durante el mapeo de las capacidades a las necesidades actuales de la misión.
4. Obtención de una línea de base del comportamiento de la misión basada en roles, que correlaciona el rol del usuario que ejecuta las tareas en el sistema de planificación de la misión y los activos que respaldan sus funciones. Al definir roles apropiados y al usar técnicas de aprendizaje automático, se puede establecer una firma basada en roles para el uso de activos. Al aprovechar esta información, el mapeo de la misión se puede construir sin la necesidad de acceder a la información de planificación de la misión de los demás, que generalmente es de naturaleza clasificada.

En [8] se introduce el concepto de MOE (*Measures of Effectiveness*, o Medida de Efectividad) aplicada a MIA en ciberataques. Según lo definido por I+D de Canadá, MOE se refiere a "una medida que caracteriza la eficacia operativa de una unidad o fuerza en el logro de sus objetivos durante una misión". En este mismo trabajo de Musman *et al.*, los autores declaran la necesidad de determinar el impacto de la misión de un ciberataque basándose en la duración del mismo. Esto fue apuntado por primera vez por Fortson en su tesis doctoral en 2007 [9]. De acuerdo con [8], el tiempo del ataque debe correlarse con el tiempo y la información del flujo de trabajo de la misión con el fin de inferir con precisión las consecuencias en el MOE. Con este objetivo, los modelos de misión deberían capturar la información dependiente del tiempo de una misión, incluida la importancia relativa de sus diversas tareas a lo largo del tiempo.

Musman *et al.* presentan un trabajo dirigido a modelar misiones y sus relaciones con activos TIC; por ejemplo, atributos de confidencialidad, integridad o disponibilidad. Su modelo amplía BPMN (*Business Process Modeling Notation*) para capturar estas relaciones. Utilizando sus modelos, se puede calcular el impacto de los ciberataques en la misión y apoyar a quienes toman las decisiones durante la misión. Sin embargo, su propuesta requiere una intervención manual para reflejar las dependencias de la ciber-misión, así como para actualizar el modelo frente a ciberataques.

Durante la evaluación del impacto de la misión es importante identificar los activos clave que más contribuyen para apoyar los procesos y servicios de la misión. Estos

activos reciben el nombre de CKT (*Cyber Key Terrain*). En un trabajo reciente, Price *et al.* exploran el uso de la toma de decisiones multi-atributo o MADM (*Multi-Atribute Decision Making*) para identificar los activos más criticos con respecto a ciertas misiones y subsistemas de apoyo [10]. La metodología propuesta es una variante de TOPSIS (*Technique for Order of Preference by Similitary to Ideal Solution*), que es un método MADM que emplea el procedimiento analítico jerárquico (AHP - *Analytic Hierarchy Process*) como criterio para establecer los pesos entre las diferentes tareas que componen la misión. Se puede definir como basado en procesos, ya que requiere la participación de diferentes expertos en la materia (SME) a lo largo de una serie de cuestionarios de escala empleados para llevar a cabo el AHP.

En [11], los autores argumentan que no es necesaria la completa comprensión de un ataque para poder evaluar el impacto efectivo sobre la misión. Proponen, así, un sistema llamado *Cyber-Argus* que aprovecha las contribuciones de Damico *et al.* [12] y Matheus *et al.* [13] para la construcción de un sistema de ontologías para simplificar las tareas de evaluación de impacto de la misión. Como se presenta en [14] (véase Figura 1), ARGUS se compone de cuatro fases principales:

- Modelado de la misión, para el cual se emplea como entrada un modelado BPMN que traduce a un modelo ontológico propio implementado siguiendo el estándar OWL (*Ontology Web Language*), siendo necesaria la intervención humana para la realización de dicho modelo BPMN.
- Modelado de la arquitectura de la red, realizado de forma automática. Para ello emplean consultas SNMP a dos tablas del *Management Information Base* (MIB) de cada *host* en la red. El fin de estas consultas es extraer los servicios que cada dispositivo tiene hospedados, en lugar de obtener esta información del tráfico de red como se hace en otros trabajos [7].
- Recopilación de información cibernética y de la misión. En esta tercera fase se emplea una serie de reglas escritas mediante *Semantic Web Rule Language* (SWRL) para extraer información de relación entre tareas, coste y temporal.
- Desarrollo de la evaluación del impacto, en la cual se lleva a cabo (en tiempo real) la fusión de la información obtenida de los tres apartados anteriores junto con datos en tiempo real de tráfico y *syslog* para dar como resultado una medida de la situación de la misión en cada momento. En [15] se explica detalladamente cómo se realiza dicha fusión de información.

Alexander Kott indica en [16] que el paradigma basado en modelado y simulación (M&S) es el acercamiento más a largo plazo y reutilizable para ciberdefensa. Tal paradigma se basa en la creación y validación de mecanismos de modelado de la organización cuya misión está sujeta a evaluación, la misión (o misiones) en sí misma y los sistemas CIS vulnerables que respaldan la misión. Los modelos se utilizan luego para simular o recrear los ciberataques y las contramedidas defensivas asociadas, así como las operaciones del sistema, incluida la evaluación del impacto de la misión. En este trabajo se hace referencia a [17], en el cual se hace una

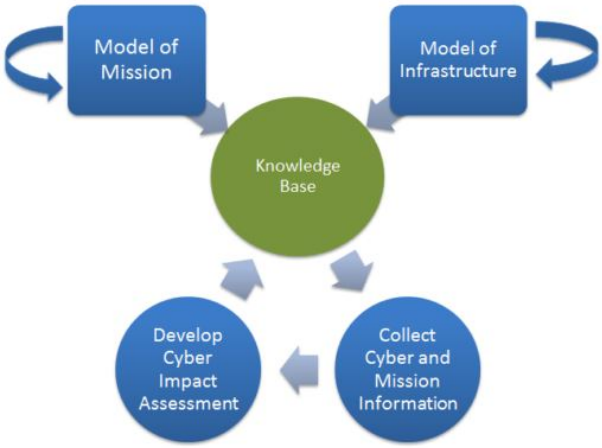


Figura 1. Fases de operación de ARGUS.

revisión del estado de (M&S) así como recomendaciones para integrarlo en ciberdefensa, indicando que los principales retos actualmente son: modelado del comportamiento del atacante y modelado del comportamiento humano, atendiendo no sólo al del atacante sino que también al del defensor.

III. Propuestas de trabajo y objetivos

En el marco global anteriormente expuesto, y partiendo del análisis de las propuestas y desarrollos a la fecha existentes en el ámbito de MIA, el presente trabajo de tesis doctoral se enfoca en el problema del **mapeo de dependencias entre activos y servicos de la red y los objetivos de misión**. Los objetivos particulares planteados para su desarrollo son los siguientes:

1. Buscar nuevos modelos efectivos de mapeo de dependencias de misión bajo constantes cambios en la infraestructura CIS, así como en los objetivos de la misión. El modelo precisaría poder capturar, además, dependencias no obvias para expertos en el área.
2. El modelo deberá admitir el razonamiento automático para acomodar otras capacidades. Por ejemplo, se debería permitir un análisis más profundo hacia la evaluación de impacto en la misión. A este fin, los modelos deberán definirse de una manera estructurada y formal.
3. Desarrollar métricas para estimar el nivel de cobertura del mapeo, e identificar los activos clave (CTK) que tienen mayor impacto en la misión.
4. Prototipado y validación de los modelos y métricas propuestos a través de la disposición de bancos de pruebas adecuados.
5. Los modelos y prototipos resultantes deben admitir análisis de tiempo real en órdenes de magnitud de Big Data, siendo la eficiencia y el rendimiento parámetros clave en su desarrollo.

Para abordar estos objetivos, se plantea como aproximación inicial emplear el esquema de relaciones utilizado en fusionador de información de Cyber-Argus, presentado en [15] (véase Figura 2). Siguiendo este modelo existen varios niveles de relación:

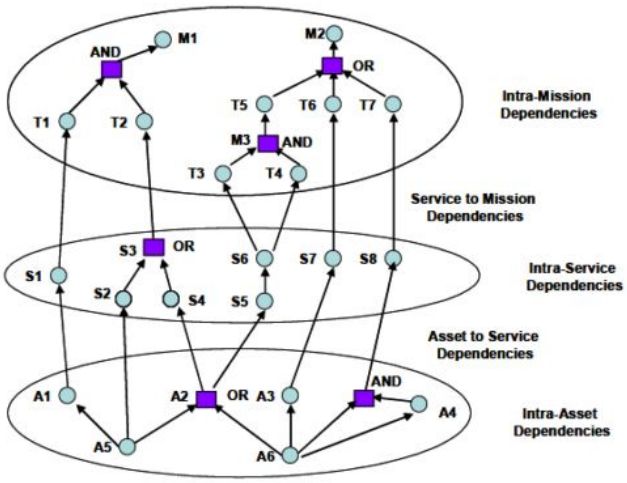


Figura 2. Esquema relaciones para impacto en misión [15].

- Relaciones entre activos. Por ejemplo, la relación que tiene un servidor con los *routers* que le permiten comunicarse.
- Un servicio se sustenta sobre uno o varios activos, siendo esta otro tipo de relación de dependencia. El caso de un servicio de hospedaje web desplegado sobre un *cluster* de servidores puede ser un ejemplo de servicio que depende de varios activos.
- En el nivel de servicios existen también relaciones de dependencia. Así, en el ejemplo del servicio web, este dependerá a su vez de un servicio de DNS para poder cumplir su cometido.
- La relación entre servicio y tarea es posiblemente la que entraña mayor nivel de dificultad de cara a realizar una traducción automática, siendo abordada tradicionalmente por SME.
- A nivel de tarea, la relación entre las mismas está bastante bien establecida en BPMN, por lo que no se espera tener dificultad para emplear dicha relación en la propagación de estados entre tareas. Un ejemplo de relación entre tareas podría ser la necesidad de tener comunicación entre un aeropuerto y un avión (tarea 1) para poder realizar un aterrizaje (tarea 2).
- Sobre el nivel de tarea se encuentra la misión, que no es más que un conjunto de tareas con un objetivo común. Al igual que el nivel de tarea, este viene bien definido en BPMN, no representando en consecuencia un problema a nivel de propagación de estados entre misión-tarea.

Gracias a estas relaciones, se permite ver cómo afecta un incidente en capas bajas, por ejemplo la caída de un *host* por fallo eléctrico, al desempeño de la misión. O, de forma opuesta, ver hasta dónde puede degradarse un servicio para alcanzar una tasa de éxito de un determinado porcentaje en una misión. Con este tipo de modelo relacional se posibilita una visión del estado en el nivel de abstracción que se precise en cada momento, permitiendo presentar la información necesaria en cada momento y cumpliendo así una de las especificaciones deseadas por la OTAN para este tipo de sistemas.

Como se ha comentado anteriormente, el **descubrimiento de servicios de forma automática**, así como la relación entre ellos, es un campo bastante maduro. Por ello se pretende abordar dicho **mapeo** empleando uno de los métodos que ha arrojado resultados más satisfactorios, analizando las latencias de flujos de tráfico de red. Una vez realizado este mapeo, sería interesante explorar la idea de **traducir dicho mapa a simuladores de redes y servicios** existentes como los empleados para planificación de redes. De esta forma, se podría emplear la capacidad de estos para estudiar las consecuencias a nivel de servicios de diferentes ataques, permitiendo escalar esta información a tareas con las relaciones entre servicios y tareas.

Respecto al **descubrimiento de CTK**, se busca explorar una alternativa que busque una aproximación mixta. Este enfoque pretende lograr poder localizar que activo, o servicio es crítico dependiendo del planteamiento de la misión, de igual forma que ver como un cambio en algún activo puede afectar a la misión. Con esta meta, el planteamiento es el siguiente: por un lado se buscan cuáles son los activos críticos para cada servicio. Esta perspectiva recibe el nombre de abajo-arriba y nos permite conocer como se relaciona cada servicio con el resto de servicios y activos, facilitando conocer el estado actual de un servicio conociendo el estado de sus nodos relacionados. Por otro lado, se busca conocer cuáles son los **servicios más relevantes** para la misión en función de qué tareas sustentan estos servicios y la prioridad de dichas tareas dentro de la misión (enfoque arriba-abajo). De esta forma se pretende poder indicar, desde la perspectiva de la planificación de la misión, que tarea es más relevante y por tanto que servicios asociados a la misma lo son. Finalmente, y gracias a las relaciones entre los elementos definidos anteriormente, se realiza la unión de ambas visiones (arriba-abajo y abajo-arriba) obteniendo como resultado una **categorización de los activos o servicios** (depende de la presentación de la información que nos interese en cada momento) según su relevancia para la misión global, teniendo en cuenta además el estado actual de la infraestructura CIS. De estos dos enfoques, es el de abajo-arriba el que requiere más atención y estudio, pues el de nivel de tarea y su relación de importancia con la misión forma parte de la descripción de BPMN.

Todas estas tareas se desarrollarán con carácter prioritario y secuencial en las próximas fechas durante el desarrollo de la tesis doctoral pretendida.

AGRADECIMIENTOS

Este trabajo ha sido financiado parcialmente por el Gobierno de España, con fondos FEDER, a través del proyecto TIN2017-83494-R.

REFERENCIAS

[1] T. Moye, R. Sawilla, R. Sullivan, and P. Lagadec, "Cyber defence situational awareness," Request for Information (UNCLASSIFIED) MNCD2-WP2-0044(1.3), NATO Communications and Information Agency, May 2015.

[2] N. I. A. Group, "Final Report of NIAG Study Group 206 on Cyber Defence Situational Awareness," Technical Report (UNCLASSIFIED) NIAG-D(2016)0027 (PFP), Dec. 2016.

[3] EDA: "EDA 14.cat.op.070. Industry Support to Cat B Project - Cyber Situational Awareness Package (CySAP)," 2014-2015.

[4] EDA: "EDA 16.cat.op.078. target architecture and system requirements for an enhanced Cyber Situational Awareness (CySA)," 2017.

[5] Army: "Federal contract opportunity W15QKN-16-Z-04EM for research and development in the physical, engineering, and life sciences (except biotechnology) at Army Contracting Command ACC - NJ (W15QKN)." Tender, Apr. 2016.

[6] A. E. Schulz, M. C. Kotson, and J. R. Zipkin, "Cyber network mission dependencies (unclassified)," Tech. Rep. 1189, Massachusetts Institute of Technology, Lincoln Laboratory, Lexington, Massachusetts, 2015.
[7] X. Chen, M. Zhang, Z. M. Mao, and P. Bahl, "Automating Network Application Dependency Discovery: Experiences, Limitations, and New Solutions," 8th USENIX Symp. Oper. Syst. Des. Implement., pp. 117–130, 2008.
[8] S. Musman, A. Temin, M. Tanner, R. Fox, B. Pridemore, E. Armistead, and E. Cowan, "Evaluating the impact of cyber attacks on missions," in Proc. of the 5th International Conference on Information Warfare and Security, (Ohio, USA), pp. 446–456, Apr. 2010.
[9] L. W. Fortson, "Towards the development of a defensive cyber damage and mission impact methodology," Master's thesis, Air Force Institute of Technology, Ohio, USA, 2007.
[10] P. Price, N. A. Leyba, M. Gondreey, Z. Staples, and T. Parker, "Asset criticality in mission reconfigurable cyber systems and its contribution to key cyber terrain," in Proc. of the 50th International Conference on Systems Sciences (HICSS 2017), pp. 446–456, 2017.
[11] A. de Barros Barreto, P. C. G. Costa, and E. T. Yano, "Using a semantic approach to cyber impact assessment," in Proc. of the 8th Conference on Semantic Technologies for Intelligence, Defense, and Security (STIDS 2013), (Fairfax, Virginia, USA), p. 101108, Nov. 2013.
[12] A. D'Amico, L. Buchanan, J. Goodall, and P. Walczak, "Mission impact of cyber events: Scenarios and ontology to express the relationships between cyber assets, missions, and users," Tech. Rep. OMB No. 0704-0188, AFRL/RIEF, 2009.
[13] MATHEUS, C. J. et al. SAWA: an Assistant for Higher-Level Fusion and Situation Awareness.Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications, Orlando, USA, v. 5813, p. 75 - 85, 2005
[14] A. D. B. Barreto, P. C. G. Costa, and E. T. Yano, "A semantic approach to evaluate the impact of cyber actions on the physical domain," CEUR Workshop Proc., vol. 966, pp. 64–71, 2014.
[15] A. De, B. Barreto, P. Costa, and M. Hieb, "Cyber-ARGUS: Modeling C2 Impacts of Cyber Attacks," in 19th ICCRTS, 2014, pp. 1–24.
[16] A. Kott, "The significance of model-driven paradigms in cyber security: an introduction," J. Def. Model. Simul., vol. 15, no. 1, pp. 3–4, 2018.
[17] M. Lange et al., "Recommendations for Model-Driven Paradigms for Integrated Approaches to Cyber Defense," 2017.

Definición de procedimientos para fabricar honeypots IoT basados en criterios de búsqueda

Antonio Acien
NICS Lab
Universidad de Málaga
acien@lcc.uma.es

Ana Nieto
NICS Lab
Universidad de Málaga
nieto@lcc.uma.es

Gerardo Fernandez
NICS Lab
Universidad de Málaga
gerardo@lcc.uma.es

Javier Lopez
NICS Lab
Universidad de Málaga
jlm@lcc.uma.es

***Resumen*—Con la revolución tecnológica que ha supuesto la Internet de las Cosas (Internet of Things, IoT) se han presentado escenarios donde la preocupación por la seguridad en dicho entorno es cada vez más relevante. Están comenzando a surgir vulnerabilidades en varios dispositivos, y los sistemas trampa son una excelente manera de lidiar con este problema. En este trabajo se analizan soluciones para honeypots en el entorno IoT (y en otros que se puedan adaptar) para sentar las bases de una metodología que permita el despliegue de honeypots IoT.**

***Index Terms*—IoT, honeypot, seguridad, botnet**

I. Introducción

La IoT está cada vez más presente en la vida diaria, afectando al sector público y empresarial, y a los usuarios que día a día delegan en sus objetos personales parte de su rutina. Las estimaciones de dispositivos conectados para 2020 varían entre diferentes estudios, pero en todo caso superan los veinte mil millones [1]. Cada vez se les da mayor poder y responsabilidad a estos dispositivos (vehículos autónomos, *smart locks* o cerraduras inteligentes), y los usuarios son más dependientes de ellos, haciendo que sean muy suculentos para los atacantes.

El panorama que presentan los dispositivos IoT es muy diverso, pues utilizan tecnologías muy distintas, y realizan funciones totalmente diferentes entre ellos. Cada uno supone una puerta de entrada distinta a ataques y amenazas, y teniendo en cuenta lo integrados que están en la sociedad actual, su seguridad debe ser un aspecto muy a tener en cuenta. A día de hoy, es imposible predecir con exactitud el efecto que tendría un ataque dirigido a la presente infraestructura de red.

Aunque la seguridad en dispositivos IoT se lleva considerando desde varios años atrás, debido a ataques a routers y dispositivos empotrados [2], fue la cantidad de *botnets* (redes de bots) que comenzaron a aparecer lo que hizo saltar las alarmas. Tras ciertos ataques por parte de botnets que usaban dispositivos IoT (Gafgyt, BrickerBot, Tsunami), fue a finales de 2016 cuando Mirai, el que más estragos causó, tuvo lugar, lanzando un ataque de denegación de servicio distribuido (DDoS) que dejó inoperativos sitios como Amazon, Github y Twitter[3]. A medida que la preocupación acerca de estos ataques ha ido creciendo, el presupuesto destinado a seguridad específica en IoT también lo ha ido haciendo, así como la cantidad de malware dirigido a esta plataforma[4].

Teniendo en cuenta el riesgo que supone que la IoT sea un escenario heterogéneo con el que es difícil lidiar, y que tenga un papel tan crucial en muchos casos, es necesario promover medidas para evitar ataques tan pronto como sea posible, entenderlos y analizarlos en un entorno seguro, ya que, como hemos visto, no es realista pensar que no van a ocurrir, o que las medidas presentes a día de hoy son suficientes. En otras palabras, se necesita atraer a los atacantes a sistemas trampa, de manera que obtengamos código que analizar con seguridad. Estos sistemas son los *honeypots* y *honeynets*.

Disponer de honeypots es especialmente útil a la hora del *logging* (registro) de eventos, ya que muchos ataques se esfuerzan en eliminar su rastro, y en el aspecto económico, ya que la emulación y simulación de dispositivos nos permite el despliegue de una gran cantidad de nodos sin un gran coste.

II. Trabajo relacionado

Hasta donde sabemos, no hay trabajos que presenten una metodología para desplegar honeypots destinados para el entorno IoT. Algunos sondeos han estudiado exhaustivamente los honeypots existentes, estableciendo clasificaciones y enfoques, y sí que se encuentran algunas soluciones enfocadas a IoT [5].

Estas soluciones no contemplan el análisis de los requisitos necesarios para desplegar los honeypots, y aquí es donde se centra este artículo: identificar correcta y claramente el conjunto de pasos y procedimientos a seguir para desplegar exitosamente honeypots en el entorno IoT. Este éxito se conseguirá si el sistema es lo suficientemente suculento para atraer atacantes de los que extraer información, y lo suficientemente realista como para no ser detectado como un sistema trampa.

Sin embargo, los honeypots IoT existentes presentan unas características interesantes. Al ser un campo reciente, la cantidad de soluciones es limitada, pero propuestas como IoTPOT [6], SIPHON [7] o IoTCandyJar [8] ofrecen arquitecturas completas de despliegue, las cuales tienen particularidades como redirigir varios honeypots de baja interacción (más simples y menos costosos, pero también menos realistas) a uno o varios de alta, dando una apariencia más real de cara a los atacantes sin aumentar en gran medida el coste. Otros, en cambio, conectan los honeypots a servidores cloud para simular que están distribuidos por el mundo, o usan las respuestas a comandos de otros dispositivos IoT conectados a Internet para elaborar mensajes realistas. Hay algunos trabajos que se basan en estos, haciendo implementaciones de código abierto o mejoras en el soporte de protocolos [9].

Asimismo, hay otros honeypots que no cuentan con una arquitectura detrás que les dé soporte, pero que pueden ser útiles en el entorno IoT, ya que simulan dispositivos como cámaras [5] u otros que cuentan con vulnerabilidades concretas (routers vulnerables a Mirai [10] o que tienen el

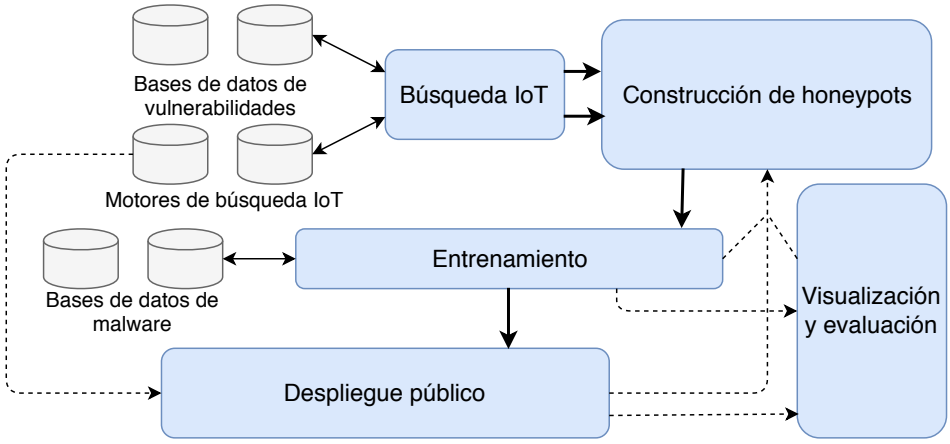


Figura 1. Metodología H-IoT

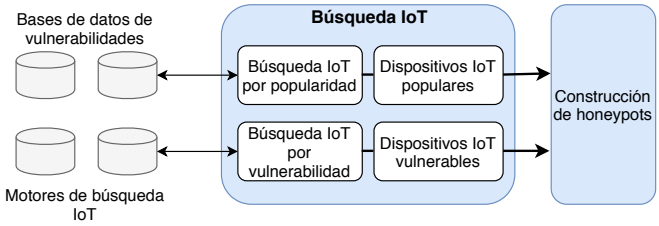


Figura 2. Fase de búsqueda IoT

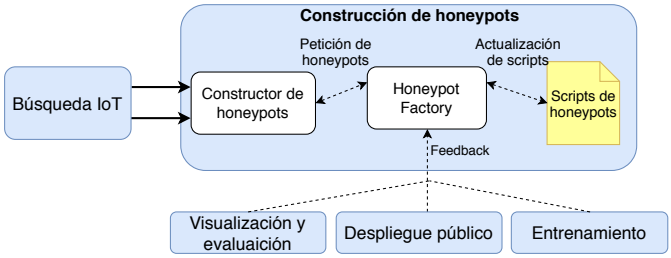


Figura 3. Fase de construcción de honeypots

protocolo TR-064 en un puerto abierto [11]), o bien se centran en ciertos protocolos, personales [12] o industriales [13].

III. Metodología

Aunque, como hemos visto, hay soluciones de honeypots IoT disponibles, éstas por sí solas no son suficientes. No basta con saber cómo se lleva a cabo el despliegue de una herramienta, sino que se necesita saber qué dispositivos simular de cara a los atacantes (cuáles pueden ser más atractivos, o cuáles pueden presentar un comportamiento más interesante), cómo evaluar su funcionamiento, y cómo sacar de esta evaluación información útil para futuras mejoras. A estos interrogantes pretende dar respuesta la metodología aquí propuesta, que va más allá de una simple arquitectura, abogando por cubrir la totalidad del proceso.

III-A. Búsqueda IoT

Los primeros desafíos abordados son hacer el honeypot atractivo para los atacantes, y hacer que el conjunto de honeypots desplegado sea representativo del contexto IoT. El primer punto tiene que hallar un compromiso en el realismo del sistema: si se trata de un sistema muy simple, puede ser obvio que es una trampa para el atacante. Sin embargo, si es un sistema complejo, puede perder interés como sistema vulnerable. El segundo desafío tiene la problemática de la heterogeneidad del entorno IoT, ya que los dispositivos son muy distintos entre sí y tienen comportamientos y respuestas muy diversos. Un honeypot que no satisfaga estas condiciones puede ser demasiado específico, o quedar rápidamente obsoleto.

Para cumplir estas condiciones se han usado varias herramientas. En primer lugar, se han utilizado motores de búsqueda especializados en IoT, tales como SHODAN, Censys o Wigle. Estos motores escanean la red y buscan dispositivos IoT, indexándolos para poder buscarlos en base a tags, localización, sus mensajes de bienvenida (*banners*) y otros criterios. Usándolos se han buscado marcas, modelos y tipos de dispositivos. También se ha hecho uso de portales de venta, tales como Amazon, u otros más especializados, como Network Webcams. Aquí se han observado los dispositivos más vendidos en distintas categorías (routers, drones, cámaras IP...), para ver qué modelos son más populares. Una vez se han obtenido los dispositivos populares, tanto en cantidad de unidades conectadas como en ventas, mediante los medios descritos, se procede a la búsqueda de los mismos en portales de vulnerabilidades. En esta ocasión se han usado Exploitee.rs, ExploitDB y CVE-Details, donde se detallan las vulnerabilidades conocidas para cada modelo y versión, su gravedad, y si hay maneras de explotarlas disponibles.

III-B. Constructor de honeypots

Este paso en la metodología se centra en la elaboración de honeypots usando la información recabada en la fase anterior: dispositivos IoT populares, y dispositivos IoT vulnerables, coincidiendo en ocasiones ambos aspectos. Aunque la implementación de honeypots puede depender de muchos factores (su tipo, su dependencia del hardware...), es importante mantener un repositorio de soluciones ya implementadas, para aumentar la eficiencia. Esto podría ahorrar mucho tiempo si, por ejemplo, se tuviera que implementar un honeypot de características similares a uno ya existente, pero cambiando ciertos valores o parámetros. Este escenario probablemente se dé, ya que los resultados de las fases posteriores, retroalimentarán a estas, para perfilar los honeypots ya existentes.

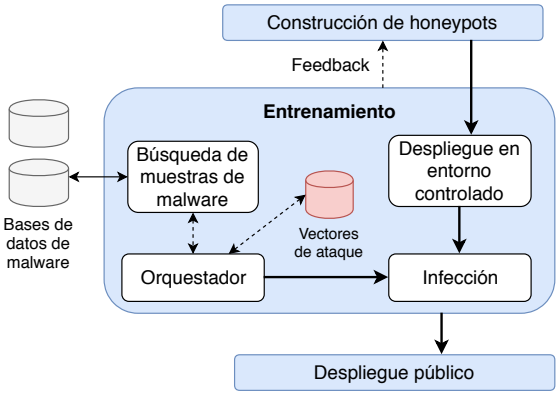


Figura 4. Fase de entrenamiento

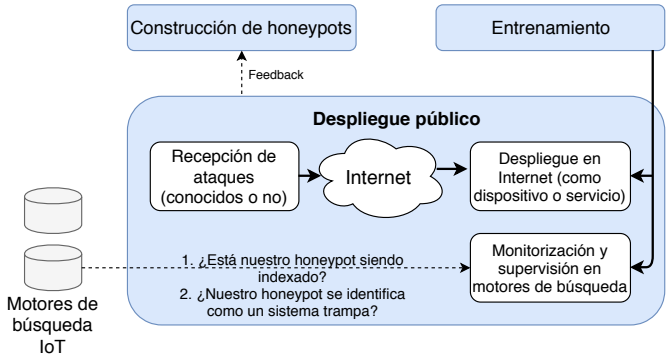


Figura 5. Fase de despliegue público

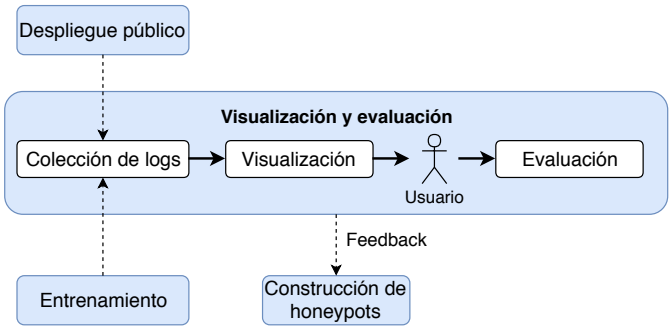


Figura 6. Fase de validación y evaluación

Para clasificar y catalogar estas soluciones, se recomienda usar una Honeypot Factory (Fábrica de honeypots) durante esta fase. Este componente se usará para controlar el acceso y las modificaciones a los archivos que describan la configuración de los honeypots. A su vez, el componente recibirá feedback para clasificar mejor los honeypots y mejorar el Honeypot Builder. Además, el despliegue de los honeypots con esta fábrica, será mucho más eficiente, ya que permitirá desplegar un nuevo honeypot en caso de que uno de los que estén operando se corrompa de alguna manera. En este trabajo, la infección de los sistemas se lleva a cabo de forma manual, por lo que tanto esta fase como la anterior no se ponen en funcionamiento. Sin embargo, son necesarias para que al desarrollar honeypots en el futuro siguiendo esta metodología, éstos cubran todo el entorno IoT y sean útiles.

III-C. Entrenamiento

Durante esta fase, se lleva a cabo el despliegue de los honeypots en un entorno controlado para probarlo. Para ello, el honeypot se infecta usando malware conocido. En algunos casos, infectar un dispositivo no es en absoluto trivial. Uno de los propósitos de este paso en la metodología es evaluar la viabilidad de un honeypot para ser infectado. Hay que destacar que, en algunos casos, es deseable que el honeypot sea lo más vulnerable posible (por ejemplo, para recabar información sobre ataques automatizados u oportunistas), pero en otros debería ser un desafío para el atacante, ya que el honeypot si no presenta unos niveles de seguridad mínimos, las sospechas sobre estar atacando a un sistema trampa crecen.

III-D. Despliegue público

El feedback más útil al sistema implementado mediante esta metodología se obtiene de esta fase. Durante estos pasos, el honeypot se despliega expuesto directamente a internet, disponible para los atacantes. Aunque pueda parecer paradójico, ya que conectar dispositivos IoT directamente a Internet los expone a muchos riesgos, es justo esto lo que motiva este trabajo. Uno de los problemas a los que se puede enfrentar el honeypot desplegado es pasar desapercibido frente a los atacantes, debido a la alta densidad de aparatos IoT disponibles. Y aún peor, si el honeypot es identificado como tal por el atacante, éste podría difundir la información a otros atacantes. El trabajo realizado en las fases previas trata de prevenir esto. Hay que tener en cuenta que se necesita promocionar el sistema trampa para hacer que los motores de búsqueda IoT indexen los honeypots desplegados. Estos sistemas deben ser comprobados a fin de detectar si se ha llevado a cabo correctamente esta indexación, por lo que se requiere una monitorización constante durante esta fase. Por último, los sistemas deben ser realistas. La complejidad de este punto es crucial, puesto que si, por ejemplo, se despliega un honeypot de un dispositivo industrial en un sistema SCADA, para imitar correctamente su funcionamiento, habría que implementar intercambios de comandos con controladores que, en muchas ocasiones, siguen protocolos propietarios, lo cual sería muy costoso de llevar a cabo a la perfección.

III-E. Validación y evaluación

Esta fase es paralela a las de entrenamiento y despliegue público. Sin embargo, los resultados de ambas fases tienen que estar claramente separados y clasificados. Esto nos facilitará saber si los resultados esperados antes de realizar el despliegue se corresponden con los reales obtenidos más tarde. En particular, esta fase está pensada para hacer que la solución sea usable por un administrador o investigador que use el sistema y pueda observar los resultados de manera clara y sencilla. Se ha de enfatizar que los logs pueden ser recogidos en diferentes formatos y esto puede representar un problema a la hora de la visualización, por lo que la traducción entre diferentes formatos se tiene en cuenta en esta fase, para que las herramientas puedan procesarlos correctamente. La metodología también considera el feedback que provee el usuario para mejorar y configurar las plataformas de los honeypots, asegurando así su mantenimiento.

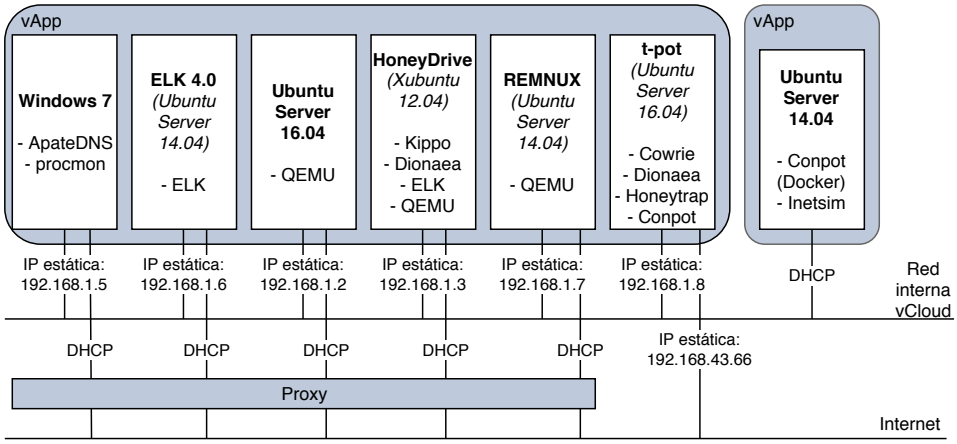


Figura 7. Entorno de despliegue vCloud

IV. Resultados preliminares

En octubre de 2016, un malware específicamente enfocado a dispositivos IoT llamado Mirai causó un ataque distribuido de denegación de servicio contra el proveedor de DNS estadounidense Dyn, haciendo que miles de dispositivos (como routers, DVRs, cámaras IP, etc.) que conservaban las credenciales de acceso por defecto del fabricante formasen parte de una botnet a merced de servidores de *command and control*, que modificaron su comportamiento. La magnitud de este ataque no tenía precedentes y afectó a servicios muy populares. Recientemente, el malware ha dado el salto a plataformas Windows, aunque comportándose de forma distinta. Anteriormente, sólo los dispositivos IoT que fueran visibles de forma pública (expuestos con una IP, como muchas cámaras de seguridad o routers de empresas, por ejemplo) eran susceptibles a Mirai. Ahora, si el malware entra en un equipo Windows, los dispositivos que se encuentren en su red local, incluso tras un firewall, se pueden ver comprometidos [14]. Dada la novedad y repercusión de este ataque, y su aspecto desafiante, se ha usado en la prueba de concepto presentada en este artículo, la cual muestra una infección con Mirai en un equipo Windows dentro de un entorno controlado de pruebas. Se han desplegado otros honeypots en este mismo entorno para llevar a cabo pruebas que quedan fuera del enfoque de este trabajo. Las muestras de malware han sido obtenidas de VirusTotal y Malwr, y el procedimiento seguido para desplegar la infección se ha llevado a cabo según el proceso detallado en SecureList, y se puede ver en Figura9.

IV-A. Entorno

La prueba de concepto se lleva a cabo en el entorno vCloud, el cual permite virtualización, definición de redes y la captura de *snapshots* para restaurar las máquinas a un estado anterior. Además, es un sistema que se puede aislar, de manera que no permite que los ataques desplegados en el mismo se extiendan fuera de él.

Los honeypots desplegados y sus conexiones se muestran en Figura7. Se puede ver cómo se conectan a una red interna que no tiene acceso a Internet. En esta configuración, hay una máquina dedicada específicamente a la centralización de *logs* (registros) para la fase de evaluación y validación, la cual se ha implementado usando el stack ELK, compuesto

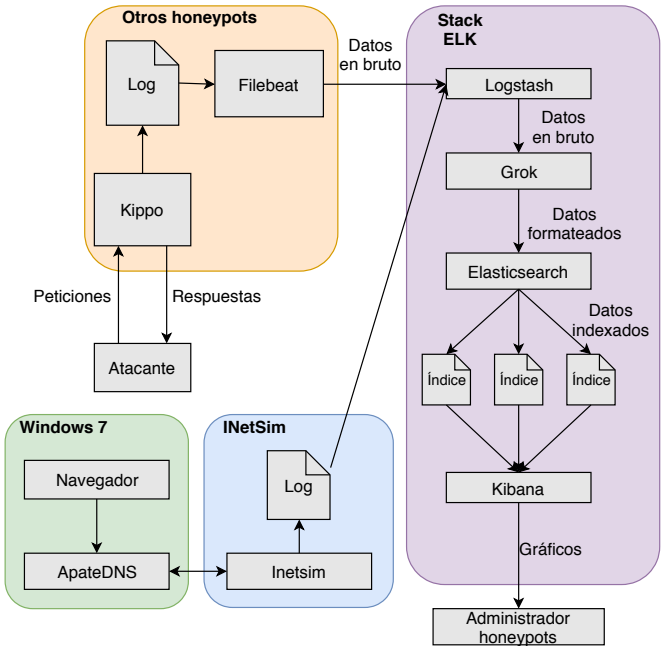


Figura 8. Integración de ELK con los honeypots

por ElasticSearch, Logstash y Kibana. Se han escogido estas aplicaciones por su sencillez de uso, la amplia documentación y la disponibilidad de plugins que aumentan su funcionalidad. La interacción entre estos sistemas es bastante directa: Logstash recibe los registros de las demás máquinas, ElasticSearch los indexa, y Kibana recoge los datos y los presenta de una manera gráfica que resulta sencilla y comprensible. Algunos de los plugins usados han sido Filebeat, para enviar los logs de los honeypots, y Grok, para el *parsing* de los eventos recibidos en los registros. El flujo de la información de los honeypots a la máquina de registro se detalla en Figura8.

Una vez los honeypots estén correctamente conectados a la máquina de centralización de logs, para poder almacenarlos y obtener información de ellos correctamente, y el entorno esté configurado para evitar que el ataque se propague, se puede llevar a cabo la infección.

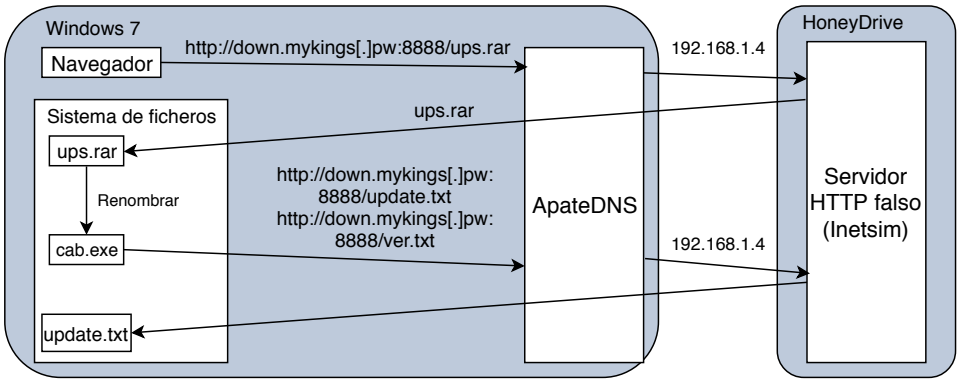


Figura 9. Proceso de infección con Mirai

IV-B. Infección

Dado que el primer paso de la infección es descargar un ejecutable malicioso, y las máquinas no tienen conexión a Internet, tiene que disponerse un servidor HTTP falso. Éste se ha situado en otra máquina del entorno con la herramienta INetSim. Los archivos que se necesitan para la infección se alojan en él, y las peticiones a las URLs donde se alojan los ejecutables de Mirai se redirigen a este servidor mediante ApateDNS. Por tanto, la máquina Windows descarga los archivos como si contara con una conexión.

Una vez se obtiene el primer archivo, se renombra y ejecuta, tal y como se detalla en la guía de infección. Si la ejecución se lleva a cabo con privilegios de administrador, ésta cambia las direcciones DNS primaria y secundaria de la máquina a 114.114.114.114 y 8.8.8.8, respectivamente. En cambio, si se ejecuta como un usuario sin estos permisos, no realiza ningún cambio.

Tras esto, el proceso intenta conectarse de nuevo con URLs del mismo dominio donde se alojaba el ejecutable para descargar los archivos *ver.txt* y *update.txt*, estando disponible el primero en el las bases de datos con muestras de malware a las que teníamos acceso, pero no el segundo. La salida de consola de esta ejecución es "*DNS set ok. ver different web:1.0.0.8 local:, needs update...*", refiriéndose la última parte a que falta el archivo *update.txt*, por lo cual la ejecución se detiene.

Si el ejecutable intenta descargar los archivos sin la herramienta ApateDNS redirigiendo las peticiones y sin conexión a Internet, el mensaje que se muestra es "*get file list failed, exit*", justo antes de detenerse.

Todos los eventos en este proceso se registran en *logs* que se envían al *stack* ELK. Los mensajes se envían mediante el plugin Filebeat, se procesan mediante Grok, se indexan y visualizan. Los mensajes de este proceso se separan del resto con lo que Kibana denomina *patrones*, unos prefijos que ayudan a diferenciar el origen de los *logs*. Además, el tráfico generado en las peticiones HTTP a los servidores que alojan los archivos de Mirai y el intercambio de los mismos se ha registrado mediante la herramienta de captura de paquetes Wireshark.

V. Discusión y trabajo futuro

Hay muchos más casos cubiertos por la metodología de los que se muestran en la prueba de concepto de este trabajo. Un paso fundamental sería la definición de honeypots IoT basada en los criterios discutidos en este trabajo, ya que la decisión de desplegar un honeypot para Mirai fue basada más en su reciente impacto que en los resultados obtenidos al seguir la metodología, pero definir un conjunto de honeypots IoT que desplegar para capturar ataques desconocidos es un objetivo a conseguir en el trabajo futuro. Aunque ya se está trabajando en este aspecto, no podemos dar detalles del proceso hasta que se obtengan datos sustanciales, ya que los resultados podrían verse afectados si se comparte información específica del mismo.

La fase de entrenamiento es esencial antes de que se considere siquiera el despliegue público de los honeypots, por lo que es en lo que se ha centrado este trabajo. Sin embargo, algunas cuestiones que no han sido consideradas en el trabajo pero merece la pena mencionar son las siguientes.

- **Infección desde una botnet Mirai.** En lugar de descargar el malware y realizar la infección manualmente, sería interesante provocar que la máquina se infecte desde una botnet de Mirai ya existente. Este enfoque permitiría ver cómo la máquina es indexada por los mecanismos de monitorización de dispositivos con Mirai. No sería sencillo, puesto que significaría dejar la máquina expuesta al exterior y contribuiría a posibles propagaciones.
- **Infección a red de dispositivos IoT desde Windows.** Crear una honeynet para este caso específico, en el que el equipo Windows esté conectado con dispositivos IoT y ver cómo se infectan.
- **Emulación de dispositivos IoT con QEMU.** Mediante una plataforma de emulación, y disponiendo de firmware de los dispositivos a tratar, podrían crearse máquinas que simularan ser dispositivos IoT vulnerables a modo de honeypot, pero con una interacción completa. Este procedimiento es complejo, porque en muchas ocasiones el firmware presenta una fuerte dependencia de la plataforma hardware a la que se destina (arquitecturas de los procesadores, sistemas de archivos...).
- **Conexión con ELK para que muestre resultados del ataque y sistema de inteligencia.** Consiste en hacer que desde HoneyDrive se envíen las acciones a la máquina ELK, y que el sistema de inteligencia se dé cuenta basándose en el comportamiento (peticiones, mensajes, timing) de que se trata de Mirai. Ya que este malware en concreto se destruye automáticamente si el equipo

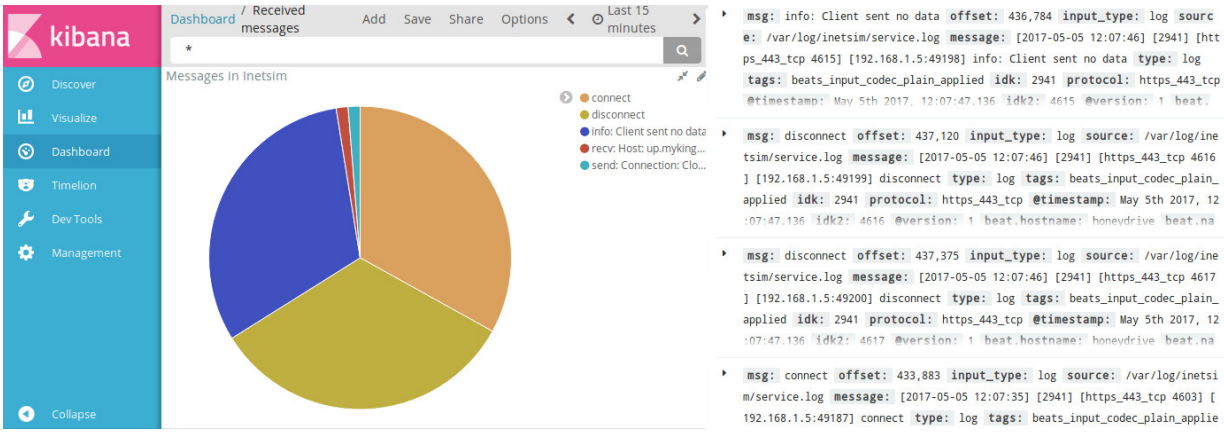


Figura 10. Recopilación y visualización de mensajes de logs en Kibana

se queda sin conexión tras la infección, decidir aislar automáticamente el equipo Windows 7 (infectado) hasta que se copie el código de Mirai, las evidencias y aquellos ficheros relevantes para el análisis del malware.

- **Desplegar "infección" de Hajime o dejarse infectar por una botnet.** Como se ha descrito anteriormente, Hajime se trata de una botnet que basa su funcionamiento en el mismo patrón que Mirai: entra a los dispositivos que conserven las credenciales por defecto del fabricante mediante telnet, usando prácticamente la misma lista de usuarios y contraseñas. Sin embargo, una vez entra, bloquea el acceso desde varios puertos, para evitar ataques vía telnet o TR-064. Al contrario que Mirai, no tiene servidores de *command and control*, sino que se basa en un esquema *peer-to-peer*, lo que hace más difícil detener su avance, y no tiene (por el momento) capacidad de efectuar ataques de DDoS [15].

VI. Conclusiones

A lo largo del trabajo, se ha detallado cómo el escenario IoT no sólo es novedoso y desafiante, sino una realidad palpable. La mayoría de las personas que utilizan internet en su vida diaria contarán con uno o varios dispositivos IoT, e interactuarán con una infinidad a lo largo del día, muchas veces sin ser siquiera conscientes de ello. Los dispositivos IoT no han sido diseñados con las suficientes medidas de seguridad, y se les han dado roles muy importantes, tales como preservar nuestra privacidad o seguridad. La situación resultante es un aparato muy vulnerable llevando a cabo tareas muy sensibles. La dificultad para obtener muestras de malware procedentes de ataques IoT, junto con la posibilidad de simular o emular plataformas sin contar con dispositivos reales, la capa de aislamiento que proveen, el registro de eventos (dado que los ataques IoT tienden a borrar su rastro) y otras particularidades de este escenario, hacen que los honeypots sean la herramienta perfecta para analizar el entorno de seguridad IoT. Aunque las herramientas de honeypots IoT están disponibles, se debe conocer antes qué honeypots se deben desplegar, cómo se puede atraer a los atacantes, y cómo se pueden mejorar los sistemas en base a la información obtenida, por lo que es necesaria una metodología que tenga en cuenta todas las partes de este proceso. Cada fase se detalla meticulosamente en la metodología propuesta, y tienen relevancia justificada a lo largo del proceso, como se demuestra en la prueba de concepto llevada a cabo. En esta prueba se despliega uno de los ataques IoT más devastadores en un entorno controlado, mostrando los datos obtenidos al realizarse dicho proceso, el cual se ha llevado a cabo siguiendo esta metodología.

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Economía y Competitividad mediante los proyectos IoTest (TIN2015-72634-EXP) y SMOG (TIN2016-79095-C2-1-R). El segundo autor ha sido financiado por INCIBE a través del programa de ayudas para la excelencia de los equipos de investigación avanzada en ciberseguridad.

Referencias

[1] T. Danova, "Morgan stanley: 75 billion devices will be connected to the internet of things by 2020," *Business Insider*, vol. 2, 2013.
[2] R. Roman, P. Najera, and J. Lopez, "Securing the internet of things," *Computer*, vol. 44, no. 9, pp. 51–58, 2011.
[3] G. Kambourakis, C. Kolias, and A. Stavrou, "The mirai botnet and the iot zombie armies," in *Military Communications Conference (MILCOM), MILCOM 2017-2017 IEEE*. IEEE, 2017, pp. 267–272.
[4] B. Insider, "This one chart explains why cybersecurity is so important," *Retrieved August*, vol. 16, p. 2016, 2016.
[5] M. Nawrocki, M. Wählisch, T. C. Schmidt, C. Keil, and J. Schönfelder, "A survey on honeypot software and data analysis," *arXiv preprint arXiv:1608.06249*, 2016.
[6] Y. M. P. Pa, S. Suzuki, K. Yoshioka, T. Matsumoto, T. Kasama, and C. Rossow, "Iotpot: analysing the rise of iot compromises," *EMU*, vol. 9, p. 1, 2015.
[7] J. D. Guarnizo, A. Tambe, S. S. Bhunia, M. Ochoa, N. O. Tippenhauer, A. Shabtai, and Y. Elovici, "Siphon: Towards scalable high-interaction physical honeypots," in *Proceedings of the 3rd ACM Workshop on Cyber-Physical System Security*. ACM, 2017, pp. 57–68.
[8] T. Luo, Z. Xu, X. Jin, Y. Jia, and X. Ouyang, "Iotcandyjar: Towards an intelligent-interaction honeypot for iot devices," *Black Hat*, 2017.
[9] P. Krishnaprasad, "Capturing attacks on iot devices with a multi-purpose iot honeypot," Ph.D. dissertation, PhD thesis, INDIAN INSTITUTE OF TECHNOLOGY KANPUR, 2017.
[10] A. Radice, "Playing with a mirai honeypot: Mtpot," 2017.
[11] M. Wang, J. Santillan, and F. Kuipers, "Thingpot: an interactive internet-of-things honeypot," 2017.
[12] S. Dowling, M. Schukat, and H. Melvin, "A zigbee honeypot to assess iot cyberattack behaviour," in *Signals and Systems Conference (ISSC), 2017 28th Irish*. IEEE, 2017, pp. 1–6.
[13] A. Jicha, M. Patton, and H. Chen, "Scada honeypots: An in-depth analysis of conpot," in *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on*. IEEE, 2016, pp. 196–198.
[14] K. Lab, "A windows-based spreader for mirai malware has been discovered," 2017.
[15] S. Edwards and I. Profetis, "Hajime: Analysis of a decentralized internet worm for iot devices," *Rapidity Networks*, vol. 16, 2016.

MEDEA: Monitorizando el espacio nulo para la detección de anomalías en sistemas industriales

Ekhi Zugasti, Mikel Iturbe, Iñaki Garitano, Urko Zurutuza
Depto. de Electrónica e Informática
Mondragon Unibertsitatea
Goiru 2, E-20500 Arrasate-Mondragón
Email: {ezugasti,miturbe,igaritano,uzurutuza}@mondragon.edu

***Resumen*—Los entornos industriales han cambiado ampliamente desde la concepción de las redes primitivas y aisladas hasta la conectividad total actual, donde la interconexión entre diferentes dispositivos y servicios se ha convertido en necesidad de negocio. Sin embargo, las prácticas de seguridad no han evolucionado al mismo nivel, debido parcialmente a las prácticas de seguridad heredadas de cuando este tipo de redes no estaban interconectadas y la inexistencia de medidas de seguridad en los nodos que las componen. En este trabajo, presentamos un sistema de detección de anomalías para entornos industriales que monitoriza las magnitudes físicas del proceso para detectar intrusiones. El sistema está basado en la monitorización del espacio nulo, que a su vez está basado en la identificación de subespacios estocásticos (SSI). La propuesta se valida utilizando el célebre proceso Tennessee-Eastman.**

***Index Terms*—detección de anomalías; redes industriales; sistemas de control industrial; espacio nulo;**

I. Introducción

Los Sistemas de Control Industrial (SCI) forman el núcleo de las redes industriales, y como tal, son los agentes especializados en automatizar, controlar y monitorizar procesos industriales de diversa índole. Los controladores lógicos programables (PLC), son, quizás, el ejemplo más representativo de este tipo de sistemas, como responsable habitual del primer grado de control en el proceso y de reenvio de datos de proceso a los sistemas y servidores presentes en capas superiores.

Históricamente, la seguridad de estos sistemas se ha basado en la seguridad por oscuridad y el aislamiento de las redes de proceso y control de las redes de las tecnologías de información (IT). Sin embargo, con la llegada del paradigma de la digitalización de la industria, las necesidades de negocio han llevado a que este tipos de redes estén cada vez más interconectadas, para proveer servicios de valor añadido como pueden ser mantenimiento predictivo o la optimización en tiempo real. Además, esta interconexión ha llevado a su vez a que las tecnologías utilizadas por los SCI sean cada vez más estándares y públicas.

Por lo tanto, las dos bases en las que se ha basado la estrategia de protección de estos sistemas ya no son útiles para preservar su seguridad. Aunque la creciente interconexión de SCI es algo relativamente reciente, los largos tiempos de vida del que disponen estos dispositivos han hecho que gran parte del parque en funcionamiento no dispone de medidas de seguridad ni ha sido diseñado con la interconectividad en mente [1].

Así, después de ataques sonados como Stuxnet [2] y Duqu [3], los sistemas de deteccion de intrusiones (SDI) y en particular, los sistemas de detección de anomalías (SDA), han sido los mecanismos de seguridad que más atención han recibido por parte de la comunidad científica para hacer frente a estos ataques.

En este artículo presentamos MEDEA, un SDA que monitoriza las magnitudes físicas del proceso en sí para detectar intrusiones. En las siguientes secciones presentamos los trabajos relacionados en el campo de la detección de intrusiones en SCI, el marco conceptual del SDA presentado y su validación con un caso de uso y los resultados obtenidos. Por último, se extraen las conclusiones.

II. Trabajos relacionados

El campo relativo a la detección de intrusiones en entornos industriales ha sido ampliamente estudiado por la comunidad científica. Tradicionalmente, la mayoría de las propuestas se han centrado en la capa de red, adaptando aproximaciones basadas en IT al mundo industrial, o utilizando la naturaleza estática del tráfico industrial para la detección de anomalías (se pueden consultar los siguientes artículos [4], [5], [6] para un análisis sistemático de las contribuciones en este campo).

Más recientemente, se ha pasado de la monitorización de la red a la monitorización de magnitudes físicas para la detección de ataques. Urbina y col. [7] y Ding y col. [8] analizaron y clasificaron los trabajos realizados en este campo. La mayoría de los trabajos analizados están basados en utilizar predicciones, es decir, predicen el siguiente valor (basándose principalmente en modelos auto-regresivos y sistemas dinámicos lineales) y luego comparan el valor obtenido con la predicción. Si la diferencia entre ellos sobrepasa un umbral establecido, se clasifica el evento como anómalo.

Otras propuestas se han basado en métodos guiados por datos, donde no se construye ningún modelo ni se realiza ninguna predicción. Dichas propuestas se han basado en diferentes técnicas, como contramedidas de causalidad basadas en la entropía de transferencia [9], control estadístico multivariante de procesos [10] y clusterización [11], [12].

Complementando dichas propuestas, MEDEA provee de un método guiado por datos, con un único estadístico de detección (en el caso de [10], [9], es necesario monitorizar dos estadísticos de detección), es capaz de detectar ataques en contínuo, (pero no en tiempo real) contra los métodos basados en clusters [11], [13], donde la detección se realiza basándose en un conjunto de datos histórico.

III. Detección de anomalías basado en el espacio nulo

Los entornos industriales son, en esencia, entornos multivariantes, donde los sistemas de control industrial monitorizan una variedad de magnitudes físicas, que luego se utilizan para el control de proceso o proveer de servicios de más alto nivel.

Por lo tanto, es natural utilizar un sistema de detección de anomalías (SDA) que tenga en cuenta esta naturaleza multivariante para detectar anomalías. Aquí, proponemos un SDA basado en el método de detección del espacio nulo. Este metodo ya ha sido utilizado con éxito en otros campos, como puede ser la detección de daños. En este trabajo, adaptamos la detección basada en el espacio nulo para la detección de intrusiones en entornos industriales.

El método del espacio nulo está basado en la identificación de subespacios estocásticos (SSI) [14] y utiliza las señales de medición del proceso (dadas por la sensórica etc.) como entrada. La matriz de identificación utilizada para este método de detección de anomalías es la matriz Hankel.

La respuesta estocástica de un sistema que es dependiente del tiempo puede ser considerada de la siguiente forma:

$$\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_m] \tag{1}$$

donde $\boldsymbol{y}_k$ es el vector de mediciones, formado por m mediciones del proceso en un momento dado.

La matriz Hankel puede ser calculada de dos formas: basándose en covarianzas o en datos. La basada en covarianzas se construye de la siguiente forma:

$$\boldsymbol{H}_{p,q} = \begin{bmatrix} \boldsymbol{\Lambda}_1 & \boldsymbol{\Lambda}_2 & \boldsymbol{\Lambda}_3 & \ldots & \boldsymbol{\Lambda}_q \\ \boldsymbol{\Lambda}_2 & \boldsymbol{\Lambda}_3 & \ldots & \ldots & \vdots \\ \boldsymbol{\Lambda}_3 & \ldots & \ldots & \ldots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \boldsymbol{\Lambda}_{p+1} & \ldots & \ldots & \ldots & \boldsymbol{\Lambda}_{p+q} \end{bmatrix} \tag{2}$$

donde p y q son los parámetros definidos por el usuario y $\boldsymbol{\Lambda}_i$ representa una estimación de la matriz de correlaciones entre las diferentes mediciones recibidas.

La matriz por bloques Hankel definida en SSI es un conjunto de matrices creada a partir del desplazamiento de la matriz de datos original. Se pueden estimar diferentes $\boldsymbol{\Lambda}_i$ a partir de un conjunto de $\boldsymbol{y}_k$ mediciones, de la siguiente forma:

$$\boldsymbol{\Lambda}_i = \left(\frac{1}{N-i-1}\right) \sum_{k=1}^{N-i} \boldsymbol{y}_{k+i} \boldsymbol{y}_k^t \tag{3}$$

En este modelo, la detección de anomalías se basa en la extracción del espacio nulo. Para encontrar este espacio nulo, la matriz Hankel se descompone en valores singulares:

$$\boldsymbol{H}_{pq} = \boldsymbol{U}_H \boldsymbol{S}_H \boldsymbol{V}_H^* \tag{4}$$

donde $\boldsymbol{U}_H$ es una matriz unitaria, $\boldsymbol{S}_H$ es una matriz diagonal rectangular con números reales no negativos en la diagonal y $\boldsymbol{V}_H^*$ (la conjungada traspuesta de $\boldsymbol{V}_H$) es una matriz unitaria. Los elementos diagonales $S_{i,j}$ de $\boldsymbol{S}_H$ son conocidos como los valores singulares de $\boldsymbol{H}_{p,q}$. Las columnas de $\boldsymbol{U}_H$ y $\boldsymbol{V}_H^*$ corresponden al vector singular izquierdo y derecho, respectivamente. El espacio nulo de la matriz Hankel ($\boldsymbol{U}_{H0}$) es una matriz que cumple con la siguiente propiedad:

$$\boldsymbol{U}_{H0}^t \boldsymbol{H}_{pq} = [\boldsymbol{0}] \tag{5}$$

Esta hipótesis nula es la base de MEDEA. Básicamente, las mediciones tomadas del proceso en condiciones de operación normales deben cumplir esta hipótesis. Por lo tanto, el residuo de una serie de mediciones a examinar puede ser definido de esta forma:

$$\boldsymbol{R} = \boldsymbol{U}_{H0}^t \boldsymbol{H}_{ij} \tag{6}$$

Por lo tanto, utilizando un $\boldsymbol{U}_{H0}$ calculado durante una fase de entrenamiento en condiciones libres de ataques, los residuos obtenidos serán mínimos. Sin embargo, si los residuos se calculan cuando tiene lugar una disrupción del proceso o está siendo atacado, los valores de los residuos serán significativos.

$\boldsymbol{U}_{H0}$ contiene el número máximo de vectores columna independientes que cubren el espacio nulo de $\boldsymbol{H}_{p,q}$. Para poder encontrarlo, es pues necesario encontrar los valores singulares $S_{i,j}$ que son iguales a cero y luego tomar los vectores singulares izquierdos que corresponden a dichos valores singulares nulos.

El modelo básico se construye utilizando este espacio nulo, pero para dar variabilidad al modelo es necesario disponer de más datos de entrada (mediciones de proceso). Para ello, en la fase de entrenamiento, utilizamos varios conjuntos de datos libres de ataques para dar más variabilidad al modelo inicial. Esta implementación del algoritmo se puede encontrar en [15].

Este algoritmo devuelve un indicador que definiremos como el indicador de anomalía (IA), que mide la distancia euclidiana entre el espacio nulo y el residuo de los datos monitorizados. Para poder determinar si está aconteciendo un evento anómalo o no, es necesario determinar un umbral para el IA que al sobrepasarlo, el evento se clasifica como anómalo. Este umbral se calcula en la fase de entrenamiento, utilizando datos libres de ataques. Se calculan los indicadores $\boldsymbol{IA}_H$ de varias mediciones y se establece el umbral θ de la siguiente forma:

$$\theta = \mu(\boldsymbol{IA}_H) + 3\sigma(\boldsymbol{IA}_H) \tag{7}$$

donde μ y σ son la media y la desviación estándar, respectivamente.

IV. Caso de estudio

En esta sección presentamos el marco de validación utilizado para evaluar el sistema de detección de anomalías (SDA) MEDEA, lo que incluye el proceso utilizado para la creación de datos como el modelo de ataque.

IV-A. El proceso Tennessee-Eastman

El proceso Tennessee-Eastman (TE) es un modelo de un proceso químico real, presentado por Downs y Vogel [16] como proceso de referencia para evaluar diferentes estrategias de control.

Modelado a partir de un proceso de la compañía química Eastman, el TE produce dos productos líquidos a partir de cuatro reactivos gaseosos. Hay, además un producto secundario y un inerte. Los reactivos se transportan mediante tres flujos que van a parar al reactor. Ahí, los reactivos reaccionan y forman los productos líquidos. Después, el proceso purifica

el producto separándolo de las impurezas que forman los reactivos sobrantes. El proceso se monitoriza utilizando 41 variables medidas (XMEAS) y 12 variables manipuladas (XMV). De las variables medidas, 22 de ellas son leídas contínuamente, mientras que del resto se toman muestras a intervalos fijos. Estas últimas variables de medición están relacionadas con el control de calidad del producto final y no tienen ningún impacto en el control del proceso. Para obtener más información acerca del proceso, el lector puede dirigirse a la publicación original [16].

Si bien el TE se desarrolló para fomentar la comparación entre diferentes propuestas de control, su utilidad y la falta de modelos físicos en general, han potenciado su uso en la investigación de ciberseguridad de los sistemas de control [17], [18], [12], [11].

En este trabajo se ha utilizado la estrategia de control propuesta por Larsson y col. [19], junto con el modelo de aleatoriedad presentado por Krotofil y col. [12]. El modelo utilizado, llamado DVCP-TE, está publicado y es de código abierto. [1]

IV-B. Modelo de ataque

Los modelos de adversario utilizados en este escenario son los propuestos por Krotofil y col. [12]. Consideramos que el adversario es capaz de leer y manipular tráfico de red entre los controladores y el proceso físico, pudiendo realizar ataques de integridad y de denegación de servicio (DoS).

En un ataque, consideramos una variable atacada $Y_i'(t)$ en el momento t, $0 \leq t \leq T$ como sigue, donde T es la duración de la simulación y T_a el intervalo arbitrario de ataque. Definimos así un ataque de integridad:

$$Y_i'(t) = \begin{cases} Y_i(t), & \text{para } t \notin T_a \\ Y_i^a(t), & \text{para } t \in T_a \end{cases} \tag{8}$$

donde $Y_i^a(t)$ es el valor modificado de la variable que ha inyectado el atacante.

De forma similar, durante un ataque DoS, el atacante es capaz de parar la comunicación, por lo que no es posible que la señal llegue al actuador o el controlador. Krotofil y col. [12] definen un ataque DoS que empieza en t_a de la siguiente forma:

$$Y_i^a(t) = Y_i(t_a - 1) \tag{9}$$

donde Y_i^a es el último valor recibido antes del comienzo del ataque.

La tabla I muestra los ataques realizados. Las variables escogidas reflejan diferentes magnitudes físicas presentes en el TE. Todas las simulaciones tienen una longitud de 72h, excepto en las simulaciones donde se para la ejecución debido a que deja de operar en límites seguros, y el ataque comienza después de las primeras 24 horas. En el caso de ataques de integridad, hemos calculado el valor medio de la variable en la fase de entrenamiento y hemos realizado un ataque de integridad inyectando un valor 10 % mayor al de la media, que no cambió hasta el final de la simulación. El ratio de muestreo es de 100 observaciones por hora.

[1] https://github.com/satejnik/DVCP-TE

V. Resultados

Los resultados obtenidos para los ataques mostrados en la tabla I, se muestran en las figuras 1, 2, 3, 4 y 5. En cada una de las figuras, se muestra el valor del indicador de anomalía (IA) durante la duración de la simulación. La línea horizontal discontinua roja corresponde al valor del umbral θ. El conjunto de datos de entrenamiento corresponde a una simulación de 72 horas de duración del proceso TE, libre de ataques y de perturbaciones.

Por cada figura, se han establecido una serie de ventanas de cinco horas de duración, donde las observaciones de cada ventana (500 por ventana para el ratio de muestreo utilizado) son utilizados para calcular el indicador de anomalía o de ataque. Por ello, el IA se calcula cada cinco horas y su valor corresponde a las mediciones obtenidas en las cinco horas anteriores. Por ejemplo, cuando el IA se muestra en la hora 20, contiene la información de los sensores obtenida entre las horas 15–20.

Las figuras muestran que el método es capaz de detectar todos los ataques (tanto de integridad como DoS), ya que el IA cruza el valor del umbral establecido en todos los casos.

De las simulaciones realizadas, el ataque que concierne a XMEAS1 (mostrado en la figura 1) es el único en terminar en la hora 72, la longitud original de la simulación. En el resto de ataques, el proceso llega a sus límites de seguridad antes de finalizar la simulación y activa el apagado automático del proceso.

En el caso de los ataques mostrados en las figuras 2, 3 y 5, el apagado del proceso es casi inmediato, mientras que en el caso relativo al DoS sobre XMEAS14, el algoritmo de control mantiene el proceso en marcha hasta la hora 40.

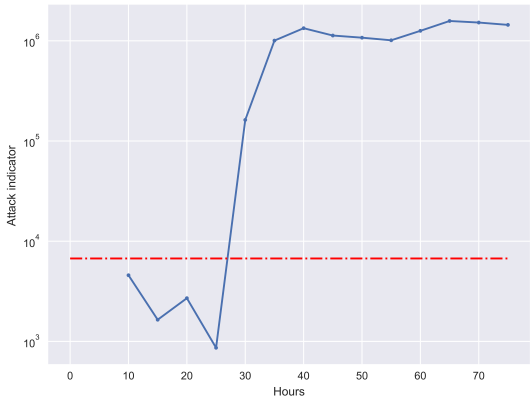


Figura 1. Detección de ataque de integridad sobre XMEAS1

VI. Conclusiones

La detección de ataques en entornos industriales es aún un desafío abierto, donde se hace necesario mejorar propuestas existentes en el campo de la detección de intrusiones para poder hacer frente a amenzadas derivadas de la interconexión de las redes industriales. Hemos presentado un sistema de detección de anomalías que se basa en el análisis del espacio nulo para detectar anomalías a nivel de campo mediante el análisis de señales provenientes de sensores. El SDA propuesto calcula un indicador de anomalía en un conjunto de lecturas comprobando si éste cumple la hipótesis nula de una

Núm. de variable	Nombre de variable	Valor medio	Unidades	Tipo de ataque	Valor inyectado
XMEAS1	Flujo de entrada A	0.265	kscmh	Integridad	0.292
XMEAS8	Nivel de reactor	65.001	%	Integridad	71.5
XMEAS9	Temperatura de reactor	122.90	°C	DoS	N/A
XMEAS14	Flujo inferior del separador	25.35	m^3h^{-1}	DoS	N/A
XMEAS17	Flujo inferior de la columna de separación	22.89	m^3h^{-1}	Integridad	25.18

Tabla I
ATAQUES REALIZADOS

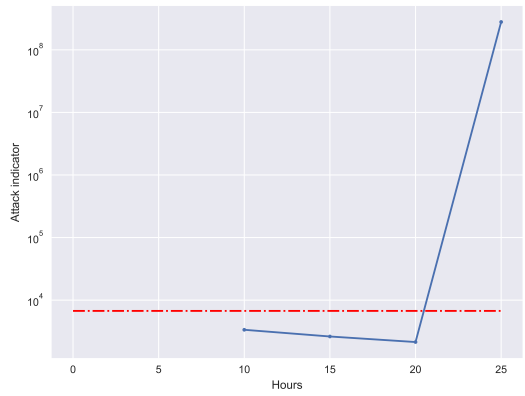


Figura 2. Detección de ataque de integridad sobre XMEAS8

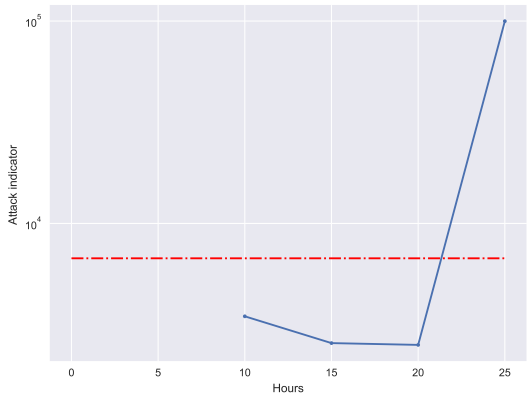


Figura 5. Detección de ataque de integridad sobre XEMAS17

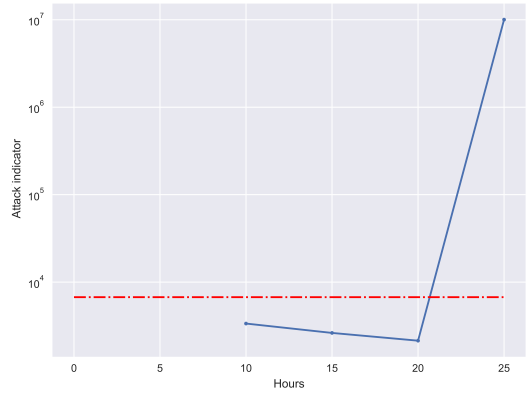


Figura 3. Detección de DoS sobre XMEAS9

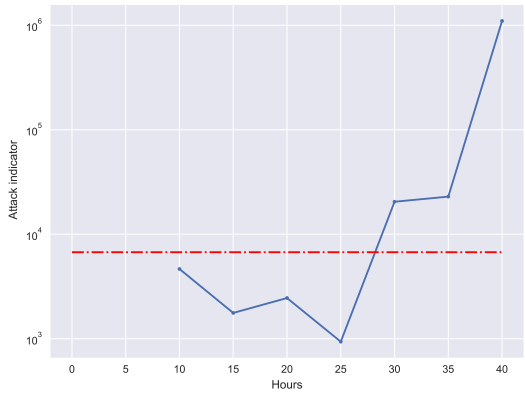


Figura 4. Detección de DoS sobre XMEAS14

captura libre de datos. La propuesta se ha validado utilizando el proceso Tennessee-Eastman y los resultados preliminares muestran que el sistema es capaz de detectar ataques de integridad y DoS.

AGRADECIMIENTOS

Este trabajo 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. Este trabajo está parcialmente financiado por el programa Elkartek del Gobierno Vasco a través del proyecto KK-2018/00076. Finalmente, Iñaki Garitano está parcialmente financiado por la beca número "INCIBEC-2015-02495", correspondiente a las "Ayudas para la Excelencia de los Equipos de Investigación avanzada en ciberseguridad" del Instituto Nacional de Ciberseguridad de España (INCIBE).

REFERENCIAS

[1] M. Cheminod, L. Durante, and A. Valenzano, "Review of Security Issues in Industrial Networks," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 1, pp. 277–293, 2013.

[2] N. Falliere, L. O. Murchu, and E. Chien, "W32.Stuxnet dossier," *White paper, Symantec Corp., Security Response*, 2011.

[3] B. Bencsáth, G. Pék, L. Buttyán, and M. Félegyházi, "Duqu: Analysis, detection, and lessons learned," in *ACM European Workshop on System Security (EuroSec)*, 2012.

[4] I. Garitano, R. Uribeetxeberria, and U. Zurutuza, "A review of SCADA anomaly detection systems," in *Soft Computing Models in Industrial and Environmental Applications, 6th International Conference SOCO 2011*. Springer, 2011, pp. 357–366.

[5] R. Mitchell and I. Chen, "A Survey of Intrusion Detection Techniques for Cyber Physical Systems," *ACM Computing Surveys*, vol. 46, no. 4, April 2014.

[6] B. Zhu, A. Joseph, and S. Sastry, "A taxonomy of cyber attacks on SCADA systems," in *Internet of things (iThings/CPSCom), 2011 international conference on and 4th international conference on cyber, physical and social computing*. IEEE, 2011, pp. 380–388.

[7] D. I. Urbina, J. Giraldo, A. A. Cardenas, J. Valente, M. Faisal, N. O. Tippenhauer, J. Ruths, R. Candell, and H. Sandberg, "Survey and new directions for physics-based attack detection in control systems. NIST GCR 16-010," National Institute of Standards and Technology, Tech. Rep., Nov 2016.

[8] D. Ding, Q.-L. Han, Y. Xiang, X. Ge, and X.-M. Zhang, "A survey on security control and attack detection for industrial cyber-physical systems," *Neurocomputing*, vol. 275, pp. 1674–1683, 2018.

[9] D. Shi, Z. Guo, K. H. Johansson, and L. Shi, "Causality countermeasures for anomaly detection in cyber-physical systems," *IEEE Transactions on Automatic Control*, 2017.

[10] M. Iturbe, J. Camacho, I. Garitano, U. Zurutuza, and R. Uribeetxeberria, "On the feasibility of distinguishing between process disturbances and intrusions in process control systems using multivariate statistical process control," in *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshop (DSN-W)*. Toulouse, France: IEEE, Jun. 2016, pp. 155–160.

[11] I. Kiss, B. Genge, and P. Haller, "A clustering-based approach to detect cyber attacks in process control systems," in *Industrial Informatics (INDIN), 2015 IEEE 13th International Conference on*, July 2015, pp. 142–148.

[12] M. Krotofil, J. Larson, and D. Gollmann, "The Process Matters: Ensuring Data Veracity in Cyber-Physical Systems," in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, ser. ASIA CCS '15. ACM, 2015, pp. 133–144.

[13] M. Krotofil and A. A. Cárdenas, "Resilience of process control systems to cyber-physical attacks," in *Nordic Conference on Secure IT Systems*. Springer, 2013, pp. 166–182.

[14] P. Van Overschee and B. De Moor, *Subspace identification for linear systems: Theory–Implementation–Applications*. Springer Science & Business Media, 1996.

[15] E. Zugasti, A. G. González, J. Anduaga, M. A. Arregui, and F. Martínez, "Nullspace and autoregressive damage detection: a comparative study," *Smart Materials and Structures*, vol. 21, no. 8, p. 085010, 2012.

[16] J. J. Downs and E. F. Vogel, "A plant-wide industrial process control problem," *Computers & Chemical Engineering*, vol. 17, no. 3, pp. 245–255, 1993.

[17] T. McEvoy and S. Wolthusen, "A plant-wide industrial process control security problem," in *Critical Infrastructure Protection V*. Springer, 2011, pp. 47–56.

[18] A. A. Cárdenas, S. Amin, Z.-S. Lin, Y.-L. Huang, C.-Y. Huang, and S. Sastry, "Attacks against process control systems: risk assessment, detection, and response," in *Proceedings of the 6th ACM symposium on information, computer and communications security*. ACM, 2011, pp. 355–366.

[19] T. Larsson, K. Hestetun, E. Hovland, and S. Skogestad, "Self-Optimizing Control of a Large-Scale Plant: The Tennessee Eastman Process," *Industrial & Engineering Chemistry Research*, vol. 40, no. 22, pp. 4889–4901, 2001.

Un esquema de intercambio de clave a tres bandas postcuántico

Josep M. Miret
Univ. de Lleida
miret@matematica.udl.cat

Daniel Sadornil
Univ. de Cantabria
daniel.sadornil@unican.es

Juan G. Tena
Univ. de Valladolid
tena@agt.uva.es

Javier Valera
Univ. de Lleida
jvalera@matematica.udl.cat

***Abstract*—Hoy en día existen algoritmos cuánticos que resuelven los problemas de la factorización y el logaritmo discreto en tiempo polinómico. Este hecho ha motivado la búsqueda de nuevas propuestas criptográficas resistentes a posibles ataques cuánticos. En esta dirección, hace unos años, De Feo, Jao y Plût propusieron un nuevo esquema de intercambio de clave entre dos participantes que se cree que es *quantum-resistente*. En este trabajo extendemos el esquema anterior a un número cualquiera de participantes explicando con más detalle el caso en el que son tres.**

***Index Terms*—Curva elíptica · Supersingular · Isogenia · Intercambio de clave · Criptografía postcuántica.**

I. Introducción

En 1976, W. Diffie y M. E. Hellman [1] propusieron el primer esquema de intercambio de clave entre dos participantes, cuya seguridad se basa en un problema matemático computacionalmente difícil relacionado con el problema del logaritmo discreto (DLP). Concretamente, es el conocido *Computational Diffie-Hellman Problem* (CDHP):

Dados un generador g de un grupo multiplicativo cíclico G de orden n y dos elementos g^a, $g^b \in G$, con $1 < a, b < n$ escogidos aleatoriamente, calcular g^{ab}.

Más adelante, A. Joux [2] utilizó emparejamientos de curvas elípticas para diseñar un esquema de intercambio de clave entre tres participantes. En este caso, la seguridad se basa en el denominado *Bilinear Diffie-Hellman Problem* (BDHP):

Dados un generador P de un grupo aditivo cíclico G_1 de orden n, un emparejamiento e de $G_1 \times G_1$ a un grupo multiplicativo G_2 tal que $e(P,P) \neq 1$ y tres elementos aP, bP, $cP \in G_1$, con $1 < a, b, c < n$ escogidos aleatoriamente, calcular $e(P,P)^{abc}$.

Estos problemas requieren que el DLP en los grupos que intervienen sea computacionalmente intratable. No obstante, P. W. Shor [3] presentó en 1994 un algoritmo cuántico que resuelve en tiempo polinómico el DLP en un cuerpo finito primo. Más aún, dicho algoritmo puede generalizarse a cualquier cuerpo finito y también al grupo de puntos de una curva elíptica. Desde entonces, han ido apareciendo propuestas de esquemas criptográficos basadas en nuevos problemas que se creen resistentes a ataques cuánticos.

Recientemente, L. De Feo, D. Jao y J. Plût [4] presentaron un protocolo de intercambio de clave entre dos participantes mediante isogenias de curvas elípticas supersingulares. Su seguridad radica en la dificultad de encontrar una isogenia entre dos curvas elípticas supersingulares isógenas. Mientras que para las curvas elípticas ordinarias sí existe un algoritmo subexponencial para calcular una isogenia [5], para las supersingulares tal algoritmo no se conoce.

En este trabajo extendemos el protocolo de De Feo, Jao y Plût a tres participantes y, más en general, a n participantes.

Denotamos en lo sucesivo por $\mathbb{F}_q$ un cuerpo finito con q elementos y característica $p \geq 5$ y por $\overline{\mathbb{F}}_q$ su clausura algebraica.

II. Curvas elípticas

Una curva elíptica E definida sobre $\mathbb{F}_q$ es una curva sin puntos singulares que supondremos dada por una ecuación de Weierstrass de la forma

$$y^2 = x^3 + ax + b, \qquad a, b \in \mathbb{F}_q.$$

El j-invariante de E es

$$j(E) = \frac{-1728(4a)^3}{\Delta}$$

con $\Delta = -16(4a^3 + 27b^2)$.

El conjunto de puntos $(x, y) \in \mathbb{F}_q \times \mathbb{F}_q$ que satisfacen la ecuación de E junto con un punto especial $\mathcal{O}$, llamado punto del infinito, forman el conjunto $E(\mathbb{F}_q)$ de puntos $\mathbb{F}_q$-racionales de E. Este conjunto tiene estructura de grupo abeliano con una operación de suma cuyo elemento neutro es $\mathcal{O}$ (véase [6, capítulo 2] o [7, capítulo III]). Mediante el uso de esta operación es posible multiplicar un entero m por un punto $P \in E(\mathbb{F}_q)$:

$$mP = \begin{cases} \underbrace{P + \cdots + P}_{m \text{ veces}} & \text{si } m > 0, \\ \mathcal{O} & \text{si } m = 0, \\ \underbrace{(-P) + \cdots + (-P)}_{m \text{ veces}} & \text{si } m < 0, \end{cases}$$

siendo $-P$ el opuesto de P, es decir, $P + (-P) = \mathcal{O}$.

Si $\#E(\mathbb{F}_q)$ denota el cardinal de $E(\mathbb{F}_q)$, entonces

$$\#E(\mathbb{F}_q) = q + 1 - t$$

con $|t| \leq 2\sqrt{q}$. Si p divide a t, entonces E es supersingular. En caso contrario, E es ordinaria.

El subgrupo de m-torsión, $m > 0$, de $(E(\mathbb{F}_q), +)$ se define como

$$E[m](\mathbb{F}_q) = \{P \in E(\mathbb{F}_q) \mid mP = \mathcal{O}\}.$$

Este subgrupo, si no es trivial, o bien es cíclico o bien es de rango 2.

Sea E' una curva elíptica definida sobre $\mathbb{F}_q$ con coeficientes a' y b'. Las curvas elípticas E y E' son isomorfas sobre $\mathbb{F}_q$ si y sólo si

$$a = u^4 a' \qquad \text{y} \qquad b = u^6 b'$$

para algún $u \in \mathbb{F}_q - \{0\}$. Si lo son, entonces $j(E) = j(E')$.

Para más información véase [6] y [7].

III. Isogenias de curvas elípticas

Sean E y E' dos curvas elípticas definidas sobre $\mathbb{F}_q$. Una isogenia entre E y E' es una aplicación racional

$$\mathcal{I} : E \to E'$$

tal que $\mathcal{I}(\mathcal{O}) = \mathcal{O}'$. El núcleo de la isogenia $\mathcal{I}$ es

$$\ker \mathcal{I} = \{P \in E(\overline{\mathbb{F}}_q) \mid \mathcal{I}(P) = \mathcal{O}'\}.$$

La isogenia $\mathcal{I}$ es $\mathbb{F}_q$-racional si y sólo si $\mathcal{I}(E(\mathbb{F}_q)) \subseteq E'(\mathbb{F}_q)$. Las curvas elípticas E y E' son isógenas sobre $\mathbb{F}_q$ si y sólo si entre ambas existe una isogenia no constante $\mathbb{F}_q$-racional, o lo que es lo mismo, si y sólo si $\#E(\mathbb{F}_q) = \#E'(\mathbb{F}_q)$ (véase [8]).

Sea G un subgrupo finito de $(E(\overline{\mathbb{F}}_q), +)$. Entonces existe una isogenia

$$\mathcal{I}_G : E \to E/G$$

tal que $\ker \mathcal{I}_G = G$ (véase [7, proposición III.4.12]). El grado de $\mathcal{I}_G$ es igual al cardinal de su núcleo. Si los puntos de G son $\mathbb{F}_q$-racionales, entonces $\mathcal{I}_G$ también lo es. Los coeficientes de E/G y las ecuaciones de $\mathcal{I}_G$ pueden calcularse utilizando las fórmulas de Vélu [9].

IV. Esquema De Feo-Jao-Plût

En esta sección explicamos el esquema de intercambio de clave propuesto por De Feo, Jao y Plût [4].

Sean P_1 y P_2 los dos participantes que intervienen en el esquema. Lo primero que hacen P_1 y P_2 es buscar un número primo p de la forma $\ell_1^{e_1}\ell_2^{e_2} f \pm 1$, siendo ℓ_1 y ℓ_2 dos números primos distintos y pequeños que no dividen a f (por ejemplo, $\ell_1 = 2$ y $\ell_2 = 3$). Una vez han fijado p, lo siguiente que hacen es construir una curva elíptica supersingular E definida sobre $\mathbb{F}_{p^2}$ tal que

$$E[\ell_i^{e_i}](\mathbb{F}_{p^2}) \simeq \mathbb{Z}/\ell_i^{e_i}\mathbb{Z} \times \mathbb{Z}/\ell_i^{e_i}\mathbb{Z}, \qquad i = 1, 2.$$

Notemos que $E[\ell_i^{e_i}](\mathbb{F}_{p^2})$ contiene un total de $\ell_i^{e_i-1}(\ell_i + 1)$ subgrupos cíclicos de orden $\ell_i^{e_i}$ (cada uno de ellos es el núcleo de una isogenia de grado $\ell_i^{e_i}$). Después, para cada ℓ_i, calculan dos puntos $T_i^{(1)}$ y $T_i^{(2)}$ de orden $\ell_i^{e_i}$ tales que

$$E[\ell_i^{e_i}](\mathbb{F}_{p^2}) = \langle T_i^{(1)}, T_i^{(2)} \rangle.$$

Por último, cada participante P_i escoge aleatoriamente dos números $m_i^{(1)}$ y $m_i^{(2)}$ en el intervalo $[0, \ell_i^{e_i} - 1]$ de manera que ambos no sean simultaneamente divisibles por ℓ_i. A excepción de estos cuatro valores (dos para cada participante), todos los demás son públicos. Ahora que ya sabemos cuales son los parámetros iniciales, lo siguiente que vamos a explicar es el esquema en sí.

El esquema De Feo-Jao-Plût consta de dos rondas. En la primera de ellas, cada participante P_i calcula la isogenia

$$\mathcal{I}_{G_{i,1}} : E \to E/G_{i,1}$$

de núcleo

$$G_{i,1} = \langle m_i^{(1)} T_i^{(1)} + m_i^{(2)} T_i^{(2)} \rangle.$$

Notemos que $\mathcal{I}_{G_{i,1}}$ tiene grado $\ell_i^{e_i}$. Escribamos, por simplicidad, $\mathcal{I}_{i,1}$ y $E_{i,1}$ en lugar de $\mathcal{I}_{G_{i,1}}$ y $E/G_{i,1}$. Entonces el participante P_1 calcula los puntos

$$T_{2,1,1}^{(1)} = \mathcal{I}_{1,1}(T_2^{(1)}) \quad \text{y} \quad T_{2,1,1}^{(2)} = \mathcal{I}_{1,1}(T_2^{(2)})$$

y se los envía a P_2 junto con la curva elíptica $E_{1,1}$. El participante P_2, por su parte, calcula los puntos

$$T_{1,2,1}^{(1)} = \mathcal{I}_{2,1}(T_1^{(1)}) \quad \text{y} \quad T_{1,2,1}^{(2)} = \mathcal{I}_{2,1}(T_1^{(2)})$$

y se los envía a P_1 junto con la curva elíptica $E_{2,1}$. De esta manera finaliza la primera ronda.

En la segunda ronda, P_1 calcula $E_{2,1}/G_{1,2}$ con

$$G_{1,2} = \langle m_1^{(1)} T_{1,2,1}^{(1)} + m_1^{(2)} T_{1,2,1}^{(2)} \rangle$$

y P_2 calcula $E_{1,1}/G_{2,2}$ con

$$G_{2,2} = \langle m_2^{(1)} T_{2,1,1}^{(1)} + m_2^{(2)} T_{2,1,1}^{(2)} \rangle.$$

Entonces la clave secreta que comparten P_1 y P_2 es

$$j = j(E_{2,1}/G_{1,2}) = j(E_{1,1}/G_{2,2})$$

ya que $E_{2,1}/G_{1,2}$ y $E_{1,1}/G_{2,2}$ son isomorfas sobre $\mathbb{F}_q$.

Este esquema basa su seguridad en el siguiente problema:

Supersingular Computational Diffie-Hellman Problem (SSCDHP): *dadas las curvas elípticas supersingulares $E_{1,1}$ y $E_{2,1}$ y los puntos $\mathcal{I}_{1,1}(T_2^{(1)})$, $\mathcal{I}_{1,1}(T_2^{(2)})$, $\mathcal{I}_{2,1}(T_1^{(1)})$ e $\mathcal{I}_{2,1}(T_1^{(2)})$, encontrar el j-invariante j de la curva elíptica supersingular*

$$E/\langle m_1^{(1)} T_1^{(1)} + m_1^{(2)} T_1^{(2)}, m_2^{(1)} T_2^{(1)} + m_2^{(2)} T_2^{(2)} \rangle.$$

Actualmente, dicho problema es computacionalmente intratable, conjeturándose que no existen algoritmos cuánticos polinómicos para resolverlo.

V. Esquema a tres bandas

En esta sección presentamos un esquema de intercambio de clave a tres bandas. Tal esquema es una ampliación del propuesto por De Feo, Jao y Plût.

En dicho esquema, los tres participantes P_1, P_2 y P_3, en primer lugar, tienen que acordar una serie de parámetros públicos así como generar sus parámetros privados. El primer parámetro público que acuerdan es un primo

$$p = \ell_1^{e_1}\ell_2^{e_2}\ell_3^{e_3} f \pm 1,$$

donde ℓ_i, $i = 1, 2, 3$, son tres primos pequeños distintos tales que $\ell_i \nmid f$. El segundo es una curva elíptica supersingular E definida sobre $\mathbb{F}_{p^2}$ tal que cada $E[\ell_i^{e_i}](\mathbb{F}_{p^2})$ es isomorfo a $(\mathbb{Z}/\ell_i^{e_i}\mathbb{Z})^2$. Finalmente, para cada subgrupo de $\ell_i^{e_i}$-torsión de $(E(\mathbb{F}_{p^2}), +)$ acuerdan dos generadores $T_i^{(1)}$ y $T_i^{(2)}$. En cuanto a los parámetros privados, cada participante P_i genera de forma aleatoria dos números $m_i^{(1)}$ y $m_i^{(2)}$ en el intervalo $[0, \ell_i^{e_i} - 1]$ de manera que ambos no sean simultaneamente divisibles por ℓ_i.

Nuestro esquema consta de tres rondas. Lo que sigue es lo que hace cada participante en cada una de ellas.

– Ronda 1 –

Cada participante P_i realiza los siguientes dos pasos:

(1) Calcula la isogenia

$$\mathcal{I}_{G_{i,1}} : E \to E/G_{i,1}$$

de núcleo

$$G_{i,1} = \langle m_i^{(1)} T_i^{(1)} + m_i^{(2)} T_i^{(2)} \rangle.$$

(2) Realiza las siguientes asignaciones:

$$\mathcal{I}_{i,1} \leftarrow \mathcal{I}_{G_{i,1}}, \quad E_{i,1} \leftarrow E/G_{i,1}.$$

- Participante P_1:

  (3) Calcula los puntos

  $$T_{2,1,1}^{(1)} = \mathcal{I}_{1,1}(T_2^{(1)}), \quad T_{2,1,1}^{(2)} = \mathcal{I}_{1,1}(T_2^{(2)}),$$
  $$T_{3,1,1}^{(1)} = \mathcal{I}_{1,1}(T_3^{(1)}), \quad T_{3,1,1}^{(2)} = \mathcal{I}_{1,1}(T_3^{(2)}).$$

  (4) Envía a P_2 los puntos anteriores junto con $E_{1,1}$.

- Participante P_2:

  (3) Calcula los puntos

  $$T_{3,2,1}^{(1)} = \mathcal{I}_{2,1}(T_3^{(1)}), \quad T_{3,2,1}^{(2)} = \mathcal{I}_{2,1}(T_3^{(2)}),$$
  $$T_{1,2,1}^{(1)} = \mathcal{I}_{2,1}(T_1^{(1)}), \quad T_{1,2,1}^{(2)} = \mathcal{I}_{2,1}(T_1^{(2)}).$$

  (4) Envía a P_3 los puntos anteriores junto con $E_{2,1}$.

- Participante P_3:

  (3) Calcula los puntos

  $$T_{1,3,1}^{(1)} = \mathcal{I}_{3,1}(T_1^{(1)}), \quad T_{1,3,1}^{(2)} = \mathcal{I}_{3,1}(T_1^{(2)}),$$
  $$T_{2,3,1}^{(1)} = \mathcal{I}_{3,1}(T_2^{(1)}), \quad T_{2,3,1}^{(2)} = \mathcal{I}_{3,1}(T_2^{(2)}).$$

  (4) Envía a P_1 los puntos anteriores junto con $E_{3,1}$.

– Ronda 2 –

- Participante P_1:

  (1) Calcula la isogenia

  $$\mathcal{I}_{G_{1,2}} : E_{3,1} \to E_{3,1}/G_{1,2}$$

  de núcleo

  $$G_{1,2} = \langle m_1^{(1)} T_{1,3,1}^{(1)} + m_1^{(2)} T_{1,3,1}^{(2)} \rangle.$$

  (2) Realiza las siguientes asignaciones:

  $$\mathcal{I}_{1,2} \leftarrow \mathcal{I}_{G_{1,2}}, \quad E_{1,2} \leftarrow E_{3,1}/G_{1,2}.$$

  (3) Calcula los puntos

  $$T_{2,1,2}^{(1)} = \mathcal{I}_{1,2}(T_{2,3,1}^{(1)}) \quad \text{y} \quad T_{2,1,2}^{(2)} = \mathcal{I}_{1,2}(T_{2,3,1}^{(2)}).$$

  (4) Envía a P_2 los puntos anteriores junto con $E_{1,2}$.

- Participante P_2:

  (1) Calcula la isogenia

  $$\mathcal{I}_{G_{2,2}} : E_{1,1} \to E_{1,1}/G_{2,2}$$

  de núcleo

  $$G_{2,2} = \langle m_2^{(1)} T_{2,1,1}^{(1)} + m_2^{(2)} T_{2,1,1}^{(2)} \rangle.$$

  (2) Realiza las siguientes asignaciones:

  $$\mathcal{I}_{2,2} \leftarrow \mathcal{I}_{G_{2,2}}, \quad E_{2,2} \leftarrow E_{1,1}/G_{2,2}.$$

  (3) Calcula los puntos

  $$T_{3,2,2}^{(1)} = \mathcal{I}_{2,2}(T_{3,1,1}^{(1)}) \quad \text{y} \quad T_{3,2,2}^{(2)} = \mathcal{I}_{2,2}(T_{3,1,1}^{(2)}).$$

  (4) Envía a P_3 los puntos anteriores junto con $E_{2,2}$.

- Participante P_3:

  (1) Calcula la isogenia

  $$\mathcal{I}_{G_{3,2}} : E_{2,1} \to E_{2,1}/G_{3,2}$$

  de núcleo

  $$G_{3,2} = \langle m_3^{(1)} T_{3,2,1}^{(1)} + m_3^{(2)} T_{3,2,1}^{(2)} \rangle.$$

  (2) Realiza las siguientes asignaciones:

  $$\mathcal{I}_{3,2} \leftarrow \mathcal{I}_{G_{3,2}}, \quad E_{3,2} \leftarrow E_{2,1}/G_{3,2}.$$

  (3) Calcula los puntos

  $$T_{1,3,2}^{(1)} = \mathcal{I}_{3,2}(T_{1,2,1}^{(1)}) \quad \text{y} \quad T_{1,3,2}^{(2)} = \mathcal{I}_{3,2}(T_{1,2,1}^{(2)}).$$

  (4) Envía a P_1 los puntos anteriores junto con $E_{3,2}$.

– Ronda 3 –

- Participante P_1:

  (1) Calcula la curva elíptica $E_{3,2}/G_{1,3}$ con

  $$G_{1,3} = \langle m_1^{(1)} T_{1,3,2}^{(1)} + m_1^{(2)} T_{1,3,2}^{(2)} \rangle.$$

  (2) Calcula $j_1 = j(E_{3,2}/G_{1,3})$.

- Participante P_2:

  (1) Calcula la curva elíptica $E_{1,2}/G_{2,3}$ con

  $$G_{2,3} = \langle m_2^{(1)} T_{2,1,2}^{(1)} + m_2^{(2)} T_{2,1,2}^{(2)} \rangle.$$

  (2) Calcula $j_2 = j(E_{1,2}/G_{2,3})$.

- Participante P_3:

  (1) Calcula la curva elíptica $E_{2,2}/G_{3,3}$ con

  $$G_{3,3} = \langle m_3^{(1)} T_{3,2,2}^{(1)} + m_3^{(2)} T_{3,2,2}^{(2)} \rangle.$$

  (2) Calcula $j_3 = j(E_{2,2}/G_{3,3})$.

Al finalizar las tres rondas, la clave secreta que comparten P_1, P_2 y P_3 es

$$j = j_1 = j_2 = j_3.$$

La seguridad de este esquema radica en el siguiente problema:

Dadas las 6 curvas elípticas supersingulares y los 18 puntos que se envían los participantes en las tres rondas, encontrar el j-invariante j de la curva elíptica supersingular

$$E/\langle W_1, W_2, W_3 \rangle,$$

siendo $W_i = m_i^{(1)} T_i^{(1)} + m_i^{(2)} T_i^{(2)}$.

Para este problema, teniendo en cuenta que es una variante del SSCDHP, tampoco existe un algoritmo cuántico que lo resuelva en tiempo polinómico.

El esquema que acabamos de explicar puede ser generalizado a n participantes P_1, P_2, ..., P_n. La configuración inicial de tal esquema sería casi idéntica a las dos que ya hemos visto (p sería de la forma $\ell_1^{e_1}\ell_2^{e_2}\cdots\ell_n^{e_n}f\pm 1$, etc.). En dicho esquema se realizarían n rondas. En la ronda k, con $k\neq n$, cada participante P_i calcularía una curva elíptica, una isogenia $\mathcal{I}_{i,k}$ de grado $\ell_i^{e_i}$ y las imágenes por $\mathcal{I}_{i,k}$ de $2(n-k)$ puntos. Todos estos "objetos" serían intercambiados excepto las isogenias. En la ronda n, cada participante P_i calcularía una curva elíptica. En esta ronda no se intercambiaría nada. En total, por lo tanto, cada participante P_i calcularía n curvas elípticas, $n-1$ isogenias y $(n-1)n$ imágenes.

Sea $c=\text{máx}(e_1\ell_1,e_2\ell_2,\ldots,e_n\ell_n)=e\ell$ con ℓ primo. Supongamos, por simplicidad, que todas las isogenias $\mathcal{I}_{i,k}$ tienen grado ℓ^e. El coste computacional de calcular una isogenia $\mathbb{F}_q$-racional de grado ℓ vía las fórmulas de Vélu, así como el de calcular la imagen de un punto por dicha isogenia, es de $O(\ell)$ operaciones en $\mathbb{F}_q$. Teniendo en cuenta que una isogenia $\mathbb{F}_q$-racional $\mathcal{I}$ de grado ℓ^e puede descomponerse en e isogenias de grado ℓ, el coste computacional de calcular $\mathcal{I}$, así como el de aplicar $\mathcal{I}$ a un punto, es de $O(c)$ operaciones en $\mathbb{F}_q$. Suponiendo que el coste de calcular una curva elíptica vía las fórmulas de Vélu puede ser omitido, por todo lo explicado, el coste computacional del esquema con n participantes sería de $O((n-1)^2nc)$ operaciones en $\mathbb{F}_{p^2}$.

El aumento tanto del coste computacional como del número de comunicaciones necesarias en el esquema hace este poco práctico para n elevado. Cabe notar, sin embargo, que tal situación de n grande no parece previsible en un acuerdo de claves privadas.

VI. Ejemplo

Consideremos el primo

$$p=10^{20}+151=\ell_1^{e_1}\ell_2^{e_2}\ell_3^{e_3}f-1$$

con

$$(\ell_1,e_1)=(2,3),\quad (\ell_2,e_2)=(3,2),\quad (\ell_3,e_3)=(7,2)$$

y $f=28344671201814059$. Sea

$$\mathbb{F}_{p^2}=\mathbb{F}_p[\imath]=\frac{\mathbb{F}_p[x]}{(x^2+1)}$$

y sea E la curva elíptica definida sobre $\mathbb{F}_{p^2}$ de ecuación

$$y^2=x^3+x.$$

Denotemos por $x_i^{(k)}$ e $y_i^{(k)}$ la abscisa y la ordenada del punto $T_i^{(k)}$ de orden $\ell_i^{e_i}$. Entonces

$$\begin{aligned}
x_1^{(1)} &= 32708564414958782596,\\
y_1^{(1)} &= 41010880281127735051,\\
x_2^{(1)} &= 65941116997463769656,\\
y_2^{(1)} &= 30150112841634796621,\\
x_3^{(1)} &= 95326767949366434843,\\
y_3^{(1)} &= 83886855001252171605,
\end{aligned}$$

$$\begin{aligned}
x_1^{(2)} &= 67291435585041217555,\\
y_1^{(2)} &= 41010880281127735051\imath,\\
x_2^{(2)} &= 34058883002536230495,\\
y_2^{(2)} &= 30150112841634796621\imath,\\
x_3^{(2)} &= 4673232050633565308,\\
y_3^{(2)} &= 83886855001252171605\imath.
\end{aligned}$$

Sean

$$\begin{aligned}
(m_1^{(1)},m_1^{(2)}) &= (3,7),\\
(m_2^{(1)},m_2^{(2)}) &= (2,5),\\
(m_3^{(1)},m_3^{(2)}) &= (11,23).
\end{aligned}$$

Entonces, el j-invariante que comparten P_1, P_2 y P_3 es

$$j=89502672579038817856\imath+37665441899604681193.$$

Finalmente, explicamos cómo hemos obtenido los puntos $T_i^{(k)}$. Para obtener cada $T_i^{(1)}$ hemos seguido los siguientes pasos:

1. Escoger aleatoriamente un punto R_i de $E(\mathbb{F}_p)$.
2. Calcular el punto $S_i=((p+1)/\ell_i^{e_i})R_i$.
3. Comprobar si existe un entero k en el intervalo $[0,e_i-1]$ tal que $\ell_i^kS_i=\mathcal{O}$.
   a. Si existe, entonces volver al paso 1.
   b. Si no existe, entonces $T_i^{(1)}=S_i$.

Los puntos $T_i^{(2)}$ los hemos obtenido aplicando a cada $T_i^{(1)}$ la siguiente aplicación de distorsión (véase [10, Sección IX.13]):

$$\sigma(x,y)=(-x,\imath y).$$

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Economía, Industria y Competitividad mediante los proyectos MTM2014-55421-P y MTM2017-83271-R y por la Generalitat de Catalunya mediante el grupo 2017SGR 1158.

References

[1] W. Diffie y M. E. Hellman: "New Directions in Cryptography", en *IEEE Trans. Inf. Theor.*, vol. 22, n. 6, pp. 644–654, 1976.
[2] A. Joux: "A one round protocol for tripartite Diffie-Hellman", en *Journal of Cryptology*, vol. 17, n. 4, pp. 263–276, 2004.
[3] P. W. Shor: "Algorithms for quantum computation: Discrete logarithms and factoring", en *35th Annual Symposium on Foundations of Computer Science*, pp. 124–134, 1994.
[4] L. De Feo, D. Jao y J. Plût: "Towards quantum-resistant cryptosystems from supersingular elliptic curve isogenies", en *Journal of Mathematical Cryptology*, vol. 8, n. 3, pp. 209–247, 2014.
[5] S. Galbraith y A. Stolbunov: "Improved Algorithm for the Isogeny Problem for Ordinary Elliptic Curves", 2011.
[6] A. J. Menezes: *Elliptic curve public key cryptosystems*, Kluwer Academic Publishers, 1998.
[7] J. H. Silverman: *The arithmetic of elliptic curves*, Graduate Texts in Mathematics, vol. 106, Springer-Verlag, 1986.
[8] J. Tate: "Endomorphisms of Abelian Varieties over Finite Fields", en *Inventiones math.*, vol. 2, pp. 134–144, 1966.
[9] J. Vélu: "Isogénies entre courbes elliptiques", en *C. R. Acad. Sci. Paris Sér. A-B*, vol. 273, pp. A238–A241, 1971.
[10] I. F. Blake, G. Seroussi y N. P. Smart: *Advances in Elliptic Curve Cryptography*, London Mathematical Society Lecture Note Series, vol. 317, Cambridge University Press, 2005.

Sistema de Comunicaciones Seguras para Posicionamiento en Interiores

Iván Santos-González, Alexandra Rivero-García y Pino Caballero-Gil
Departamento de Ingeniería Informática y de Sistemas
Universidad de La Laguna
{jsantosg, ariverog, pcaballe}@ull.edu.es

***Resumen*—Este trabajo describe una solución alternativa para el problema de localización en interiores en lugares donde no es posible utilizar dispositivos GPS o estos no son lo suficientemente precisos. Esta nueva propuesta se basa en diferentes métodos de recolección de información que proporcionan información útil sobre la ubicación en este tipo de lugares. En particular, el uso de la tecnología de NFC (Near Field Communication) en combinación con una unidad de medición inercial a través de teléfonos móviles, permite resolver el problema de la localización en interiores sin incrementar los costes. Para solucionar dicho problema, se ha implementado una aplicación Android para demostrar la aplicabilidad de la solución propuesta. Además se ha añadido una capa de seguridad debido a la importancia de proteger la información de posicionamiento y evitar así la posibilidad de trazabilidad de los usuarios del sistema. Para ello, se ha seleccionado la curva elíptica de FourQ para generar una clave compartida utilizando el protocolo Diffie-Hellman para curvas curva elípticas. A continuación, la clave generada se utiliza para cifrar todas las comunicaciones mediante el algoritmo cifrado en flujo SNOW 3G. El sistema desarrollado ofrece resultados prometedores.**

***Index Terms*—Curvas Elípticas, Cifrado en Flujo, Posicionamiento en Interiores, Trazabilidad.**

I. Introducción

Uno de los lugares donde la ubicación no ofrece resultados aceptables hoy en día es en interiores. Es bien sabido que en exteriores el uso del sistema de posicionamiento GPS proporciona una localización precisa, pero su uso en interiores no es posible. La necesidad de un sistema de localización en interiores es esencial en grandes espacios como centros comerciales o aeropuertos. La forma tradicional de resolver este problema ha sido colocar mapas estáticos en diferentes puntos del edificio indicando dónde te encuentras dentro del plano del edificio. La principal desventaja de este tipo de puntos de información es que no son accesibles en todos los espacios del edificio.

Es sabido también que los teléfonos inteligentes son cada vez más esenciales en nuestra vida diaria porque no sólo utilizamos el smartphone para hacer llamadas telefónicas o enviar mensajes SMS (Short Message Service), sino también para realizar otras tareas como tomar imágenes, grabar vídeos, leer correos, ubicarnos con el GPS (Global Positioning System) o navegar por Internet. Debido a ello, en los últimos años se han propuesto diferentes soluciones de localización en interiores basadas en smartphones.

La propuesta que aquí se presenta se basa en el uso de dos tecnologías: la tecnología NFC (Near Field Communication) [1], que proporciona un posicionamiento de corto alcance, y una IMU (Inertial Measurement Unit) [2] instalada en el pie del usuario y que proporciona cambios inerciales para seguir el movimiento del usuario. El uso de estas tecnologías nos permite proporcionar una posición en tiempo real en un smartphone utilizando un mapa interior del edificio. Un aspecto que es muy importante en este tipo de sistemas y que no suele ser estudiado es su seguridad, y en particular, la imposibilidad de seguimiento de los usuarios. Para evitar esto, la propuesta presentada añade el uso del protocolo de intercambio de claves de Diffie-Hellman [3] utilizando la curva elíptica FourQ para generar una clave compartida y el algoritmo de cifrado en flujo SNOW 3G [4] para cifrar todas las comunicaciones entre la IMU y el smartphone.

Este trabajo se estructura de la siguiente manera. La Sección 2 describe algunos preliminares. El sistema propuesto se define en la Sección 3. La sección 4 introduce algunas características de la seguridad del sistema. Por último, algunas conclusiones y cuestiones pendientes cierran este documento.

II. Preliminares

Durante los últimos años se han presentado diferentes propuestas en el campo del uso de las IMUs para el seguimiento del movimiento y/o posición de los usuarios en diferentes situaciones. Existen diferentes tipos de IMU, pero tradicionalmente, las que se han usado para seguir el movimiento y/o la posición han sido las de 6 o 9DoF (Degrees of Freedom). Una IMU de 6 DoF usualmente tiene un acelerómetro de 3 DoF y un giróscopio de 3 DoF. El acelerómetro se utiliza para medir la aceleración de los movimientos de la IMU en los sistemas de coordenadas x, y, z, que pueden transformarse fácilmente en velocidad a través de la primera integral de la aceleración, y en posición relativa a través de la segunda integral de la aceleración. Por lo tanto, se puede utilizar para medir los cambios en la velocidad y la posición respectivamente. Un problema que suele aparecer al obtener velocidad y posición mediante el uso de la integral es que si no se elimina el error constante intrínseco de la medida original, la aceleración, este se convierte en un error lineal en la velocidad y en un error cuadrático en la posición, hecho que haría inutilizable el sistema. El giróscopio mide la orientación en los sistemas de coordenadas x, y, z. Un IMU 9 DoF tiene un acelerómetro de 3 DoF y un giroscopio de 3 DoF y añade un magnetómetro de 3 DoF. El magnetómetro es un sensor que mide el campo magnético y que normalmente se utiliza para obtener la orientación global debido al campo magnético de la tierra. Una guía completa de las fuentes de error más comunes en el uso de IMUs para sistemas de posicionamiento y sus efectos sobre el rendimiento de la navegación se puede encontrar en [5].

Este tipo de sistemas son ampliamente utilizados para el seguimiento de los movimientos en el espacio, por lo que en los últimos años se han presentado diferentes propuestas. Un método que se utiliza habitualmente y que se basa en las mediciones de los sensores mencionados anteriormente es el Dead-Reckoning. Este método consiste en el uso de diferentes algoritmos basados en ecuaciones trigonométricas fáciles para obtener la posición real de un objeto o persona, a través de operaciones basadas en el rumbo y la velocidad de navegación. Hay múltiples algoritmos que implementan Dead-Reckoning [6]. En dicho trabajo se presenta un estudio comparativo de diferentes algoritmos de Pedestrian Dead-Reckoning. Un algoritmo Pedestrian Dead-Reckoning es básicamente un algoritmo que calcula el movimiento de una persona mediante la detección de pasos, la estimación de la longitud de las zancadas y las direcciones de movimiento. Los resultados obtenidos en ese trabajo muestran cómo esta técnica ofrece resultados prometedores con una tasa media de errores en la estimación de la longitud de la zancada de alrededor del 1 % y una estimación por debajo del 5 % en la distancia total recorrida. Otro método que se aplica normalmente para mejorar el rendimiento y reducir el error de deriva en las mediciones de los sensores es el método basado en el uso de filtros estáticos y adaptativos, y uno de los filtros más utilizados es el filtro Kalman [7].

III. Sistema Propuesto

El sistema desarrollado consiste en una aplicación Android que muestra en un mapa interior la posición actual del usuario. Este plano interior del edificio debe ser proporcionado previamente por el personal de administración del edificio con su escala correcta para ser añadido al sistema. El sistema utiliza dos tecnologías diferentes para realizar esta función.

Por un lado, la tecnología NFC se utiliza en la entrada y en algunos puntos específicos de los lugares interiores para establecer una posición inicial del usuario en el mapa. Se eligió la tecnología NFC para este propósito porque es una tecnología de comunicación de corto alcance sin errores en la estimación de la posición inicial. La selección de esta tecnología en lugar de otra más barata como los códigos QR se debe a que la tecnología NFC es más fácil de proteger que los códigos QR como se muestra en la Sección IV.

Por otro lado, el uso de una IMU situada en el pie del usuario supone un punto de referencia estático. El uso de una IMU ubicada en el pie del usuario es una forma más precisa de recopilar datos que el smartphone, ya que es estática y produce menos ruido que el uso de los sensores del smartphone. La IMU se utiliza para recopilar datos sobre los sensores del acelerómetro, giroscopio y magnetómetro, que se envían al smartphone del usuario mediante el uso de la tecnología BLE (Bluetooth Low Energy) [8]. En el smartphone del usuario, los datos del sensor se procesan mediante el uso del algoritmo de filtrado Madgwick. Este algoritmo proporciona una orientación precisa del usuario en forma de cuaternión [9], que proporciona una orientación absoluta a partir de una relativa. A continuación, el cuaternión se utiliza para orientar la posición en el mapa de interiores.

Finalmente, se ha utilizado una estimación de la longitud de los pasos para realizar un estudio exhaustivo de los diferentes métodos. Como método inicial, decidimos utilizar una forma sencilla de calcular la longitud del paso en centímetros, l, que se puede mostrar en la ecuación (1), donde h representa la altura en centímetros del usuario y k es una constante que es de 0,415 para los hombres y 0,413 para las mujeres [10].

$$l = hk \tag{1}$$

En futuras versiones del sistema, se implementarán métodos de estimación de la longitud de los pasos más eficientes, precisos y complejos. Además, se llevará a cabo un estudio comparativo de la precisión de los diferentes métodos. En la figura 1 se pueden ver algunas capturas de pantalla del prototipo desarrollado.

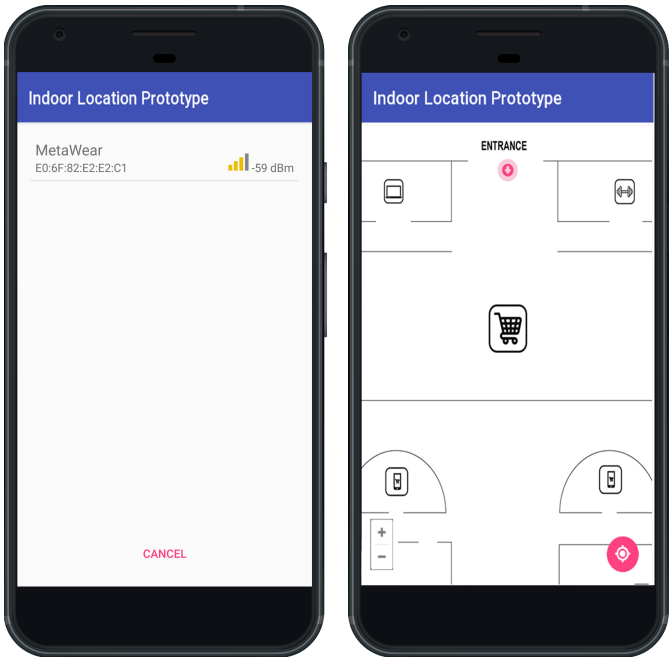


Figura 1. Prototipo de Aplicación

El rendimiento general del sistema puede mostrarse en la figura 2. Los pasos que un usuario del sistema realiza durante su uso son:

1. El paso inicial del sistema consiste en poner la altura del usuario la primera vez que utiliza la aplicación para calcular la longitud del paso.
2. El usuario escanea la etiqueta NFC situada en la entrada. La etiqueta NFC contiene algunos códigos de identificación que representan el edificio, la entrada y la planta donde se encuentra. La posibilidad de poner más etiquetas NFC alrededor del edificio está abierta para algunos casos en los que un usuario olvidó realizar el escaneo de la tarjeta NFC en la entrada. Esta información es importante para situar al usuario en el lugar correcto dentro del edificio y de la planta del mismo.
3. Una vez que la etiqueta NFC ha sido leída, el usuario puede ver su posición inicial sobre el mapa de planta.
4. Desde este momento, la IMU comienza a recopilar datos y a enviarlos al smartphone del usuario, que es el responsable de operar con ellos. El algoritmo Madgwick se utiliza para obtener la orientación en tiempo real. Con el cuadrante obtenido por algoritmo y la longitud

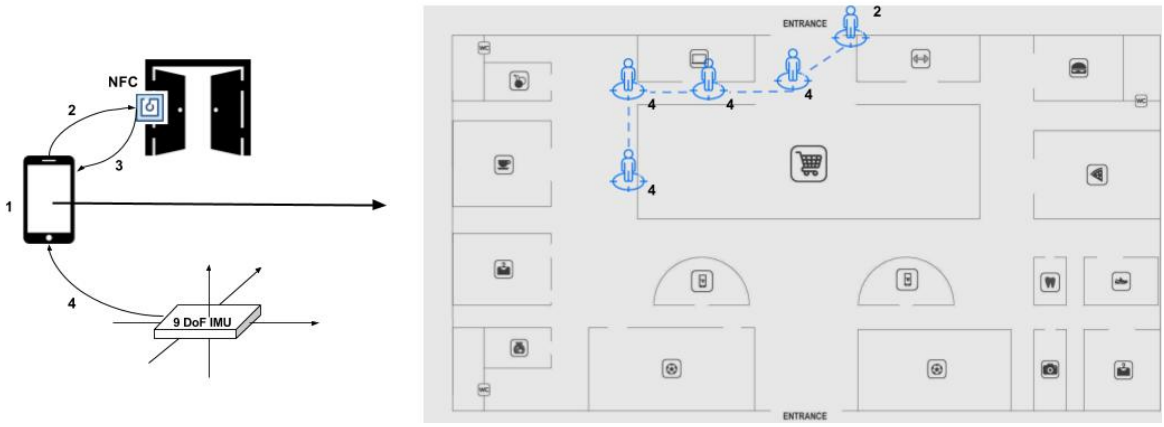


Figura 2. Esquema General del Sistema

del paso, la posición del usuario se muestra en el smartphone en tiempo real.

III-A. Sistema de Posicionamiento

La parte principal del sistema de posicionamiento interior es la parte relacionada con los datos recogidos por IMU y su tratamiento. En nuestro sistema de posicionamiento se ha utilizado una CPRO IMU Metawear que recoge las medidas de los 3 sensores, acelerómetro, giroscopio y magnetómetro, obteniendo unidades de g, $grados/segundos$ y $Tesla$ respectivamente. Las especificaciones completas de la unidad IMU se pueden ver en [11]. Esta unidad IMU transmite los datos a través de BLE. Los datos recogidos se transmiten en tiempo real al smartphone donde se realiza el tratamiento de las diferentes variables, incluyendo la conversión de las unidades del acelerómetro, g, a m/s^2, y las unidades del giroscopio, $grados/segundos$, a rad/s. Luego, se aplica el filtro Madgwick para obtener el cuaternión que representa la deriva, inclinación y alabeo. Con estos datos, la detección del paso y la longitud del paso, la posición del usuario se muestra sobre el mapa cada vez que da un paso.

Se han realizado diferentes estudios sobre el uso de filtros en los datos obtenidos de las IMU para mejorar la calidad y reducir el ruido en los datos [12] [13] [14], mostrando que el filtro Madgwick es el más apropiado en este tipo de sistemas. En este trabajo, como trabajo complementario, decidimos implementar pruebas de tres de los filtros más utilizados, un filtro Kalman, un filtro Mahony y el mencionado filtro Madgwick. La representación de la deriva, inclinación y alabeo obtenidos en las pruebas se muestran en las figuras 3, 4 y 5, respectivamente. En todas las pruebas, el filtro Kalman se representa en color verde, el filtro Mahony en azul y el Madgwick en naranja. Durante las diferentes pruebas, el filtro Magdwick muestra una mejor precisión comparando la posición real con la posición mostrada en la aplicación Android.

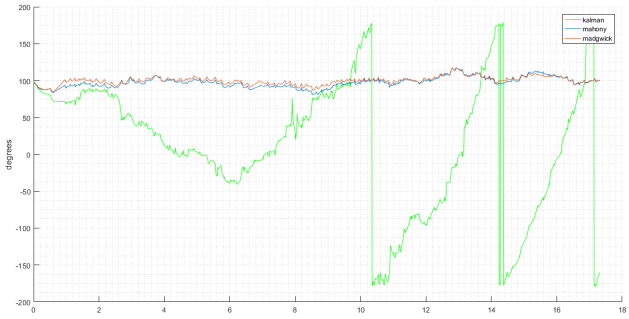


Figura 3. Filtros de Deriva

Figura 4. Filtros de Inclinación

Figura 5. Filtros de Alabeo

IV. Seguridad del Sistema

En los sistemas de localización, la seguridad es un aspecto muy importante. Una aplicación vulnerable puede implicar un problema de privacidad y, en particular, de trazabilidad. La trazabilidad de un usuario puede suponer un gran problema para los usuarios de una aplicación, ya que un atacante puede

saber en todo momento dónde se encuentra un usuario y así realizar ataques derivados. Por ejemplo, si el atacante sabe que no estamos en casa, podría robar en la casa, o vender nuestros datos a diferentes compañías para permitir que nos envíen, por ejemplo, publicidad de comida si el usuario está en la cocina, o publicidad de champú si el usuario está en el baño, etc.

Para proteger el sistema, decidimos utilizar dos métodos diferentes. Por un lado, utilizamos el algoritmo de cifrado en flujo SNOW 3G para cifrar todas las comunicaciones y el protocolo ECDH (Elliptic Curve Diffie-Hellman) utilizando la curva elíptica FourQ para generar una clave compartida. Por otro lado, decidimos cambiar la clave secreta cada vez que se reinicia la aplicación para usarla como un protocolo de sesión. El uso de estos dos métodos, ECDH y SNOW 3G, se ha utilizado comúnmente en diferentes documentos y su seguridad está ampliamente probada [15] [16].

IV-A. FourQ ECDH

El uso de curvas elípticas en criptografía ha sido ampliamente discutido durante los últimos años y las ventajas que tienen con respecto a la criptografía tradicional tanto en longitud de clave como en requerimientos computacionales son bien conocidas. La curva FourQ es una nueva curva elíptica desarrollada por Microsoft Research [17], que cumple con los requisitos del NIST para la selección de curvas elípticas de nueva generación. Estos requisitos son que las nuevas curvas deben, al menos, mantener el nivel de seguridad de las anteriores y ser altamente eficientes en sus implementaciones software y hardware. Esta curva produce resultados prometedores, como demuestran diferentes estudios presentados por Microsoft Research, y ofrece mejoras en los tiempos de cómputo de las pruebas realizadas en ordenadores tradicionales. Para saber si las mejoras mostradas en las pruebas de Microsoft Research son posibles también en dispositivos portátiles, donde la arquitectura del procesador es totalmente diferente, ya que los ordenadores suelen utilizar una arquitectura x32 o x64 mientras que los smartphones y dispositivos portátiles suelen utilizar la arquitectura ARM, decidimos portar la implementación realizada por Microsoft Research al lenguaje Java para utilizarla en dispositivos Android. Para ello se realizó una librería java y se comparó el tiempo de cómputo ejecutando el protocolo ECDH entre FourQ, la curva NIST P-256 [18], y la curva Curve25519 [19]. Los resultados de esta comparación se pueden ver en la Tabla I.

Tabla I
COMPARACIÓN DE TIEMPOS DE EJECUCIÓN EN PROTOCOLO ECDH

Curva	Tiempo
Curve25519	721 ms
Curve NIST P-256	1876 ms
Curve FourQ	417 ms

Los tiempos de cómputo mostrados en la Tabla I muestran que la curva elíptica de FourQ también ofrece mejoras interesantes en dispositivos portátiles. En particular, esta curva es 2 veces más rápida que la curva de nueva generación Curve25519 y entre de 4-5 veces más rápida que la curva NIST P-256. El uso de esta curva puede ser un avance importante en la seguridad de la IoT debido a una menor longitud de las claves y a una mayor eficiencia, hechos que son especialmente importantes en este tipo de dispositivos con bajas capacidades de cómputo y almacenamiento.

IV-B. SNOW 3G

SNOW 3G es un algoritmo de cifrado de flujo designado en 2006 como base para la protección de la integridad y el cifrado de la tecnología UMTS. Gracias a que el algoritmo satisface todos los requisitos impuestos por el 3GPP (3rd Generation Partnership Project) con respecto a los recursos de tiempo y memoria, fue seleccionado como base de los algoritmos UEA2 (UMTS Encryption Algorithm 2) y UIA2 (UMTS Integrity Algorithm 2) [20] [21].

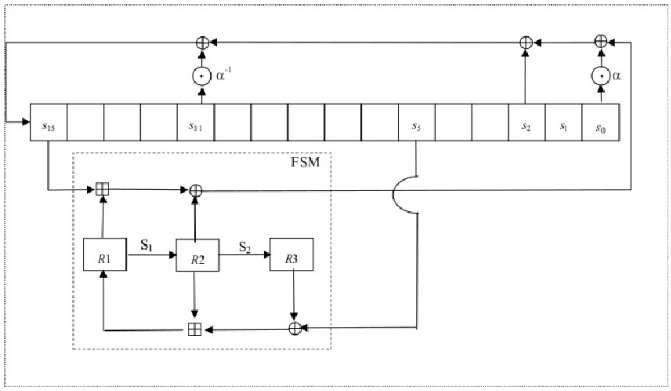


Figura 6. Esquema General de SNOW 3G

El algoritmo SNOW 3G deriva del algoritmo SNOW 2 y utiliza claves de 128 bits y un vector de inicialización para generar en cada iteración 32 bits de secuencia cifrante. Por un lado, el LFSR utilizado en este algoritmo tiene 16 etapas denominadas s0, s1, s2... s15 con 32 bits cada una. Por otro lado, el FSM (Finite State Machine) utilizado se basa en tres registros de 32 bits denominados R1, R2 y R3 y utiliza dos cajas de sustitución llamadas S1 y S2. La operación de combinación utiliza un XOR y un módulo de adición 2^{32}, como podemos ver en la figura 6.

SNOW 3G tiene dos modos de ejecución: el modo de inicialización y el modo de generación de secuencia cifrante. Primero, el modo de inicialización se ejecuta sin producir ninguna secuencia cifrante. A continuación, se ejecuta el modo de generación de secuencia cifrante. En general, el número de iteraciones que se deben ejecutar el algoritmo depende del número de palabras de 32 bits que queramos cifrar.

V. CONCLUSIONES

En este trabajo se presenta un nuevo sistema de localización y posicionamiento interior que ofrece resultados prometedores. La combinación de diferentes tecnologías nos permite obtener la localización en interiores con un alto nivel de precisión. El uso de una IMU de bajo coste hace que el sistema pueda ser utilizado por muchas personas en un futuro próximo. Durante las simulaciones se ha desarrollado un prototipo de aplicación Android para recoger la información

del IMU, proceder a los diferentes cálculos y mostrar la ruta sobre el mapa de interiores. La seguridad de este tipo de sistemas es esencial, por lo que se han implementado diferentes protocolos y algoritmos de seguridad para evitar la posible trazabilidad del usuario por parte de un atacante malicioso. Se trata de un trabajo en progreso, por lo que todavía hay varias líneas abiertas. La primera de ellas es el estudio de otros algoritmos de fusión de sensores que podrían encajar mejor en el sistema desarrollado. Otra mejora podría ser el uso de los sensores del smartphone en lugar de la IMU para realizar el posicionamiento. Por otro lado, también son necesarias pruebas de seguridad y ataques controlados para mejorar la seguridad del sistema.

AGRADECIMIENTOS

Investigación parcialmente financiada con TESIS2015010102, TESIS-2015010106, el Ministerio Español de Economía y Competitividad , Fondos Europeos FEDER, y la Fundación Cajacanarias, bajo los Proyectos TEC2014-54110-R, RTC-2014-1648-8, MTM2015-69138-REDT y DIG02-INSITU.

REFERENCIAS

[1] V. Coskun, K. Ok, and B. Ozdenizci, *Near field communication (NFC): From theory to practice*. John Wiley & Sons, 2011.

[2] M. M. Morrison, "Inertial measurement unit," Dec. 8 1987, uS Patent 4,711,125.

[3] N. Koblitz, "Elliptic curve cryptosystems," *Mathematics of computation*, vol. 48, no. 177, pp. 203–209, 1987.

[4] A. Kircanski and A. M. Youssef, "On the sliding property of snow 3g and snow 2.0," *IET Information Security*, vol. 5, no. 4, pp. 199–206, 2011.

[5] W. Flenniken, J. Wall, and D. Bevly, "Characterization of various imu error sources and the effect on navigation performance," in *Ion Gnss*, 2005, pp. 967–978.

[6] A. R. Jimenez, F. Seco, C. Prieto, and J. Guevara, "A comparison of pedestrian dead-reckoning algorithms using a low-cost mems imu," in *Intelligent Signal Processing, 2009. WISP 2009. IEEE International Symposium on*. IEEE, 2009, pp. 37–42.

[7] R. Van Der Merwe and E. A. Wan, "The square-root unscented kalman filter for state and parameter-estimation," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 6. IEEE, 2001, pp. 3461–3464.

[8] C. Gomez, J. Oller, and J. Paradells, "Overview and evaluation of bluetooth low energy: An emerging low-power wireless technology," *Sensors*, vol. 12, no. 9, pp. 11 734–11 753, 2012.

[9] B. K. Horn, "Closed-form solution of absolute orientation using unit quaternions," *JOSA A*, vol. 4, no. 4, pp. 629–642, 1987.

[10] I. Bylemans, M. Weyn, and M. Klepal, "Mobile phone-based displacement estimation for opportunistic localisation systems," in *Mobile Ubiquitous Computing, Systems, Services and Technologies, 2009. UBICOMM'09. Third International Conference on*. IEEE, 2009, pp. 113–118.

[11] MambientLab, "Metawear specifications," https://mbientlab.com, accessed: 08/01/2017.

[12] S. O. Madgwick, A. J. Harrison, and R. Vaidyanathan, "Estimation of imu and marg orientation using a gradient descent algorithm," in *Rehabilitation Robotics (ICORR), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–7.

[13] S. Madgwick, "An efficient orientation filter for inertial and inertial/magnetic sensor arrays," *Report x-io and University of Bristol (UK)*, vol. 25, pp. 113–118, 2010.

[14] F. Alam, Z. ZhaiHe, and H. Jia, "A comparative analysis of orientation estimation filters using mems based imu," in *Proceedings of the International Conference on Research in Science, Engineering and Technology, Dubai, UAE*, 2014, pp. 21–22.

[15] S. Kumar, M. Girimondo, A. Weimerskirch, C. Paar, A. Patel, and A. S. Wander, "Embedded end-to-end wireless security with ecdh key exchange," in *Circuits and Systems, 2003 IEEE 46th Midwest Symposium on*, vol. 2. IEEE, 2003, pp. 786–789.

[16] J. Molina-Gil, P. Caballero-Gil, C. Caballero-Gil, and A. Fúster-Sabater, "Analysis and implementation of the snow 3g generator used in 4g/lte systems," in *International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*. Springer, 2014, pp. 499–508.

[17] Z. Liu, P. Longa, G. Pereira, O. Reparaz, and H. Seo, "Fourq on embedded devices with strong countermeasures against side-channel attacks," Cryptology ePrint Archive, Report 2017/434, 2017. 28, 29, Tech. Rep.

[18] M. Brown, D. Hankerson, J. López, and A. Menezes, "Software implementation of the nist elliptic curves over prime fields," *Topics in Cryptology—CT-RSA 2001*, pp. 250–265, 2001.

[19] D. J. Bernstein, "Curve25519: new diffie-hellman speed records," in *International Workshop on Public Key Cryptography*. Springer, 2006, pp. 207–228.

[20] P. Kitsos, G. Selimis, and O. Koufopavlou, "High performance asic implementation of the snow 3g stream cipher," *IFIP/IEEE VLSI-SOC*, pp. 13–15, 2008.

[21] G. Orhanou, S. El Hajji, and Y. Bentaleb, "Snow 3g stream cipher operation and complexity study," *Contemporary Engineering Sciences-Hikari Ltd*, vol. 3, no. 3, pp. 97–111, 2010.

Mejora de la seguridad en la autentificación basada en contraseñas mediante un cifrador en bloque

Rafael Álvarez Sánchez
Departamento de
Ciencia de la Computación e
Inteligencia Artificial
Universidad de Alicante
Email: ralvarez@dccia.ua.es

Alicia Andrade Bazurto
Facultad de Ingeniería,
Ciencias Físicas
y Matemática
Universidad Central del Ecuador
Email: aandrade@uce.edu.ec

Antonio Zamora Gómez
Departamento de
Ciencia de la Computación e
Inteligencia Artificial
Universidad de Alicante
Email: zamora@dccia.ua.es

***Resumen*—Las funciones hash de contraseña son funciones hash con propiedades criptográficas especializadas, dotadas de parámetros ajustables que permiten frustrar los ataques por fuerza bruta o de diccionario sobre las bases de datos que almacenan el resumen de las contraseñas de los usuarios, incluso cuando los atacantes emplean unidades de procesamiento gráfico de propósito general o hardware especializado.**

En este documento, se propone el diseño de una función hash de contraseña haciendo uso de un cifrador en bloque en modo contador y se analizan las características de rendimiento, seguridad y complejidad para el estándar de cifrado en bloque, *Advanced Encryption Standard*. Los resultados se comparan con los obtenidos al utilizar la conocida función hash de contraseña *Scrypt*, obteniendo resultados prometedores dado que se explota la ventaja que tiene AES de ser acelerado por hardware en la mayoría de los procesadores modernos.

***Index Terms*—Contraseña, PBKDF, Autenticación, AES, Hash**

I. Introducción

En la mayoría de sistemas y servicios de Internet, el principal medio de autentificación de usuario está basado en el uso de contraseñas, aun cuando existen otras técnicas aparentemente más avanzadas como las basadas en biometría. La gestión de estas contraseñas no es un tema menor y, generalmente, se procesan mediante una función hash criptográfica junto con una cadena aleatoria llamada *sal*.

El uso de unidades de procesamiento gráfico (GPUs) e, incluso, hardware específico ha tenido un gran impacto, permitiendo ataques por fuerza bruta a las bases de datos que contienen los hashes de las contraseñas. Por ello, ha sido necesario utilizar funciones de hash de contraseña especiales, con parámetros de espacio y tiempo ajustables, para hacer más costosos los ataques por fuerza bruta (véase [7]). Las funciones de derivación de claves basadas en contraseña (*password-based key derivation functions*, PBKDF) es un campo de investigación muy activo, existiendo una gran cantidad de publicaciones recientes relacionadas con estas funciones de hash de contraseña (véanse [1], [2], [3], [5], [9], [10], [11], [12], [13]) que mejoran el conocido estándar PBKDF2 (véase [8]).

El estándar de cifrado en bloque *Advanced Encryption Standard* (AES, véase [4]) es más lento de forma nativa que algunos cifradores en flujo actuales pero tiene la ventaja de ser acelerado por hardware en muchos procesadores modernos, lo que hace que sea extremadamente rápido, con velocidades superiores a 1 GB/s en sistemas actuales. Es una excelente primitiva básica sobre la que construir muchas herramientas criptográficas, incorporando en su diseño la velocidad y seguridad que aporta este estándar.

En este trabajo se describe una PBKDF (o función de hash de contraseñas), basada en el empleo de AES en modo contador (CTR) como un generador de números pseudoaleatorios (o *pseudo-random number generator*, PRNG). También analizamos su rendimiento y seguridad, y lo comparamos en términos de rendimiento en una cantidad equivalente de uso de memoria con la conocida PBKDF Scrypt (véase [9]).

El hecho de que AES esté acelerado por hardware en muchas plataformas comunes permite protegerse frente ataques basados en GPU o en hardware a medida, equilibrando el coste computacional entre defensor y atacante cuando este último tiene como objetivo recuperar las contraseñas de los usuarios. La función de hash de contraseñas propuesta puede ser especialmente útil en aplicaciones donde se requiera autentificación segura y sea importante tener un rendimiento configurable, como en dispositivos móviles, computación embebida, etc.

II. Descripción

A continuación, detallamos los parámetros de entrada a la función de hash de contraseña propuesta, los elementos esenciales que conforman el núcleo de la función, así como el pseudocódigo detallado para las etapas de inicialización y salida del algoritmo.

II-A. Parámetros

- **pass[]**, *array* de bytes de longitud arbitraria que contiene la contraseña (*password*) del usuario.
- **salt[]**, *array* de bytes de longitud arbitraria que contiene la sal aleatoria. Se recomienda un mínimo de 32 bytes.
- **plen**, entero positivo que indica la longitud (*length*) en bytes del hash de salida. Se recomienda un mínimo de 32 bytes.
- **pmem**, entero positivo que indica el número de entradas en la tabla principal. Este parámetro modula el coste espacial (memoria). Se requiere un valor mínimo de 1.
- **ptime**, entero positivo que indica el número de pasadas a través de la tabla principal. Este parámetro modula el coste temporal (*time*). El valor mínimo posible es 1.

II-B. Elementos

- **M[]**, *array* de pmem × plen bytes de longitud. Constituye el bloque de memoria principal del algoritmo.
- **out[]**, *array* de plen bytes que contiene el hash de salida.
- **fH()**, una función hash criptográfica segura con un tamaño de resumen de 256 bits utilizado para la semilla. En nuestro caso, empleamos el estándar SHA3-256.
- **fC()**, un cifrador en bloque seguro, en modo CTR. En nuestro caso, empleamos el estándar AES-128 en modo CTR. Es la primitiva criptográfica básica en el diseño de la función propuesta y se utiliza principalmente como un PRNG.

II-C. Inicialización

En el lenguaje de pseudocódigo empleado (basado en Go, véase [6]), se asume que se puede llamar a *len()* en *arrays* y obtener su longitud, evitando tener que pasar longitudes como argumentos separados.

La llamada a la función devuelve un *array* de bytes con el hash de salida y se define de la siguiente manera:

```
func Key(pass,salt []byte,plen,pmem,ptime int)
```

Para evitar colisiones en las que la concatenación directa de la contraseña y la sal produzcan como resultado los mismos datos de entrada, la semilla inicial se obtiene como el hash de la contraseña y la sal separadas por un solo byte con valor 0. De ese hash de 256 bits, los primeros 16 bytes de la semilla se usan como la clave de 128 bits para AES y los segundos 16 bytes de la semilla conforman el valor de inicialización de 128 bits para el modo CTR.

A continuación, los búferes out[] y M[] se cifran utilizando la instancia de AES inicializada. Se supone que estos almacenamientos intermedios son puestos a cero por make() antes del cifrado.

A continuación, se muestra el pseudocódigo para la etapa de inicialización:

```
fH := sha3.New256()
fH.Write(pass)
fH.Write(0)
fH.Write(salt)
seed := fH.Sum(nil)

blk := aes.NewCipher(seed[0:16])
fC := cipher.NewCTR(blk,seed[16:32])

out := make([]byte,plen)
fC.XORKeyStream(out,out)

M := make([]byte,pmem*plen)
fC.XORKeyStream(M,M)
```

II-D. Salida

El algoritmo de salida es sencillo ya que, para cada entrada en M[], se obtiene un índice pseudoaleatorio de los primeros 8 bytes (64 bits) de out[] que, a su vez, se emplea para elegir una entrada de M[] a combinar con el estado actual de out[]. Esta operación consiste en una resta de bytes (Z_{256}). Finalmente, el resultado de esta operación se cifra a out[].

El ciclo externo es el número de veces (o pasadas por M[]) que se realiza todo este proceso, siendo la salida de la función el último estado de out[]. Dado que out[] es la salida directa de un cifrado AES, toda la función hash de contraseña debe ser tan segura como AES.

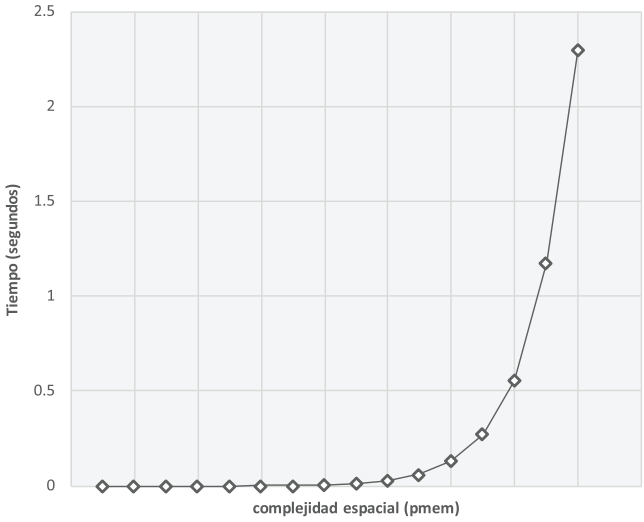


Figura 1. Rendimiento en función de la complejidad espacial.

A continuación, se muestra el pseudocódigo para la etapa de salida:

```
for t := 0; t < ptime; t++ {
  for m := 0; m < pmem; m++ {
    i := (int(out[0:8])%pmem) * plen
    for o := 0; o < len(out); o++ {
      M[i+o] -= out[o]
    }
fC.XORKeyStream(out, M[i:i+plen])
  }
}
```

III. Análisis

III-A. Rendimiento

En esta sección se analiza el rendimiento respecto al incremento de la complejidad espacial (parámetro pmem) y temporal (parámetro ptime), así como una comparación con Scrypt. Las pruebas experimentales se han implementado en el lenguaje de programación Go (versión 1.8.1) y se han llevado a cabo en una computadora con Intel Core i5 a 2,6 GHz (que admite aceleración de hardware AES) y 16 GB de RAM; tanto las contraseñas y sales empleadas han sido de 32 bytes (256 bits) de longitud, al igual que la salida, que también ha sido de 32 bytes; cada prueba se midió diez veces, tomando el mínimo como el valor final.

III-A1. Complejidad espacial: en la figura 1, se muestra el coste computacional de la función propuesta al aumentar el parámetro pmem exponencialmente de 2^8 a 2^{23} entradas de 32 bytes (lo que equivale a un uso de memoria desde 8KiB a 256MiB) y con una sola pasada (*ptime* $= 1$). Se observa que se obtienen valores muy razonables de alrededor de $0,5$ segundos o menos con pmem hasta 2^{21}. También es destacable que existe un margen significativo disponible, en caso de desear una mayor complejidad computacional.

III-A2. Complejidad temporal: el rendimiento del parámetro ptime se muestra en la figura 2. En este caso, ptime aumenta exponencialmente de 1 a 2^{15} pasadas y pmem se fija en 256 entradas de 32 bytes (que equivale a 8KiB de

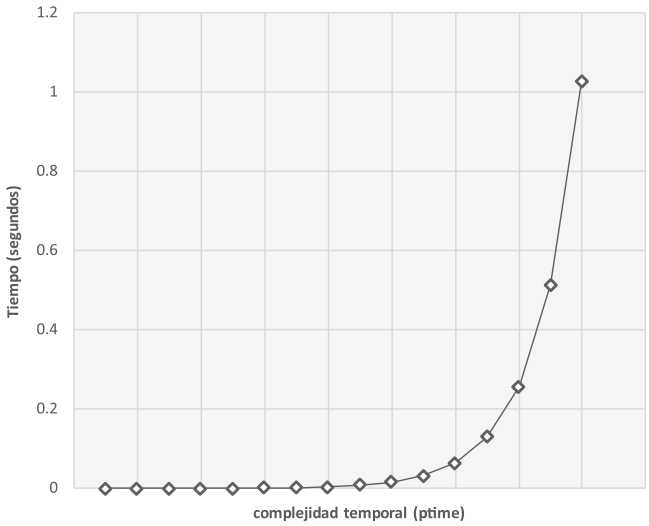


Figura 2. Rendimiento en función de la complejidad temporal.

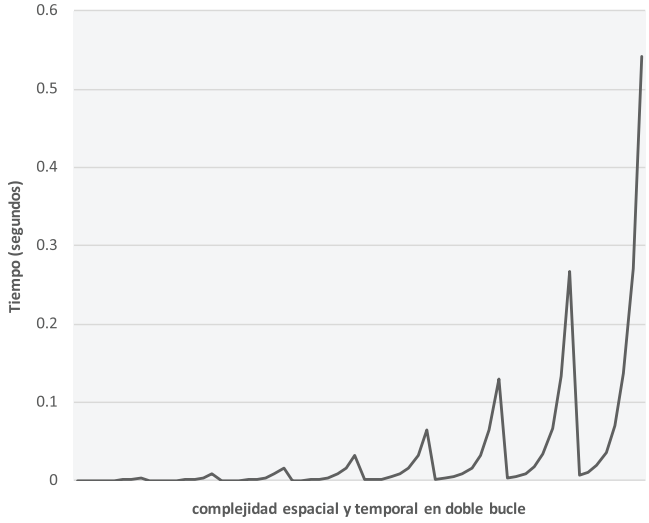


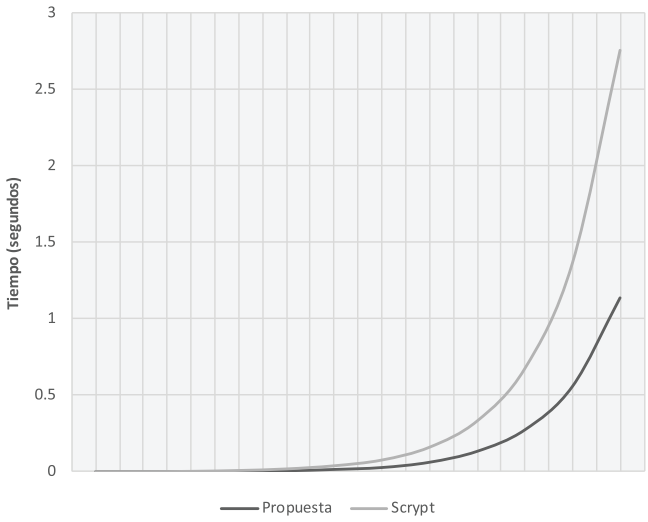
Figura 3. Rendimiento en función de ambas complejidades combinadas.

memoria). Se obtienen tiempos de ejecución razonables de alrededor de $0,5$ segundos o menos con hasta 2^{14} pasadas.

III-A3. Complejidad combinada: la figura 3 muestra el tiempo de ejecución asociado a la modulación simultánea en un doble bucle de los parámetros pmem y ptime. El número de entradas en M[] (bucle externo) va de 2^8 a 2^{15} y el número de pasadas (bucle interno) va de 1 a 2^7.

La serie esperada de curvas exponenciales crecientes se observa claramente, mostrando un buen rango de rendimiento en la modulación de ambos parámetros de complejidad. El rendimiento también es muy razonable dentro de estos parámetros, con todos los tiempos alrededor de $0,5$ segundos o menos. La cantidad máxima de memoria procesada es de 128MiB con pmem=2^{15} y ptime=2^7.

III-A4. Comparación con Scrypt: una de las funciones de hash de contraseñas más utilizadas en los servicios actuales de autentificación mediante contraseña es la PBKDF Scrypt (véase [9]), que ha sido diseñada para defenderse de los ataques mediante GPU y permite utilizar una cantidad



pmem	Propuesta (segundos)	Scrypt (segundos)
2^8	0.000391	0.000774
2^9	0.000780	0.001513
2^{10}	0.001562	0.003001
2^{11}	0.003431	0.006577
2^{12}	0.007240	0.013007
2^{13}	0.016000	0.024674
2^{14}	0.028405	0.049614
2^{15}	0.062817	0.098736
2^{16}	0.134064	0.198680
2^{17}	0.274017	0.400811
2^{18}	0.558009	0.806953
2^{11}	1.137899	1.620456

Figura 4. Comparativa con Scrypt.

configurable de memoria. Se puede observar en la figura 4 que la función propuesta en este trabajo (basada en AES-CTR) es significativamente más rápida que Scrypt al procesar cantidades equivalentes de memoria. Para obtener los resultados mostrados en la gráfica, se ha hecho uso de la librería estándar de Go que implementa AES (que aprovecha la aceleración de hardware) y la de Scrypt (que no está acelerada por hardware y utiliza la versión de ronda reducida de Salsa20/8 en su núcleo). El resultado obtenido pone de manifiesto que AES es una buena opción para diseñar funciones de hash de contraseñas en la mayoría de las máquinas modernas; por el contrario, en plataformas donde AES no esté acelerado por hardware, el rendimiento será considerablemente menos competitivo.

III-B. Seguridad

El algoritmo propuesto está diseñado, fundamentalmente, como función de hash de contraseñas con un objetivo principal: la autentificación de usuarios mediante contraseña; al tiempo, proporciona algunas contramedidas contra ataques que hacen uso de unidades de procesamiento gráfico rápidas o de hardware especializado. Las principales trabas que interpone contra este tipo de ataques son el uso de memoria (evitando, así, el paralelismo) y el tiempo requerido para obtener el hash, explotando el hecho de que AES se implementa en hardware en la mayoría de los procesadores modernos (véanse [3], [5], [10], [12], [13]); estas trabas no

las ofrecen las funciones hash criptográficas comunes, que se ejecutan en estas plataformas especializadas mucho más rápido (varios órdenes de magnitud), permitiendo ataques por fuerza bruta satisfactorios.

Otra aplicación que se le puede dar es convertir una contraseña de longitud variable en una clave de longitud variable, con excelentes propiedades de avalancha o también como un esquema de prueba de trabajo en los protocolos con cadenas de bloques (*blockchain*) usados en las criptomonedas.

En la función propuesta, la primitiva criptográfica principal es AES-128 en modo CTR, lo que permite una alta seguridad y rendimiento en la mayoría de sistemas modernos. La clave y el valor de inicialización para AES-CTR se obtienen haciendo uso del hash SHA3-256 de la contraseña y la sal proporcionadas. Tanto AES como SHA3 son primitivas que se consideran seguras y estándares actuales ampliamente empleados; el diseño propuesto debería estar acorde con estas primitivas en términos de seguridad, proporcionando un nivel de seguridad mínimo de 128 bits para prevenir ataques por fuerza bruta. En el caso de que SHA3-256 o AES-128 se rompieran, podrían utilizarse otras primitivas de seguridad equivalentes en su lugar.

IV. Conclusiones

Se ha presentado una función de hash de contraseñas basada en Advanced Encryption Standard en modo contador que aprovecha el soporte de hardware acelerado en la mayoría de sistemas modernos para prevenir ataques por fuerza bruta para la recuperación de contraseñas que emplean hardware especializado o unidades de procesamiento gráfico de propósito general.

Se ha analizado el rendimiento de la función propuesta en varios escenarios de memoria y tiempo y se ha realizado un estudio comparativo con un estándar de facto dentro de las funciones de hash de contraseñas como Scrypt, obteniéndose resultados prometedores y demostrando que AES es un candidato adecuado para este tipo de primitivas criptográficas. La seguridad general de la función diseñada es equivalente a la de AES, ya que se utiliza como una primitiva central en el diseño y la salida final es el resultado directo del cifrado AES.

Con respecto a futuras investigaciones, podría ser muy interesante realizar una implementación en hardware especializado para probar de forma adecuada las características de rendimiento en estas plataformas. Además, se podría considerar en el contexto de la función propuesta otras medidas de seguridad adicionales como el uso de un archivo aleatorio grande en el servidor (*server-side ROM*), poder actualizar la seguridad sin intervención del usuario final (*client-independent update*) o delegar parte del trabajo al usuario final (*server relief*), así como la evaluación comparativa del rendimiento con otras propuestas recientes.

Agradecimientos

Investigación financiada parcialmente por el Gobierno de España mediante el proyecto TEC2014-54110-R (CASUS).

Referencias

[1] Alvarez, R., Zamora, A.: Using Spritz as a Password-Based Key Derivation Function, en *Advances in Intelligent Systems and Computing*, vol. 527, pp. 518–525, 2017.

[2] Bertoni, G., Daemen, J., Peeters, M., Van Assche, G.: Cryptographic sponge functions. http://sponge.noekeon.org/, 2011.

[3] Biryukov, A., Dinu, D., Khovratovich, D.: .Argon2: the memory-hard function for password hashing and other applications", *Password Hashing Competition winner*, 2016. https://github.com/P-H-C/phc-winner-argon2/blob/master/argon2-specs.pdf

[4] Daemen, J., Rijmen, V.: AES proposal: Rijndael. In First Advanced Encryption Standard (AES) Conference (1999)

[5] Forler, C., Lucks, S., Wenzel, J.: The Catena Password-Scrambling Framework. Version 3.2, Bauhaus-Universität Weimar. (2015) https://www.uni-weimar.de/fileadmin/user/fak/medien/professuren/Mediensicherheit/Research/Publications/catena-v3.2.pdf

[6] The Go Programming Language. http://www.golang.org

[7] Hellman, M. E.: A cryptanalytic time-memory trade-off. Information Theory, IEEE Transactions on, vol. 26–4, 401–406. (1980)

[8] Kaliski, B.: PKCS #5: Password-Based Cryptography Specification Version 2.0. Internet Engineering Task Force, Network Working Group, Request for Comments (RFC) 2898 (2000) https://tools.ietf.org/html/rfc2898#section-5.2

[9] Percival, C.: Stronger key derivation via sequential memory-hard functions. BSDCan – The BSD Conference. (2009) http://www.bsdcan.org/2009/schedule/attachments/87_scrypt.pdf

[10] Pornin, T.: The MAKWA Password Hashing Function. Version 1.1. Password Hashing Competition finalist. (2015) http://www.bolet.org/makwa/makwa-spec-20150422.pdf

[11] Provos, N., Mazieres, D.: A Future-Adaptable Password Scheme. USENIX Annual Technical Conference, FREENIX track, 81–91. (1999)

[12] Simplicio, M. A., Almeida, L. C., Andrade, E. R., dos Santos, P. C. F., Barreto, P. S. L. M.: Lyra2: Password Hashing Scheme with improved security against time-memory trade-offs. IACR Cryptology ePrint Archive, 2015:136. (2015)

[13] Solar Designer: yescrypt – password hashing scalable beyond bcrypt and scrypt. Presented at PHDays'14. Openwall (2014) http://www.openwall.com/presentations/PHDays2014-Yescrypt/PHDays2014-Yescrypt.pdf

Analysis of the SegWit adoption in Bitcoin

C. Pérez-Solà, S. Delgado-Segura, J. Herrera-Joancomartí, G. Navarro-Arribas
Dep. of Information and Communications Engineering,
Universitat Autònoma de Barcelona

***Abstract*—Segregated witnesses (SegWit) is the most controversial modification performed in the Bitcoin protocol so far. Although it was a soft fork modification with backward compatibility, some miners reluctance to accept the improvement let finally to a fork in the Bitcoin blockchain that produced the new Bitcoin Cash cryptocurrency. In this paper, we describe the segregated witness upgrade providing information about its motivation and improvements in terms of security (malleability) and scalability (payment channels). Furthermore, we provide some data on its real adoption on the Bitcoin ecosystem.**

***Index Terms*—Bitcoin, cryptocurrency, segregated witness**

I. Introduction

Recently, the increase of both the popularity and the use of Bitcoin has shown its bounds regarding the ability of the system to scale with the number of users. It is obvious that a system with a unique (although replicated) register containing all system transactions (i.e. the blockchain) may present a bottleneck.

There are several ways of measuring scalability, as pointed out in [1]. From latency (the time for a transaction to be confirmed) to bootstrap time (the time it takes a new node to download and process the history necessary to validate payments) through cost per confirmed transactions, different measures can be used to evaluate the efficiency of a payment system. Croman et al. [1] gave approximations of all of these metrics for the Bitcoin network. However, probably the easiest metric to compare Bitcoin with existing global payment systems and the one that has a direct impact on scalability is the system transaction throughput. The transaction throughput can be measured by the maximum number of transactions per second that a system can deal with. Throughput is often chosen to evaluate systems because it is objective, easy to compute, and gives a metric to easily compare different payment systems. For instance, Visa reported to allow around $2{,}000$ transactions per second in normal situation [1] while reaching a peak of $56{,}000$ in a stress test [2]. Paypal manages lower values, providing 136 transactions per second on average on his payment network.[1]

Regarding Bitcoin, its throughput can be measured considering different parameters, from network communication latency to processing power of the nodes performing transaction validation. However, what limits the throughput of Bitcoin the most is the block size limit, which is currently fixed at 1MB.[2]

Limiting the block size to 1MB implies a maximum throughput of 7 transactions per second [1]. Such limit is an approximation obtained by dividing the maximum block size by the average size of Bitcoin transactions (250 bytes) and the inter-block time (10 minutes). Therefore, a block may contain, at most, $1{,}000{,}000/250 = 100$ average sized transactions, thus giving a throughput of $100/600 = 6.6$ transactions per second. Notice that such value is far behind the throughput that other payment systems, like Visa or PayPal, can deal with.

The rest of the paper is organized as follows. Section II reviews the solutions that were proposed to solve the Bitcoin scalability problem. Then, Section III presents SegWit, the solution that was finally deployed. After that, Section IV analyses the adoption of SegWit since its deployment up until today. Finally, Section V presents the conclusions.

II. Bitcoin scaling proposals

An increase on the adoption and use of the system made Bitcoin face a state in which blocks became full. Full blocks implied a longer confirmation time, since the network produced more transactions than blocks were able to handle. The lack of space in blocks also affected transaction fees, making them increase, and therefore making the currency less usable. In order to deal with the scaling problem several solutions where proposed towards dealing with the main bottleneck, the block size limit.

Some proposals suggested to increase the limit following different strategies or even propose to remove the limit. Jeff Garzik's BIP 100 [4] proposed to change the 1MB fixed limit to a new floating block size limit, where miners may increase the block size by consensus. Gavin Andresen BIP 101 [5] proposal (currently withdrawn) consisted in initially increasing block size to 8MB and doubling the size every two years for 20 years, after which the block size remains fixed. Jeff Garzik's BIP 102 [6] proposed to simply increase block size to 2MB. Pieter Wuille's BIP 103 [7] proposed to increase the maximum block size by 4.4% every 97 days until 2063, implying a 17.7% block size increase per year. Gavin Andresen's BIP 109 [8] proposed a fixed block size increase to 2MB with a new limit on the amount of data that can be hashed to compute signature hashes and a change on how the maximum number of signatures is counted. However, increasing the block size implies a change on the consensus rules of the currency, that have to be deployed via a hard fork of the protocol.

A. Protocol forks

Modification proposals in the Bitcoin protocol, even those of utter importance like the ones affecting the scalability of the system, are often difficult to tackle since they have to be deployed with extreme caution and maximum consensus. Furthermore, if changes affect the consensus mechanisms of the protocol, their implications may cause a blockchain fork and that could have a big impact in the cryptocurrency. Moreover, the collateral implications of changes need to be also

[1] PayPal Q1 2016 Results [3] reported handling 1.41B payment transactions, which leads to an estimated $1.41B/4/30/24/60/60 = 136$ transactions per second.

[2] https://github.com/bitcoin/bitcoin/blob/a6a860796a44a2805a58391a009ba22752f64e32/src/consensus/consensus.h#L9

considered beforehand to prevent unexpected consequences, specially those related to security and decentralization.

Changes in the Bitcoin consensus rules may be introduced by soft (protocol) forks or hard (protocol) forks. A soft fork is produced when the protocol rules are changed but they remain backward-compatible with the old rules, making the old nodes able to accept blocks produced by new nodes. On the contrary, hard forks make a change in the protocol that changes its behavior in some sense, and make old nodes not able to follow the protocol anymore. When a soft fork occurs, only the majority of miners have to upgrade to the new version, since old nodes can accept newly generated blocks. However, when a hard fork occurs, all nodes that want to follow the new protocol rules are forced to upgrade to the new version.

As already pointed out, increasing the block size will imply a hard fork. However, it is unclear whether an increase in the block size by its own will solve future scalability problems. In some sense arbitrarily increasing the block size is like making hard drives twice as big to deal with storage problems, instead of designing smarter hard drives able to store more information within the same physical space.

B. Off-chain Payment Channels

An alternative solution to deal with the Bitcoin scaling problem are off-chain payment channels, that intend to move most of the transactions performed in the system outside of the blockchain, reducing the number of transactions to be included in blocks and therefore indirectly increasing the throughput.

The first proposal of such a mechanism targeted unidirectional payments (payments from a payer to a payee). Its main goal was to avoid the fees that transactions in the blockchain imply and that are not affordable for micropayment transactions [9]. To set up the payment channel, a transaction is included in the blockchain as a deposit of the money that will be used in the payment channel. A refund transaction is also created, allowing the payer to recover the deposited funds if the payee does not cooperate. The refund transaction can not be included in the blockchain until a certain point in the future, and thus the channel may remain open until that moment is reached. Between the set up and the closing of the payment channel, the payer can perform multiple payments to the payee through transactions that, although build following the Bitcoin transaction format, would be transferred privately between A and B without using the Bitcoin network. Furthermore, the individual payment transactions will not appear in the blockchain: only the set up transaction that opens the channel and the last transaction that closes the channel will be broadcast through the Bitcoin P2P network and therefore included in a block.

A more recent proposal, which allows bidirectional payments between parties, was proposed by Poon and Dryja [10]: the Lightning network[3].

However, both the basic off-chain protocol and the Lightning network faced an issue with the Bitcoin version at the time, the transaction malleability problem.

[3]The Lightning network is an extensive topic to talk about, and will not fit within the paper page limit. Interested readers should refer to the original paper.

C. Transaction Malleability Problem

Bitcoin transactions are identified by their transaction id, a value computed using a double SHA256 function over the raw data that defines the transaction. Each identifier should uniquely identify a transaction, and each transaction should have a unique identifier. However, due to how Bitcoin transactions are formatted the previous statement does not always hold. Some modifications can be performed over a signed transaction that will keep the origin and destination of payment while completely modifying the transaction id. Therefore, a transaction id is not final until the transaction can not be modified anymore, that is, once is has been included in a block. In scenarios like off-chain payment channels, were some transactions are created referring to transactions that are not final, transaction malleability poses a risk in the system. It is important to mention that malleability does not affect the ability of an attacker to spend/steal the bitcoins from an already created transaction, since he can not forge the digital signature of the owner. For that reason, although transaction malleability is known back from 2011, it has never been considered as a security issue.

III. Segregated Witness

Segregated Witness (SegWit) [11] was proposed as a solution for the malleability problem, by redesigning the transaction structure to compute transaction identifiers without counting the signatures. Such modification was proposed as a soft fork of the protocol, that in addition, virtually increased the block size up to four times the limit at the time (4 MB). Moreover, by deploying SegWit, off-chain payment channels could build on top of Bitcoin, increasing the throughput virtually to no limit.

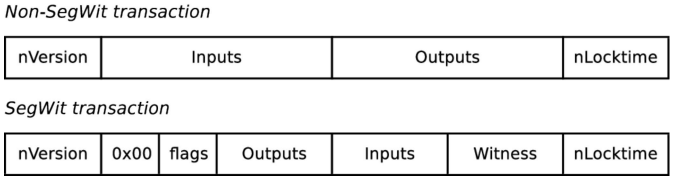


Fig. 1. Simplified non-SegWit and SegWit transaction.

The main idea behind SegWit is to detach from the transaction the information needed to validate its correctness while keeping the information related to its effects. This information comprises scripts and signatures, which are then placed in a new structure called *witness* (see Figure 1). The transaction signature includes only the fixed information making the id unique. Figure 2 show the (simplified) fields used to compute the transaction id (*tx_id*) in SegWit and non-SegWit transactions.

The witness has to be included in the block where the transaction belongs. To do that, an `OP_RETURN` output of the generation transaction of the block includes the root of a Merkle tree of the witnesses of the transactions.

As we will see in the next section, there are several types SegWit scripts. Most notably SegWit can be used *natively* or encapsulated in P2SH outputs (also called *nested* SegWit). We thus have encapsulated Pay-to-Witness-Public-Key-Hash (P2SH-P2WPKH), native Pay-to-Witness-Public-

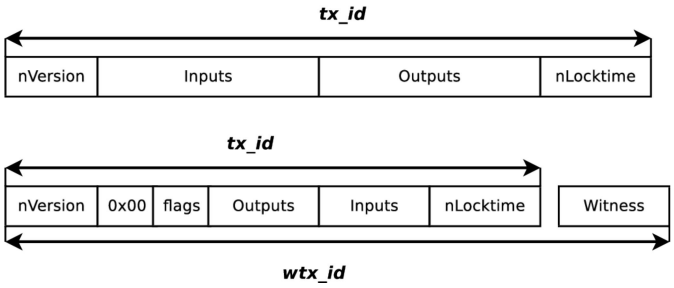


Fig. 2. Simplified SegWit transaction id vs. non-SegWit transaction id.

Key-Hash (P2WPKH), or native Pay-to-Witness-Script-Hash (P2WSH).

As a result of the use of SegWit, the size of a transaction is reduced, increasing the number of transaction that fit in a 1MB block. Fees for transaction are also reduced and computed based on the transaction *virtual size*.

SegWit also introduces a script versioning system, which allows to expand the scripting language through soft-forks.

It is true that in overall, SegWit introduces some complexity to Bitcoin, but this complexity was required in order to maintain backward compatibility and allow its deployment as a soft-work.

IV. Analysis

In this section, an analysis of the usage of segregated witness in the Bitcoin blockchain is presented. The analysis is made using BlockSci [12], an open-source software for blockchain analysis.

Segregated Witness was activated on the main Bitcoin network at block height 481,824 (on the 24th of August, 2017). Previously, it was locked-in at height 479,707, thus providing almost a two-week grace period for clients and wallets to upgrade to the new rules. Once activated, the new consensus rules started to be enforced, allowing to create transactions whose signature data was placed outside the old transaction format, with all the consequences on scalability and malleability that this implied.

As explained in the previous section, SegWit scripts may be deployed either with their native definition or encapsulated within standard BIP16 P2SH outputs. Native SegWit outputs may be identified directly by looking at their output scripts, because they follow a fixed structure. On the contrary, SegWit scripts encapsulated in P2SH outputs can not be identified until the output has been spent (and thus the redeem script has been revealed). This is because a P2SH output only contains the hash of the redeem script and, therefore, the actual redeem script is not known until the output is spent. Consequently, our analysis of SegWit usage is focused in studying, on one hand, redeem scripts found in P2SH inputs (for nested SegWit) and, on the other hand, output scripts (for native SegWit). Finally, we study the effects of deploying SegWit (both native and nested scripts) in the scalability of the system, both directly by effectively allowing bigger block sizes and indirectly by solving the malleability problem.

A. P2SH nested SegWit scripts

The firsts SegWit transactions in the Bitcoin main network correspond to P2WPKH nested within P2SH inputs and are found in block 481,824. This very same block has five transactions with this kind of inputs.[4] Three of these transactions include references to the activation of segregated witness: two of them have `OP_RETURN` outputs with messages saying hello to SegWit and a third one spends an output with a vanity address starting with `35SegWitPieWKVH...`. The first P2SH input with a P2WSH redeem script is found a little latter, in block 481,831.[5] Figure 3 shows the (2000-block moving average) percentage of inputs in each block that correspond to P2SH scripts with either P2WPKH or P2WSH redeem scripts. P2WPKH shows an still increasing tendency (accounting for almost 8% of today's inputs). Contrarily, P2WSH shows an increase during the first months (up until block 490,000) but seems to stabilize from that block onward. Altogether, nested SegWit scripts account for up to 11% of the current Bitcoin inputs in blocks.

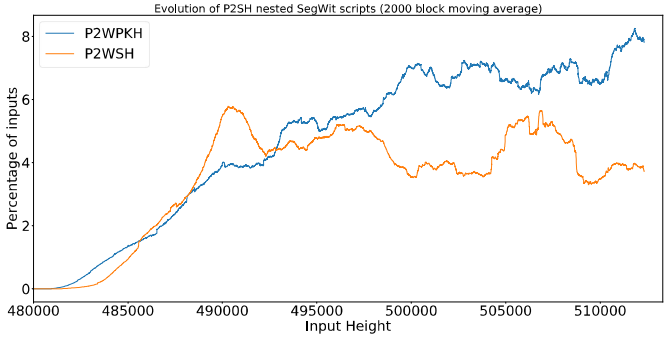


Fig. 3. Evolution of the number of P2SH nested SegWit inputs.

The average number of blocks between a P2SH SegWit input and the output it spends is 1,032 blocks, with small differences between nested P2WPKH and P2WSH inputs.

B. Native SegWit scripts

The first native SegWit outputs can be also found in block 481,824, that contains one transaction with a P2WSH[6] and two with P2WPKH outputs[7]. One of these two transactions is also spending a P2WPKH nested in P2SH input, and the other contains an `OP_RETURN` output announcing that a Japanese exchange supports SegWit. Figure 4 shows the (2000-block moving average) percentage of outputs in each block that correspond to native SegWit scripts, either P2WPKH or P2WSH witness programs. In contrast with encapsulated SegWit scripts, the adoption of native SegWit scripts is still very small, with an average of less than 0.5% of the outputs using these kinds of scripts. However, its adoption is still increasing. Also in contrast with nested SegWit scripts, native P2WSH outputs are seen more often than native P2WPKH outputs.

[4]Transactions with ids:
8f907925d2ebe48765103e6845c06f1f2bb77c6adc1cc002865865eb5cfd5c1c
b6e7c5365351ec7f8d29725afcdace8d76cf63a83bc9b58b9ff7cba61670b2be
c586389e5e4b3acb9d6c8be1c19ae8ab2795397633176f5a6442a261bbdefc3a
d09e2a5edbb6a0ac390a52a1b5292d88667f5445eb8e507441737a7bdd7157ee
dfcec48bb8491856c353306ab5febeb7e99e4d783eedf3de98f3ee0812b92bad

[5] fb5e59e65fa92719b3e24f17f20e5a2e19cf0a3643403d512d0dd806a84d510d

[6] 461e8a4aa0a0e75c06602c505bd7aa06e7116ba5cd98fd6e046e8cbeb00379d6

[7] dfcec48bb8491856c353306ab5febeb7e99e4d783eedf3de98f3ee0812b92bad
f91d0a8a78462bc59398f2c5d7a84fcff491c26ba54c4833478b202796c8aafd

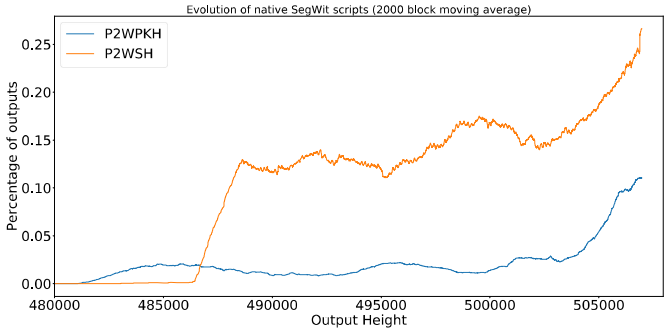


Fig. 4. Evolution of the number of native SegWit outputs.

Regarding the percentage of native SegWit scripts that were spent when this analysis was performed (block height 507,952), native P2WSH and P2WPKH present notable differences: 96.92% of P2WSH outputs have been already spent, whereas this percentage goes down to 74.21% for P2WPKH outputs.

Taking into account already spent (native SegWit) outputs only, the average number of blocks between an output and the input that spends it is 310 (the average for P2WPKH outputs is 529, whereas for P2WSH outputs is 287).

C. The impact of SegWit in the scalability of Bitcoin

SegWit inputs allow to move part or the totality (nested or native SegWit inputs, respectively) of their signature data to the witness structure. Since witness data is discounted when accounting for the total block size, SegWit allows to effectively create bigger block sizes.

Figure 5 shows the evolution of block sizes in Bitcoin. The orange dots denote base block size, that is, the size of the block without witness data. Base block size is limited to 1MB. The evolution of base block size over time shows how some blocks started to became full around height 325,000. Blue dots denote total block sizes, that is, block sizes including witness data. Since segregated witness was activated, many blocks total sizes have surpassed the 1MB limit, with block 505,253 setting a maximum of 2,217,342 bytes. However, all these blocks have a virtual size still less than 1MB, allowing them to be valid with the existing consensus rules. For instance, block 505,253 has a virtual size of 998,206 bytes. Currently (20th of April, 2018), up to 15,121 blocks have total sizes bigger than 1MB.

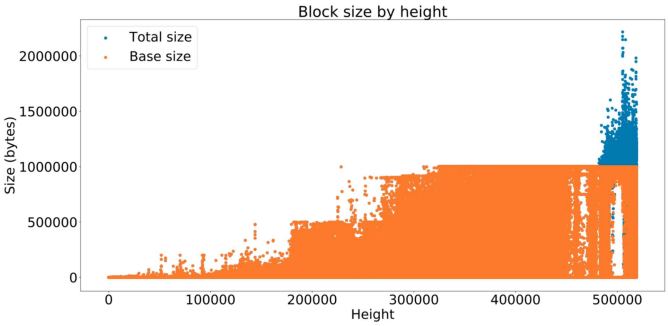


Fig. 5. Evolution of the size of blocks in Bitcoin.

Apart from the increase in block sizes, SegWit also solved the malleability problem, and thus allowed the deployment of payment channel networks such as the Lightning Network. Currently, the Lightning Network has almost 2,000 nodes and more than 5,000 channels[8].

V. Conclusions

After a very controversial scaling debate, Segregated Witness was deployed in the Bitcoin main network as a solution to deal with the scaling problems of Bitcoin while collateraly offering many other features (such as solving the malleability problem or introducing a new way to make upgrades to script).

As we have seen in this paper, since its activation SegWit has actively been used, mainly in its nested form where it is encapsulated within a classical P2SH script. The usage of native SegWit outputs is increasing, but it is still very residual. In any case, the usage of SegWit transactions has allowed, on one hand, to surpass the 1MB block limit and, on the other hand, to effectively deploy second layer solutions such as the Lightning Network.

Acknowledgements

This work is partially supported by the Spanish ministry under grant number TIN2014-55243-P.

References

[1] K. Croman, C. Decker, I. Eyal, A. E. Gencer, A. Juels, A. Kosba, A. Miller, P. Saxena, E. Shi, and E. Gün, "On scaling decentralized blockchains," in *Proc. 3rd Workshop on Bitcoin and Blockchain Research*, 2016.

[2] Visa, "56,582 transaction messages per second!" http://visatechmatters.tumblr.com/post/108952718025/56582-transaction-messages-per-second, July 2014, last accessed: June 2016.

[3] Paypal, "Paypal q1 2016 fast facts," June 2016, last accessed: June 2016.

[4] J. Garzik, "BIP 100: Making decentralized economic policy," 2015, last accessed: June 2016.

[5] G. Andresen, "BIP 101: Increase maximum block size," 2015, last accessed: June 2016.

[6] J. Garzik, "BIP 102: Block size increase to 2mb," 2015, last accessed: June 2016.

[7] P. Wuille, "BIP 103: Block size following technological growth," 2015, last accessed: June 2016.

[8] G. Andresen, "BIP 109: Two million byte size limit with sigop and sighash limits," 2016, last accessed: June 2016.

[9] M. Hearn and J. Spilman, "Bitcoin contracts," https://en.bitcoin.it/wiki/Contract, last accessed: June 2016.

[10] J. Poon and T. Dryja, "The bitcoin lightning network: Scalable off-chain instant payments," Technical Report (draft). https://lightning. network, Tech. Rep., 2015.

[11] E. Lombrozo, J. Lau, and P. Wuille, "BIP 141: Segregated witness (consensus layer)," 2015, last accessed: June 2016.

[12] H. Kalodner, S. Goldfeder, A. Chator, M. Möser, and A. Narayanan, "Blocksci: Design and applications of a blockchain analysis platform," *arXiv preprint arXiv:1709.02489*, 2017.

[8]There are many graphical explorers of the Lightning Network. A list can be found in https://gist.github.com/bretton/798ec38165ffabc719d91e0f4f67552d

Reconstrucción de la Clave PILAR utilizada por el Gobierno Civil de Málaga en 1940

Alberto Peinado
Universidad de Málaga, Andalucía Tech
E.T.S.I. Telecomunicación, Depto. Ingeniería de Comunicaciones,
Campus de Teatinos 29071 Málaga
apeinado@ic.uma.es

Abstract- **Este documento presenta la reconstrucción del cifrado de cinta móvil denominado "Clave Pilar" utilizado en 1940 en las comunicaciones entre la Dirección General de Seguridad, en Madrid, y el Gobierno Civil de Málaga. La reconstrucción se ha realizado a partir de 14 telegramas encriptados, localizados en el Archivo Histórico Provincial de Málaga, que fueron recibidos en su día por el Gobierno Civil. Como es sabido, este tipo de cifradores no ofrece gran dificultad para recuperar los mensajes. Sin embargo, la recuperación de la tabla de homófonos y de las cintas fija y móvil no es una tarea trivial. En este caso, se han recuperado utilizando un modelo basado en rotaciones cíclicas que hace innecesaria la búsqueda exhaustiva entre todas las permutaciones posibles del alfabeto empleado. De este modo, la Clave Pilar se añade a la lista de sistemas de cinta móvil utilizados en España que han sido identificados hasta la fecha.**

***Index Terms*- criptoanálisis, cifrado simétrico, homófonos, cinta móvil.**

I. Introducción

La gran mayoría de los archivos históricos españoles incluyen entre sus fondos documentos encriptados que no han sido descifrados todavía. Este es el caso del Archivo Histórico Provincial de Málaga (AHPM) perteneciente a la Consejería de Cultura de la Junta de Andalucía, que alberga 14 telegramas encriptados, bajo la signatura "[1940]. Málaga. Archivo Histórico Provincial de Málaga. Secc. Gobierno Civil. Sig. 77145/07".

Estos telegramas fueron remitidos por la Dirección General de Seguridad, en Madrid, al Gobierno Civil de Málaga, durante 1940. A pesar del mal estado de conservación con que fueron incorporados al AHPM, se observa con claridad en alguno de ellos, y se confirma tras el criptoanálisis en otros, que fueron encriptados utilizando el mismo cifrador (aunque con claves de sesión diferentes), identificado como CLAVE PILAR (ver figura 1), y que corresponde a un sistema de cifrado de cinta móvil.

Dado que en el registro más reciente de este tipo de sistemas publicado en 2016 [1] no consta esta denominación entre los numerosos sistemas de cinta móvil estudiados y analizados, se hace necesario el criptoanálisis y la reconstrucción de todo el sistema de cifrado.

Tras una primera inspección se observa que el contenido de los telegramas es principalmente texto cifrado, aunque en algunos documentos se combinan texto en claro con texto cifrado, muy al estilo de la época. En un número importante de estos telegramas se observan descifrados parciales, mediante anotaciones manuscritas bajo el texto cifrado que indican el correspondiente carácter del mensaje en claro; aspecto que facilita la recuperación de los mensajes, si bien, como es sabido, esta tarea no ofrece grandes dificultades en

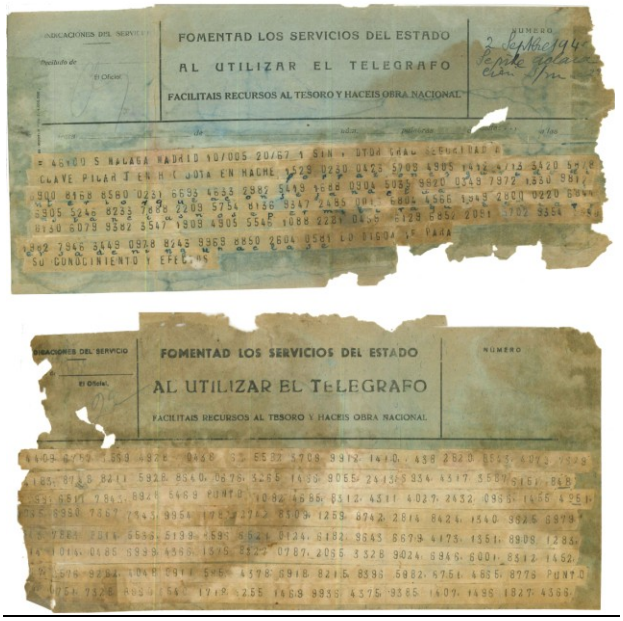


Fig. 1. Telegramas encriptados con la Clave PILAR. [1940]. Málaga. Archivo Histórico Provincial de Málaga. Secc. Gobierno Civil. Sig. 77145/07 .

este tipo de cifrado [2,3]. En consecuencia, el objetivo principal de este trabajo no es el descifrado de los mensajes sino la reconstrucción de la tabla de homófonos y de las cintas fija y móvil. Para ello, y tras un criptoanálisis exhaustivo de los mensajes, se comprueba que el conjunto de documentos no corresponde a 14 telegramas independientes, sino únicamente a 8 mensajes recibidos, que al ser transcritos utilizando los impresos oficiales de aquel tiempo ocuparon, en algunos casos, más de un ejemplar.

El método más reciente para recuperar la cinta móvil fue propuesto en 2013 por Campos *et al* [4], basado en la utilización de algoritmos genéticos para realizar búsquedas entre las 26! combinaciones de letras posibles (para un alfabeto estándar de 26 letras). Sin embargo, este método necesita conocer previamente la tabla de homófonos original.

Como alternativa, dado que no se conoce la tabla de homófonos de la CLAVE PILAR, se ha desarrollado en este trabajo un método, con un menor coste computacional, basado en un modelo del comportamiento del sistema que permite reconstruir simultáneamente la cinta móvil y la tabla de homófonos, así como una reconstrucción parcial de la cinta fija, a partir de diversos telegramas cifrados con claves de sesión distintas.

En la sección II, se describe el funcionamiento y las características principales del cifrado de cinta móvil, así como

los métodos de criptoanálisis aplicables en función de la información conocida. Las secciones III y IV describen la recuperación de los homófonos y de las cintas fija y móvil. Finalmente, las conclusiones se detallan en la sección V.

II. SISTEMAS DE CIFRADO DE CINTA MÓVIL

El sistema de cinta móvil fue propuesto por el Gobierno de España a finales del siglo XIX como método oficial para toda la Administración y estuvo en uso durante un período aproximado de 50 años [2,3]. Es un caso particular de sustitución homofónica [5] en el que todos los elementos del alfabeto en claro tienen asignados homófonos. Debido a este uso extensivo de los homófonos, bien se podría entender también como la generalización del proceso de sustitución homofónica. Las primeras implementaciones de este sistema seguían la definición inicial que ofrece Carmona [2] en 1894 en la que se utilizaban como homófonos todos los números comprendidos entre el 10 y el 99. Sin embargo, durante la Guerra Civil comenzaron a utilizarse como homófonos todos los números decimales de 2 dígitos, esto es, del 00 al 99 [3].

Como se puede observar en la figura 2, los sistemas de cinta móvil se implementan mediante una tabla impresa con tantas columnas como elementos tenga el alfabeto en claro **A**, donde se colocan los homófonos. Esta tabla se denomina *tabla de homófonos* **H,** siendo una representación gráfica de todos los conjuntos de homófonos que se utilizarán en el sistema, y constituyendo el núcleo del cifrado. Por lo tanto, una vez definida la tabla **H**, los conjuntos de homófonos no se pueden modificar, pero sí la letra a la que representan. Las modificaciones de **H** dieron lugar a cifrados distintos, que fueron identificados utilizando el término "CLAVE" [3]. Así, la CLAVE PILAR indica la utilización de una tabla de homófonos concreta, que se mantenía inalterada y en secreto.

Para determinar la asignación de los homófonos a los elementos del alfabeto en claro **A** se utilizaban generalmente dos cintas, una fija y otra móvil. La cinta móvil es la que determina la asignación de los homófonos a cada una de las letras, de manera que una vez colocada en su posición definitiva, los homófonos que se utilizan para encriptar una letra son los de la columna en la que se ha colocado dicha letra. Así, en el sistema de la figura 2, la letra "A" se encripta utilizando cualquiera de los homófonos 16, 34, 46 y 85. Por lo tanto, la cinta móvil convierte la columna $\mathbf{H}_{11}$ en el conjunto de homófonos H("A").

Una vez que la cinta móvil se posiciona no se mueve hasta que finaliza el cifrado del mensaje. Nótese que la cinta móvil contiene dos alfabetos consecutivos para facilitar el movimiento en ambos sentidos.

Por otra parte, la cinta fija tiene únicamente la misión de servir de referencia para identificar la posición en la que se coloca la cinta móvil. Esta posición es lo que en la práctica se utiliza como clave de sesión, pues permite generar distintos mensajes cifrados sin necesidad de modificar la tabla de homófonos **H**. La posición se determina indicando dos letras, una de cada cinta, que coinciden en la misma columna. En la figura 2, la posición de la cinta móvil se podría indicar como "T en F", indicando que la "T" de la cinta móvil coincide con la "F" de la fija en la misma columna.

El cambio de clave supone un cambio de posición de la cinta móvil y, en consecuencia, un cambio en el conjunto de homófonos que se utiliza con cada letra. Existen, por tanto, tantas claves como posiciones de la cinta móvil, esto es, el

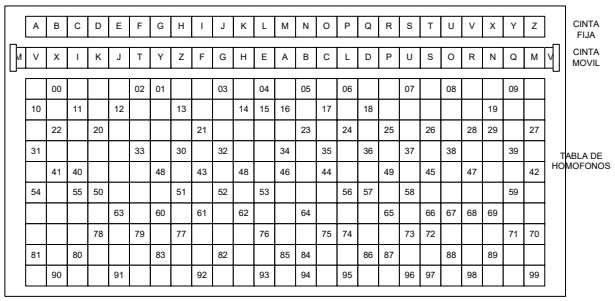


Fig. 2. Sistema de cifrado de cinta móvil.

cardinal de A, denotado como $|\mathbf{A}|$. Se define, entonces, el espacio de claves de sesión $\mathbf{K}_S$ como el conjunto de posiciones distintas de la cinta móvil con respecto a la tabla **H**, donde $|\mathbf{K}_S| = |\mathbf{A}|$. En consecuencia, se puede establecer la siguiente definición.

Definición 1. *Sea* **H** *la tabla de homófonos de un sistema de cinta móvil, y* $\mathbf{H}_i$, $1 \le i \le |\mathbf{A}|$, *las columnas de dicha tabla definidas para un alfabeto en claro* **A**. *El conjunto de homófonos* $\mathbf{H}^{(k)}(a)$, $a \in$ A, *para una clave* $k \in \mathbf{K}_S$, *queda determinado por la relación* $\mathbf{H}^{(k)}(a) = \mathbf{H}_i$, *siendo i la posición que ocupa el elemento a en la cinta móvil una vez colocada según la posición establecida por la clave k.*

Originalmente, el sistema de cinta móvil se define con una cinta fija que contiene el alfabeto español ordenado, mientras que la cinta móvil se basa en una permutación aleatoria de dicho alfabeto [2]. Años más tarde, como se indica en [3], se aplican modificaciones que utilizan alfabetos desordenados en las dos cintas.

A. Métodos de criptoanálisis

Los distintos métodos de criptoanálisis aplicables a estos sistemas se pueden clasificar en función del conocimiento de los parámetros que determinan su funcionamiento, esto es, la tabla de homófonos **H**, la cinta fija *CF*, la cinta móvil *CM* y la clave de sesión *k*. A continuación se relacionan ordenados de menor a mayor complejidad en su ejecución.

a. *Criptoanálisis con clave desconocida*. El atacante conoce todos los parámetros excepto la clave de sesión *k* utilizada para cifrar el mensaje. Este criptoanálisis es el más sencillo de resolver. Basta con colocar la cinta móvil en una posición y descifrar. Si el resultado no proporciona un resultado inteligible, se desplaza la cinta móvil una posición y se vuelve a descifrar. El número máximo de intentos es la longitud de la cinta, esto es, $|\mathbf{A}|$.

b. *Criptoanálisis con clave desconocida y cinta fija desconocida*. Es equivalente al anterior ya que la cinta fija solo se utiliza como referencia para posicionar la cinta móvil, quedando reflejada directamente en la clave *k* de sesión utilizada. En consecuencia, se debe proceder de la misma forma que en el criptoanálisis anterior, esto es, probando todas las posiciones de la cinta móvil.

c. *Criptoanálisis conocida la tabla de homófonos*. Si el atacante conoce la tabla de homófonos **H**, pero no conoce el resto de parámetros, el enfoque clásico indica que el análisis estadístico de los mensajes es lo más adecuado puesto que el efecto

que se desea conseguir con la utilización de homófonos desaparece al conocer todos los que se utilizan con cada letra.

d. *Criptoanálisis conocidos los conjuntos de homófonos*. El atacante conoce los conjuntos de homófonos $\mathbf{H}^{(k)}(a)$, pero no conoce la relación que existe con las columnas $\mathbf{H}_i$ de la tabla de homófonos **H.** Esto es equivalente a decir que se conocen las columnas que forman la tabla pero no el orden en el que deben colocarse para reconstruir la tabla **H** original. Por lo tanto, no se pueden aplicar los métodos descritos en los apartados anteriores. Esta situación se produce cuando el atacante dispone de varios mensajes que han sido cifrados con claves distintas, aunque sean desconocidas. Se considera, por tanto, que el atacante conoce las columnas de homófonos utilizadas con las letras del alfabeto en distintas situaciones (distintas claves). Esto permite recuperar simultáneamente la cinta móvil y la tabla de homófonos, si se desarrolla algún mecanismo que reordene adecuadamente las columnas.

e. *Criptoanálisis de texto cifrado*. Este es el ataque más difícil de implementar, dado que el atacante no conoce ningún elemento del criptosistema. La manera más adecuada de proceder es mediante análisis estadístico para descifrar el mensaje y obtener, al menos parcialmente, las columnas de homófonos utilizadas, y las letras asignadas a cada una de ellas.

En el caso particular de la CLAVE PILAR, el método que se aplica es una combinación de los indicados en d) y e), puesto que, en unos casos, los telegramas aparecen con indicaciones sobre el carácter en claro que corresponde a los símbolos del mensaje cifrado, lo que supone un cierto conocimiento parcial de los conjuntos de homófonos (no de la tabla) y, en otros, solo se conoce el texto cifrado.

III. Reconstrucción del conjunto de homófonos

El texto encriptado se encuentra distribuido en 14 documentos que han sido identificados como TG1 a TG14. Estos documentos constituyen un total de ocho mensajes recibidos (M1 a M8) en formato telegrama, que fueron encriptados utilizando distintas claves. La relación entre mensajes, telegramas y claves aparece en la tabla I. Esta tabla recoge, por una parte, la información en claro ubicada en las cabeceras de los telegramas, y por otra, el resultado del criptoanálisis, que ha revelado la utilización de otras claves, como ocurre con el mensaje M5, en el que la clave inicial (L en E) se modifica dos veces a lo largo del mensaje. La sincronización se realiza insertando el texto (encriptado) "CAMBIE CLAVE A" seguido de la nueva clave de sesión. Asimismo, cabe destacar que algunos mensajes son repeticiones de otros mensajes, solicitados por el receptor para confirmar la veracidad del mensaje descifrado.

La reconstrucción del conjunto de homófonos requiere el análisis de los telegramas recibidos. Para ello, se seleccionan grupos de telegramas encriptados con la misma clave (tabla I) que contengan la mayor cantidad de caracteres posible, con objeto de recuperar el mayor número posible de homófonos. Así, se elige el grupo formado por TG1, TG2 y TG3, cifrados con la clave J en H; el grupo formado por TG4 y TG5, cifrados con J en K; y el grupo formado por TG8 y TG9. Nótese que otros telegramas, como TG7, TG6 y TG10, no se incluyen en los grupos a pesar de estar encriptados con las claves mencionadas. Esto se debe a que inicialmente no se conocía dicha clave, siendo incluidos en la tabla I una vez finalizado el proceso.

Tabla I
CLAVES EMPLEADAS PARA ENCRIPTAR LOS TELEGRAMAS

Mensaje	Telegrama	Clave
M1	TG1	J en H
M2	TG2	J en H
M3	TG3+TG7	J en H
M4	TG4+TG5	J en K
M5	TG8+TG13+TG11	L en E + M en J + S en Z
M6	TG9+TG6+TG10	L en E
M7	TG12	N en S
M8	TG14	U en L

El análisis de todos estos telegramas ha permitido identificar un total de 28 columnas de homófonos distintas, que incluyen los 100 valores numéricos distintos (00-99) que se utilizan en este sistema de cifrado, con lo que se garantiza que no puede haber más columnas. En consecuencia, se concluye que el número de columnas de **H** es 28, así como las longitudes de las cintas móvil y fija, lo que corresponde con las 26 letras habituales, a las que se añaden la "Ch" y la "Ñ". Este último aspecto se deduce de un descifrado parcial en el que aparecen sendos caracteres. No obstante, a pesar de que el alfabeto incluye las letras "Ch" y "Ñ", no siempre son utilizadas en los mensajes cifrados, empleando en ocasiones la técnica habitual de sustituir la "Ñ" por "N", y la de separar la "Ch" como "C" y "H".

El resultado de esta fase de análisis es la recuperación de los conjuntos de homófonos que permiten descifrar únicamente los telegramas encriptados analizados, o cualquier otro en el que se haya aplicado la misma clave.

Con objeto de facilitar el posterior análisis, se define la tabla de homófonos equivalente $\mathbf{H}^{(k)}$ para una clave sesión k, del siguiente modo:

Definición 2. *Se denomina tabla de homófonos equivalente, o simplemente tabla equivalente,* $\mathbf{H}^{(k)}$*, al conjunto ordenado de columnas* $\mathbf{H}_i^{(k)}$, $1 \le i \le |\mathbf{A}|$*, de tal forma que* $\mathbf{H}_i^{(k)} = \mathbf{H}^{(k)}(a_i)$ *siendo* a_i *los elementos del alfabeto español ordenado de tamaño* $|\mathbf{A}|$*, es decir,* a_1 = "A", a_2 = "B", …, $a_{|\mathbf{A}|}$ = "Z".

Según esta definición, cualquier sistema de cinta móvil tiene una única tabla **H** de homófonos y $|\mathbf{A}|$ tablas $\mathbf{H}^{(k)}$ equivalentes, puesto que solo hay $|\mathbf{A}|$ claves de sesión distintas. Además, cada tabla equivalente se obtiene mediante permutación de las columnas de **H** y, en consecuencia, de las de cualquier otra tabla equivalente. Nótese que $\mathbf{H}^{(k)}$ se puede entender también como la reordenación de las columnas de **H** que permite la automatización o implementación de la fase de encriptado, en un computador, puesto que la ordenación toma como índice las letras del alfabeto, facilitando así la búsqueda de los homófonos que se deben aplicar para encriptar cada carácter. En el caso particular de una cinta móvil con alfabeto ordenado las permutaciones que convierten una tabla equivalente en otra distinta resultan ser simples rotaciones.

Tras realizar el criptoanálisis de los telegramas seleccionados, se obtienen sendas tablas equivalentes que, como muestra la tabla II, resultan incompletas, pero permiten establecer relaciones entre las mismas y, en consecuencia, entre las distintas claves de sesión. Así, por ejemplo, los homófonos utilizados para encriptar la letra "C" con la clave "J en H" son los mismos que los utilizados para encriptar la "A" con la clave "J en K" y los mismos que los utilizados para la "Z" con la clave "L en E". A pesar de ello, la falta de algunos caracteres en estas tablas dificulta notablemente la recuperación de la **H** original a partir de estas sencillas relaciones.

IV. Reconstrucción de H y de las cintas fija y móvil

El primer paso para la reconstrucción, tanto de **H** como de las cintas, es comprobar si la cinta móvil contiene un alfabeto ordenado o desordenado (aleatorio). Para ello, se comprueba si cualquiera de las tablas equivalentes recuperadas en la sección anterior se puede generar mediante rotaciones cíclicas de las demás. El descifrado de la clave "J en H" con respecto al de "J en K" no refleja este comportamiento. En el primer caso, las letras "H","I" y "J" tienen asociadas las columnas que contienen los valores 18, 13 y 16, respectivamente. Sin embargo, estos valores no aparecen en columnas consecutivas en la tabla equivalente de "J en K", sino asociados a las letras "J","G" y "U", respectivamente. Por lo tanto, se concluye que la cinta móvil no está formada por el alfabeto ordenado.

El siguiente paso consiste, entonces, en estudiar los ciclos que se forman por efecto del desplazamiento de la cinta móvil, con el fin de identificar el orden en el que aparecían las letras en la cinta móvil. Este desplazamiento, δ, se define como la rotación cíclica que se debe aplicar a la cinta móvil para pasar de la posición P_1 a la P_2, definidas por las claves k_1 y k_2, respectivamente. La figura 3 muestra un desplazamiento δ = 7 entre dos posiciones P_1 y P_2 de una cinta con alfabeto de tamaño 25, lo que permite obtener el ciclo de la permutación A-S-X-F-L-N-J-E-U-V-Z-C-R-K-H-P-M-Y-B-O-I-G-D-Q-T-A.

En general, la longitud L_δ de estos ciclos depende del desplazamiento δ y de la longitud |**A**| del alfabeto, y responde a la siguiente expresión

$$L_\delta = \frac{|A|}{\text{mcd}\,(\delta,|A|)} \quad (1)$$

Por lo tanto, al realizar la comparación entre los homófonos recuperados de distintas claves, utilizando la información de la tabla II, aparecerán ciclos que necesariamente tendrán una la longitud divisor de 28, según la *Ec* (1), esto es,

$$L_\delta \in \{2,4,7,14,28\} \quad (2)$$

Analizando los homófonos de la clave "J en K" con respecto a los de la "J en H", se obtienen los siguientes ciclos incompletos:

K-C-A-L-V-X-
S-H-J-U-T-E-F-I-G-D-M-Y-P-
O-B-Z-N-
Q-R-

Del mismo modo, los homófonos de la clave "L en E" con respecto a la clave "J en H" generan otras relaciones, también incompletas:

Tabla II
RELACIÓN ENTRE HOMÓFONOS DE DISTINTAS CLAVES

J en H	J en K	L en E	J en H	J en K	L en E
A	L	N	O	B	
B	C	R	P		D
C	A	Z	Q	R	V
D	M	F	R		
E	F	J	S	H	Y
F	I	U	T	E	H
G	D	E	U	T	S
H	J	P	V	X	C
I	G	T	X		A
J	U	Ñ	Y	P	
K	C	B	Z	N	
L	V			O	Ch
M	Y	I		Q	L
N		O		S	M

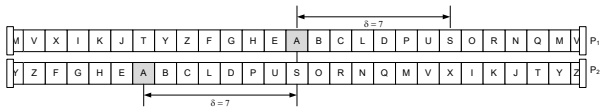


Fig. 3. Desplazamiento de la cinta móvil.

X-A-N-O-
Q-V-C-Z-
M-I-T-H-P-D-F-U-S-Y-
G-E-J-Ñ-
K-B-R-

Estas relaciones representan la secuencia de letras que se obtiene al recorrer la cinta saltando δ_1 posiciones en el caso "J en K", y δ_2 posiciones en el caso "L en E".

Observando todas las secuencias incompletas se puede identificar que las letras "M" y "P" se encuentran separadas 2 posiciones en las secuencias de "J en K" y 4 en las secuencias de "L en E". Esta relación se debe mantener para el resto de las letras, siempre que pertenezcan al mismo ciclo, lo que permite completarlos obteniendo dos ciclos de longitud $L_{\delta 1}$ = 14, para la clave "J en K" con respecto a "J en H",

S–H–J–U–T–E–F–I–G–D–M–Y–P–Ñ–S
K–C–A–L–V–X–Ch–Q–R–W–O–B–Z–N–K

y cuatro ciclos de longitud $L_{\delta 2}$ = 7 para la clave "L en E" con respecto a "J en H",

S–M–F–J–P–G–T–S
D–E–H–Y–I–U–Ñ–D
K–O–Ch–A–Z–R–V–K
C–B–Q–L–N–W–X–C

Esta reconstrucción de los ciclos se efectúa simultáneamente, ya que la reconstrucción parcial de un ciclo de "J en K" afecta a los ciclos de "L en E", y viceversa.

En consecuencia, aplicando la Ec (1) para el caso de la clave "J en K" respecto de "J en H" se obtiene que mcd(δ_1, 28) = 2, y por lo tanto,

$$\delta_1 \in \{2,6,10,18,22,26\} \quad (3)$$

Nótese que estos seis valores suponen realmente tres desplazamientos con distancias distintas, puesto que los desplazamientos 18, 22 y 26 corresponden a -2, -6 y -10, respectivamente. Es decir, el conjunto de valores posibles de δ_1 corresponde a rotaciones de 2, 6 y 10 posiciones, tanto a derecha como a izquierda. Este resultado proporciona información sobre las posiciones relativas de las letras en la cinta móvil, que reduce a tres el conjunto de posibles configuraciones, considerando únicamente las rotaciones a derecha, esto es, con desplazamientos positivos. Así, los caracteres que componen cada uno de los ciclos ocuparán posiciones en la cinta móvil separadas 2, 6 o 10 posiciones. Sin embargo, este análisis no proporciona información sobre el modo exacto en que deben combinarse los dos ciclos en la cinta. En cualquiera de los casos ($\delta_1 = 2$, $\delta_1 = 6$, $\delta_1 = 10$), el segundo ciclo se puede colocar en los huecos libres del primero, y viceversa, comenzando en cualquiera de ellos, puesto que ambos ciclos son independientes.

Es importante recordar que todas estas configuraciones son soluciones válidas para el sistema de cifrado, y que por tanto deben considerarse como cintas equivalentes que permiten descifrar los telegramas cifrados con la clave "J en K".

Por otra parte, de las relaciones "L en E" con "J en H" se obtiene que mcd(δ_2, 28) = 4, y por lo tanto,

$$\delta_2 \in \{4,8,12,16,20,24\} \quad (4)$$

Se puede comprobar fácilmente, a partir de la observación de los ciclos, que cada valor de δ_1 se corresponde con un valor diferente de δ_2, estableciéndose la siguiente relación

$$\delta_2 = -4 \cdot \delta_1 (\text{mod } 28) \quad (5)$$

No obstante, esta relación no resuelve el problema de la ubicación en la cinta de todos los ciclos, puesto que en esta ocasión los 4 ciclos de longitud 7 se han formado al dividirse los dos ciclos de longitud 14, de manera que sigue existiendo independencia entre los elementos de los dos ciclos de longitud 14.

Para avanzar en la solución del problema es necesario encontrar una situación en la que aparezcan ciclos con elementos que hasta ahora han aparecido en ciclos diferentes. Es el caso del telegrama TG13, cuyo análisis de homófonos genera, por comparación con los obtenidos para la clave "J en K", un único ciclo de longitud 28.

S–B–Y–R–G–X–E–A–J–N–Ñ–O–M–Q–I–V–T–C–H–Z–
–P–W–D–Ch–F–L–U–K–S

Aplicando la Ec (1) se tiene que $\text{mcd}(\delta_3, 28) = 1$, lo que permite acotar el valor del desplazamiento δ_3 a un conjunto de doce valores, esto es,

$$\delta_3 \in \{1,3,5,9,11,13,15,17,19,23,25,27\} \quad (6)$$

de los cuales, la mitad corresponden con las rotaciones simétricas, esto es, en sentido inverso. Así, si tomamos como referencia el valor $\delta_3 = 1$, se tiene la composición de la cinta que aparece en la figura 4.

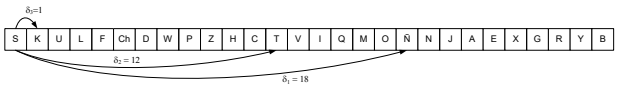


Fig. 4. Cinta móvil reconstruida.

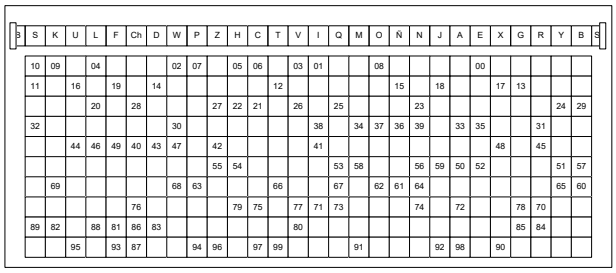


Fig. 5. Tabla de homófonos de la CLAVE PILAR.

Tabla III
DESPLAZAMIENTOS DE LAS CINTAS EQUIVALENTES
DE LA CLAVE PILAR

δ_3	δ_2	δ_1
1	12	-10
3	8	-2
5	4	6
9	-4	-6
11	-8	2
13	-12	10
-13	12	-10
-11	8	-2
-9	4	6
-5	-4	-6
-3	-8	2
-1	-12	10

En general, las relaciones entre los ciclos de los tres casos analizados se pueden representar mediante las siguientes equivalencias:

$$\delta_1 = 18 \cdot \delta_3 \,(\text{mod } 28) \quad (7)$$

$$\delta_2 = 12 \cdot \delta_3 \,(\text{mod } 28) \quad (8)$$

A partir de esta relación de equivalencia se puede establecer la tabla III en la que aparecen las 12 posibles configuraciones de la cinta, todas ellas equivalentes y válidas para poder descifrar los mensajes cifrados con la CLAVE PILAR, y que corresponden realmente a los 12 valores que puede tomar el desplazamiento δ_3. Las seis últimas son las configuraciones simétricas de las seis primeras. Por lo tanto, la tabla proporciona el método para construir 12 cintas móviles que permiten descifrar los mensajes cifrados con el sistema de cinta móvil denominado CLAVE PILAR. El método consiste en disponer la cinta con los 28 espacios en blanco y comenzar colocando todos los elementos del alfabeto según el orden establecido en el ciclo de longitud 28 obtenido anteriormente, S-B-Y-R-G-X-E-A-J-N-Ñ-O-M-Q-I-V-T-C-H-Z-P-W-D-Ch-F-L-U-K-S, o, equivalentemente, en orden inverso, cada δ_3 posiciones.

A partir de la cinta móvil reconstruida se genera la correspondiente tabla de homófonos colocando las columnas de homófonos en el orden que marcan los caracteres de la cinta. Así, tomando como referencia la cinta móvil de la figura 4, se obtiene la tabla de la figura 5.

Hay que considerar, además, que mientras no se conozca la cinta fija, cualquier rotación cíclica de la tabla H generada a partir de una cinta móvil es válida para descifrar los mensajes de la CLAVE PILAR. No obstante, el desconocimiento de la cinta fija no impide la recuperación del mensaje. Para ello, basta aplicar el método descrito en la sección II.A.b.

Tomando como referencia la tabla de homófonos de la figura 5, y teniendo en cuenta que no se conoce la cinta fija, se numeran las columnas para poder identificar con claridad el desplazamiento de la cinta móvil que permite el descifrado correcto de cada mensaje. Por lo tanto, las 28 claves de sesión quedan identificadas por la columna en la que se encuentre la letra "S", a contar desde la izquierda, tal como se indica en la figura 6. Así, la tabla IV contiene la correspondencia numérica de las claves utilizadas en los telegramas analizados, que permite una recuperación parcial de la cinta fija (figura 7).

V. CONCLUSIONES

El criptoanálisis realizado sobre el conjunto de 14 telegramas localizados en el AHPM ha permitido reconstruir el criptosistema de cinta móvil, denominado CLAVE PILAR, utilizado en las comunicaciones oficiales del Gobierno de España, en 1940, entre la Dirección General de Seguridad, en Madrid, y el Gobierno Civil de Málaga, lo que supone un ejercicio práctico de criptoanálisis acompañado de un notable esfuerzo forense para la lectura de algunos caracteres debido al mal estado de los documentos y de los errores cometidos en las transmisiones originales, tal como reflejan algunas anotaciones de los telegramas recibidos.

La reconstrucción del sistema incluye la recuperación de la cinta móvil y de la tabla de homófonos, que constituyen el núcleo de este tipo de sistemas. Más concretamente, el estudio realizado sobre estos documentos ha permitido generar 12 tablas de homófonos distintas, junto con sus cintas móviles correspondientes, que son equivalentes a la original, es decir, que permiten descifrar cualquier mensaje cifrado con la CLAVE PILAR a través de cualquiera de las 28 claves de sesión posibles. Para ello, es suficiente con probar la tabla **H** con las 28 claves de sesión, que se corresponden con las 28 posiciones de la cinta móvil sobre **H**.

Por otra parte, la información obtenida de los telegramas analizados permite realizar únicamente una reconstrucción parcial de la cinta fija, tal como refleja la figura 7.

A partir de ahora, la CLAVE PILAR pasará a formar parte de la relación de sistemas de cinta móvil registrados y analizados en la literatura.

Por último, es interesante mencionar que el texto en claro recuperado tras el criptoanálisis de los telegramas no se ha incluido en este trabajo por no constituir una contribución relevante desde el punto de vista criptográfico.

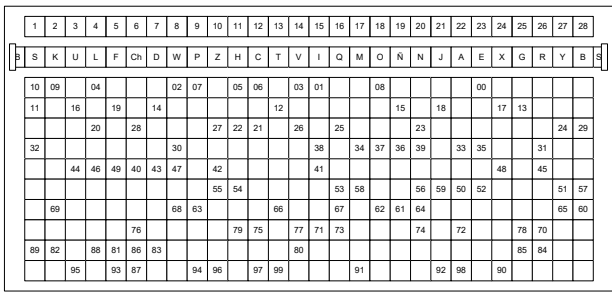


Fig. 6. Reconstrucción de la CLAVE PILAR.

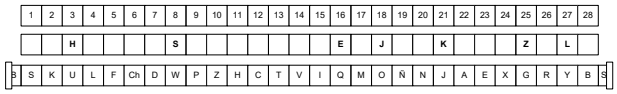


Fig. 7. Reconstrucción parcial de la cinta fija de la CLAVE PILAR.

Tabla IV
CORRESPONDENCIA NUMÉRICA
DE LAS CLAVES UTILIZADAS

Clave original	Clave numérica
J en H	11
J en K	1
L en E	13
M en J	2
S en Z	25
N en S	17
U en L	25

AGRADECIMIENTOS

Este trabajo ha sido financiado parcialmente por el Campus de Excelencia Internacional Andalucía Tech de la Universidad de Málaga y por el Ministerio de Economía, Industria y Competitividad (MINECO), la Agencia Estatal de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER, UE), a través del proyecto COPCIS, referencia TIN2017-84844-C2-1-R.

REFERENCIAS

[1] J.R. Soler Fuensanta, F.J. López Brea Espiau, *Mensajes secretos. La historia de la criptografía española desde sus inicios hasta los años 50*, Ed. Tirant lo Blanch, Valencia 2016.

[2] J. García Carmona, *Tratado de criptografía con aplicación especial al ejército*. Madrid, Sucesores de Rivadeneyra (1894). Re-editado por el Ministerio de Defensa, 2011

[3] J.R. Soler Fuensanta, F.J. López-Brea Espiau, "The Strip Cipher-The Spanish Official Method", *Cryptologia*, 31, pp.46-56, 2007.

[4] F.A. Campos, A. Gascón, J.M. Latorre, J.R. Soler, "Genetic algorithms and mathematical programming to crack the Spanish strip cipher", *Cryptologia*, 37, pp.51-68, 2013.

[5] D. Kahn, *The Codebreakers: The Story of Secret Writing*. New York, Macmillan. 1967

Adecuación al Esquema Nacional de Seguridad: Una Aproximación Metodológica

Pedro García Teodoro, José Antonio Gómez Hernández
Universidad de Granada - ETSIIT
Network Engineering & Security Group
{pgteodor,jagomez}@ugr.es

***Resumen*—El Esquema Nacional de Seguridad (ENS) persigue fundamentar la confianza de la sociedad en los sistemas de información y de las comunicaciones. A pesar de la oportunidad normativa del ENS, su adopción práctica por parte de empresas y organizaciones no está exenta de dificultades, lo que produce cierto rechazo y, en consecuencia, constituye un hándicap en la consecución del objetivo último perseguido. Entendiendo que una de las principales dificultades para la adecuación al ENS es su escasa formulación metodológica, el presente trabajo realiza un breve estudio del mismo donde, de forma clara y concisa, se exponen sus objetivos, alcance, dimensiones y procedimientos. Además, se proponen unas plantillas que recogen de forma sistemática las diferentes medidas y aspectos a tener presentes en el proceso de implantación. Nuestro ánimo con ello es facilitar la necesaria adopción del ENS en nuestra sociedad para un uso más racional y confiable de las TIC.**

***Index Terms*—Categoría, confianza, ENS, metodología, medida de seguridad, nivel de seguridad**

Tipo de contribución: *Investigación original*

I. Introducción

La dependencia social actual de las TIC es un hecho que se evidencia en base a la presencia de estas tecnologías en todos los aspectos de nuestra vida cotidiana: ocio, cultura, economía, política, etc. La cantidad y relevancia de la información manejada actualmente por los servicios y sistemas TIC se demuestra cada vez mayor, lo cual lleva aparejados numerosos y serios riesgos de seguridad [1], [2]. Sin entrar en una relación detallada de incidentes de seguridad, sírvanse citar algunos casos recientes como es el secuestro de equipos y datos ocurrido con *Wannacry* en mayo de 2017 [3] o el de exposición de datos sensibles a través de la *app* de *tracking* y localización *Strava* anunciado en enero de 2018 [4].

Esta relevancia e impacto de las TIC ha provocado la aparición de diversas normas jurídicas orientadas a la necesaria protección de servicios y sistemas que garanticen la seguridad de la información relativa a usuarios y organizaciones conforme a diversas leyes y convenios nacionales e internacionales [5], [6], [7]. Entre otras normas relacionadas, cabe citar a nivel nacional el *Esquema de Seguridad Nacional* (ENS) [8], [9], de obligado cumplimiento para entidades que trabajen con la Administración del Estado, y a nivel europeo la conocida como *Directiva NIS* [10], de aplicación general a cualquier entidad que maneje datos de carácter personal y cuyo plazo de transposición está fijado para mayo de 2018.

La interacción entre legisladores, académicos y empresarios/gestores no suele caracterizarse por ser demasiado operativa, lo cual está motivado principalmente por la naturaleza 'disjunta' de los respectivos ámbitos en la práctica. Objetivos, prioridades, 'lenguaje' incluso, e impacto económico cómo no, constituyen aspectos diferenciadores y separadores de estos mundos; mundos que, sin embargo, están abocados a entenderse y converger para la necesaria securización y avance de nuestra sociedad. Entre los escasos estudios que abordan la toma de decisiones de empresas y organizaciones en lo que respecta a la inversión en seguridad en sus entornos podemos citar a modo de ejemplo [11], [12], los cuales concluyen, como apuntábamos, la ya citada separación y necesario entendimiento de los agentes implicados en el campo.

En este contexto, el presente trabajo lleva a cabo una propuesta procedimental sobre el ENS con un triple objetivo. Por una parte, una mejor comprensión de sus objetivos y alcance. Por otra y principal, el estudio sistemático simplificado de su adopción en distintos entornos. Ello concluirá en una reducción de los costes de implantación asociados. En defintiva, la propuesta aquí presentada pretende facilitar la adopción generalizada del ENS en empresas y organizaciones.

El resto del trabajo se organiza como sigue. En la Sección II se describe a grandes rasgos el ENS, comentándose sus objetivos y líneas de actuación. Tras ello, en la Sección III se descibe la metodología desarrollada por los autores a fin de facilitar y sistematizar el estudio de la seguridad de un entorno y su adecuación al ENS. Finalmente, en la Sección IV se concluyen las aportaciones principales del trabajo y se apuntan nuevas actuaciones de futuro en el campo.

II. Esquema Nacional de Seguridad

El ENS plantea como objetivos principales:

1. Crear las condiciones necesarias de confianza en el uso de los medios electrónicos, a través de medidas para garantizar la seguridad de los sistemas, los datos, las comunicaciones, y los servicios electrónicos, que permita a los ciudadanos y a las Administraciones públicas, el ejercicio de derechos y el cumplimiento de deberes a través de estos medios.
2. 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.

Para dar cumplimiento a lo anterior se establecen:

- dimensiones de seguridad y sus niveles,
- categoría de los sistemas,
- medidas de seguridad adecuadas y la auditoría periódica de la seguridad,
- elaboración de un informe para conocer regularmente el estado de seguridad de los sistemas de información a los que se refiere el real decreto, y el

Figura 1: Medidas de seguridad recogidas en el ENS (*Fuente*: CCN).

- papel de la capacidad de respuesta ante incidentes de seguridad de la información del Centro Criptológico Nacional.

Por lo que respecta a las dimensiones de seguridad, son 5 las consideradas:

- *Autenticidad* [A]: Propiedad o característica consistente en que una entidad es quien dice ser o bien que garantiza la fuente de la que proceden los datos.
- *Confidencialidad* [C]: Propiedad o característica consistente en que la información ni se pone a disposición, ni se revela a individuos, entidades o procesos no autorizados.
- *Disponibilidad* [D]: Propiedad o característica de los activos consistente en que las entidades o procesos autorizados tienen acceso a los mismos cuando lo requieren.
- *Integridad* [I]: Propiedad o característica consistente en que el activo de información no ha sido alterado de manera no autorizada.
- *Trazabilidad* [T]: Propiedad o característica consistente en que las actuaciones de una entidad pueden ser imputadas exclusivamente a dicha entidad.

Cada una de ellas contempla 3 posibles niveles de seguridad:

- *Bajo*: Cuando las consecuencias de un incidente de seguridad que afecte a alguna de las dimensiones de seguridad supongan un perjuicio limitado sobre las funciones de la organización, sobre sus activos o sobre los individuos afectados
- *Medio*: Cuando las consecuencias de un incidente de seguridad que afecte a alguna de las dimensiones de seguridad supongan un perjuicio grave sobre las funciones de la organización, sobre sus activos o sobre los individuos afectados.
- *Alto*: Cuando las consecuencias de un incidente de seguridad que afecte a alguna de las dimensiones de seguridad supongan un perjuicio muy grave sobre las funciones de la organización, sobre sus activos o sobre los individuos afectados.

En función del nivel de seguridad concluido para el entorno, se derivará un informe del estado de la seguridad del mismo y, a partir de éste y el alcance e impacto de la actvidad realizada y los servicios en él provistos, la implementación de un conjunto de medidas de seguridad.

El ENS define un total de 75 medidas de seguridad agrupadas en 3 ámbitos (véase Figura 1, extraída de https://www.ccn-cert.cni.es/ens.html):

- *Marco organizativo*: Constituido por el conjunto de medidas relacionadas con la organización global de la seguridad.
- *Marco operacional*: Formado por las medidas a tomar para proteger la operación del sistema como conjunto integral de componentes para un fin.
- *Medidas de protección*: Centradas en proteger activos concretos, según su naturaleza y la calidad exigida por el nivel de seguridad de las dimensiones afectadas.

III. Metodología para la adecuación al ENS

Tomando como base la definición normativa del ENS y los documentos [13], [14] elaborados por el Centro Criptológico Nacional (CCN, https://www.ccn-cert.cni.es) para facilitar la adecuación al mismo de empresas y organizaciones, seguidamente se detalla una propuesta desarrollada para analizar de forma sencilla y sistemática el estado de seguridad de un sistema. Facilitada así la labor de análisis, todo el proceso completo de adecuación por parte de una organización dada resultará más rápido, barato y efectivo. Se prevé con ello una mayor aceptación e implantación del ENS y, en suma, un mayor y más rápido nivel de securización de la sociedad.

Sistematizando los procesos previstos en el ENS, los pasos a seguir en el estudio y adecuación de un sistema u organización son (Tabla I):

1. Determinación de los **servicios e información** implicados en el sistema.

Tabla I: Pasos metodológicos para la adecuación al ENS.

Proceso	Descripción
1. Servicios	Determinación de los servicios e información
2. Seguridad de los servicios	Estimación del nivel de seguridad de cada uno de los servicios
3. Categoría del sistema	Determinación de la categoría del sistema
4. Medidas	Concreción de las medidas de seguridad a adoptar
5. Informe	Estudio del estado actual del sistema en relación a estas medidas
6. Implementación	Implementación de las actuaciones pertinentes para desarrollar las medidas correspondientes

2. Estimación del **nivel de seguridad de cada uno de los servicios**.
3. Determinación de la **categoría del sistema** analizado.
4. Concreción de las **medidas de seguridad a adoptar**.
5. Informe del **estado actual del sistema** en relación a las medidas a aplicar.
6. Implementación de las **actuaciones pertinentes** para desarrollar plenamente las medidas correspondientes.

Todos estos procesos se describen a continuación. Los números 1-3 en el apartado III-A, el 4 en III-B y los 5-6 en III-C.

III-A. Categoría de un sistema

El objetivo principal del estudio de seguridad de un sistema dado es la determinación de su categoría, esto es, del nivel de seguridad requerido para el mismo. A este fin, lo primero que hemos de aclarar son los servicios o conjunto de informaciones dispuestos en el entorno. Por ejemplo, un servidor de correo electrónico, un dispositivo de conexión a Internet, una base de datos de clientes, etc. Sin pretender ser demasiado extensos en la identificación de servicios, cuanto mejor sea ésta más fiables serán los resultados derivados del análisis del sistema.

Realizado el proceso de identificación, seguidamente llevaremos a cabo la evaluación de seguridad de cada servicio desde la perspectiva del nivel de seguidad requerido en las distintas dimensiones. Dicho análisis podemos realizarlo en base a la plantilla mostrada en la Figura 2, donde especificaremos el impacto de cada dimensión de seguridad sobre 4 aspectos distintos: *(i)* objetivos de la organización, *(ii)* activos de la organización y servicio diario, *(iii)* normativa y legislación e *(iv)* individuos afectados. Dicho impacto se categorizará de acuerdo a 3 posibles niveles:

- *Bajo*: Cuando el impacto de la dimensión en cuestión sea reducido, subsanable.
- *Medio*: Cuando el impacto de la dimensión sea significativo, de difícil reparación.
- *Alto*: Cuando el impacto de la dimensión en cuestión sea muy grave, provocando la anulación de las capacidades asociadas y resultando irreparable.

A modo de apunte, sírvase indicar que un servicio de información tipo 'base de datos' suele ser altamente sensible a [C] e [I], mientras que uno tipo 'servidor web' lo es más a [D]. A partir de la información anterior, el **nivel de seguridad** global asociado al servicio analizado será el de mayor valor de los obtenidos para las distintas dimensiones.

Una vez realizado el análisis de todos los servicios que componen nuestro sistema completo, la **categoría** del mismo será una de las tres siguientes posibles:

- *ALTA*: Cuando alguna de las dimensiones estudiadas alcanza el nivel *Alto*.

Dimensión	Objetivos	Activos	Legislación	Derechos personas
Disponibilidad [D]				
Autenticidad [A]				
Integridad [I]				
Confidencialidad [C]				
Trazabilidad [T]				

Figura 2: Plantilla para determinar el nivel de seguridad de un servicio.

- *MEDIA*: Cuando alguna de las dimensiones estudiadas alcanza el nivel *Medio* y ninguna el *Alto*.
- *BÁSICA*: Cuando alguna de las dimensiones estudiadas alcanza el nivel *Bajo* y ninguna los niveles *Medio* o *Alto*.

III-B. Medidas de seguridad a aplicar

Como hemos comentado, la estimación de la categoría de un sistema resulta crucial, pues ella determinará cuáles de las 75 medidas de seguridad del ENS son de aplicación obligada según la norma. La tabla completa de **medidas de seguridad** por categoría puede consultarse en el Anexo II de [9][1], mostrándose en la Figura 3 las relativas al *Marco operacional* a modo de ejemplo. Así, no es necesaria la implantación de un *plan de continuidad* ('Continuidad del servicio') si la categoría del sistema es BAJA o MEDIA, mientras que la elaboración de un *inventario de activos* (''Explotación') es precisa en cualquier caso.

Es importante mencionar que cuando en un sistema existan servicios/informaciones que precisen la aplicación de un nivel de medidas de seguridad diferente de las del sistema principal, podemos llevar a cabo una segregación siempre que puedan delimitarse claramente la información y/o los servicios afectados.

III-C. Informe de estado del sistema y aplicación de medidas

Establecidas cuáles son las medidas de seguridad a aplicar en nuestro sistema, seguidamente procederemos a llevar a cabo un **informe detallado del estado actual** de desarrollo de cada una de ellas. Sólo podrá ser a partir que ello que podamos planificar adecuadamente futuras acciones orientadas a alcanzar los niveles exigidos en cada caso.

Con el fin de facilitar la revisión sistemática del estado actual de cada una de las medidas a considerar, en este trabajo proponemos el desarrollo y uso de unas plantillas tipo *checklist*, como la mostrada en las Figura 4 para el *Marco organizativo*. En ella se puede observar que en un primer nivel se explicitan las 4 medidas contempladas en el marco

[1]Es importante señalar que [9] recoge hasta 24 modificaciones en la tabla referida sobre la establecida en [8].

«Dimensiones				Medidas de seguridad	
Afectadas	B	M	A		
				op	**Marco operacional**
				op.pl	Planificación
categoría	aplica	+	++	op.pl.1	Análisis de riesgos
categoría	aplica	+	++	op.pl.2	Arquitectura de seguridad
categoría	aplica	=	=	op.pl.3	Adquisición de nuevos componentes
D	n.a.	aplica	=	op.pl.4	Dimensionamiento/Gestión de capacidades
categoría	n.a.	n.a.	aplica	op.pl.5	Componentes certificados
				op.acc	Control de acceso
A T	aplica	=	=	op.acc.1	Identificación
I C A T	aplica	=	=	op.acc.2	Requisitos de acceso
I C A T	n.a.	aplica	=	op.acc.3	Segregación de funciones y tareas
I C A T	aplica	=	=	op.acc.4	Proceso de gestión de derechos de acceso
I C A T	aplica	+	++	op.acc.5	Mecanismo de autenticación
I C A T	aplica	+	++	op.acc.6	Acceso local *(local logon)*
I C A T	aplica	+	=	op.acc.7	Acceso remoto *(remote login)*
				op.exp	Explotación
categoría	aplica	=	=	op.exp.1	Inventario de activos
categoría	aplica	=	=	op.exp.2	Configuración de seguridad
categoría	n.a.	aplica	=	op.exp.3	Gestión de la configuración
categoría	aplica	=	=	op.exp.4	Mantenimiento
categoría	n.a.	aplica	=	op.exp.5	Gestión de cambios
categoría	aplica	=	=	op.exp.6	Protección frente a código dañino
categoría	n.a.	aplica	=	op.exp.7	Gestión de incidentes
T	aplica	+	++	op.exp.8	Registro de la actividad de los usuarios
categoría	n.a.	aplica	=	op.exp.9	Registro de la gestión de incidentes
T	n.a.	n.a.	aplica	op.exp.10	Protección de los registros de actividad
categoría	aplica	+	=	op.exp.11	Protección de claves criptográficas
				op.ext	Servicios externos
categoría	n.a.	aplica	=	op.ext.1	Contratación y acuerdos de nivel de servicio
categoría	n.a.	aplica	=	op.ext.2	Gestión diaria
D	n.a.	n.a.	aplica	op.ext.9	Medios alternativos
				op.cont	Continuidad del servicio
D	n.a.	aplica	=	op.cont.1	Análisis de impacto
D	n.a.	n.a.	aplica	op.cont.2	Plan de continuidad
D	n.a.	n.a.	aplica	op.cont.3	Pruebas periódicas
				op.mon	Monitorización del sistema
categoría	n.a.	aplica	=	op.mon.1	Detección de intrusión
categoría	n.a.	n.a.	aplica	op.mon.2	Sistema de métricas

Figura 3: Aplicación de medidas de seguridad en relación al *Marco operacional* (de [9]).

ID	Medida	Aspectos	Estado	Comentarios
1.	**Política de seguridad**		Cumple	No existe política
	* Definición de "seguridad de la información"	Concisa y accesible	✓	
	* Responsables: información, servicio, seguridad	Roles y funciones	✓	
	* Roles	Mecanismos de coordinación	✓	Redactar el párrafo de nuevo para evitar ambigüedades
		Resolución de conflictos	✓	
2.	**Normativa de seguridad**		No cumple	No existe normativa
	* Posición ante problemas de seguridad (no contemplados en procedientos explícitos o implícitos)	Objetivos	✗	
		Usos correctos/incorrectos	✗	
		Escrita por expertos	✗	
	* Actuaciones a realizar	Realista, viable, concisa, clara	✗	
3.	**Procedimientos operativos de seguridad**		No cumple	No existen procedimientos definidos
	* Procedimiento, paso paso, de actuaciones	Responsable de la ejecución	✗	
	* Los procedimientos críticos deben estar documentados	Registro de actividad a realizar	✗	
		Identificación de situaciones anómalas	✗	
		Cubre >80% de las actividades rutinarias	✗	
		Actualización periódica	✗	
		Accesibilidad a los usuarios	✗	
		Informe de errores por parte de los usuarios	✗	
4.	**Proceso de autorización**		Parcial	No existe proceso de autorización
	* Acceso restringido al sistema de información (instalaciones, equipos, soportes y servicios)	Responsable	✓	
		Mecanismos de solicitud	✓	
		Proceso de estudio y aprobación	✳	No queda claro el diagrama
		Tipos de autorización	✗	
		Re-evaluación periódica	✗	

Figura 4: Plantilla de medidas de seguridad del *Marco organizativo*.

correspondiente: política de seguridad, normativa de seguridad, procedimientos operativos y proceso de autorización, describiéndose cada una de ellas de forma sucinta. Para ellas se deberá indicar el estado de cumplimiento existente a la

fecha del análisis: “Cumple” / “Parcial” / “No cumple”. Como complemento visual, cada uno de los tres niveles se representa mediante un color de fondo: verde para Cumple, naranja para Parcial, y rojo para No cumple.

Adicionalmente, para mejorar la granularidad y comprensión del alcance e impacto del estudio, el nivel de consecución de cada medida de seguridad se analiza en base a ciertos aspectos constitutivos. Cada uno de ellos se marcará individualmente con ✓, ∗, ✗ o en función de su consideración total, parcial o nula, respectivamente, en el despliegue de la medida. ¡Y con el mismo criterio de colores anterior! Adicionalmente, cada aspecto podrá ser comentado para una mayor aclaración en el sentido que sea oportuno.

De modo análogo, en la Figura 5 se presenta la plantilla de *checklist* para el estado de las medidas correspondientes al *Marco operacional*. Como para el *Marco organizativo*, cada una de las medidas contempladas se describe de forma breve en la plantilla para una mejor comprensión, adjuntándose su nivel de cumplimiento y los aspectos con ello relacionados.

Por su parte, dada la extensión del conjunto de medidas consideradas como *Medidas de protección* (40; véase Figura 1), la plantilla de informe correspondiente se obvía en el presente documento para su acceso a través del enlace https://nesg.ugr.es/ENS/checkmp.pdf.

Es de citar que, como se indicaba en la Figura 3, no todas las medidas de seguridad son de aplicación necesaria, sino que depende de la categoría del sistema de información. Así, a modo ilustrativo, en las tablas de estado aquí proporcionadas para el *Marco operacional* y *Medidas de protección* se marcan como *N/A* (“No Aplica”), y fondo gris, aquellas medidas que no se exigen para un sistema de categoría *BAJA*. La elaboración de los *checklists* hace que esta característica sea fácilmente automatizable en una aplicación y, en la práctica, para una empresa con un nivel medio-bajo de seguridad, es muy útil pues simplifica enormenente el proceso al descartar elementos no necesarios.

Como se ha mencionado con anterioridad, los respectivos *checklists* permiten generar de forma sencilla al tiempo que completa el informe de estado actual del sistema analizado, así como aclarar aquellos aspectos a desarrollar de cara a satisfacer los requisitos de seguridad correspondientes; esto es, las **actuaciones de mejora** subsiguientes.

IV. Conclusiones

El presente trabajo lleva a cabo una aproximación metodológica para la adecuación de sistemas al ENS, lo cual permitirá una mayor confianza social en las TIC.

Si bien es cierta la disponibilidad de herramientas que pueden utilizarse en este fin (*p.ej.*, *Pilar* o *Magerit*), ninguna de ellas es específica, al tiempo que deben ser usadas por expertos en el campo. Frente a estas opciones, nosotros proponemos un estudio sistemático basado en el simple empleo de tablas y *checklists*, que pueden ser usadas por personal no experto en el campo, de manera que se facilite (al menos en una primera instancia) el proceso de adecuación al ENS.

Con el *feedback* disponible hasta la fecha, podemos afirmar que la experiencia con empresas en este ámbito hasta el momento es positiva, pudiendo concluir sin ningún género de dudas que: *(a)* el (necesario) acercamiento y comprensión del ENS por parte de las empresas mejora, *(b)* la empresa tiene una visión general y organizada por categorías del estado de su nivel de cumplimiento del ENS, a la vez que le muestra los aspectos concretos que debe subsanar, y *(c)* la complejidad y coste asociados al proceso de adecuación con el uso de los *cheklists* se reducen de forma notable.

Como propuesta de futuro se prevé el desarrollo de las herramientas en formato *online*, de manera que se permita (por ejemplo a través de un servicio web) el avance en el estudio del estado de seguridad de un sistema al tiempo que la consulta de documentación y generación de sugerencias/recomendaciones al respecto para facilitar el proceso global.

Agradecimientos

Este trabajo ha sido financiado parcialmente por el Gobierno de España, con fondos FEDER, a través del proyecto TIN2017-83494-R.

Referencias

[1] Symantec: “Internet Security Threat Report 2017”. *Tech. Report*, 2017. Disponible en https://www.symantec.com/security-center/threat-report.

[2] Kaspersky: “Kaspersky Security Bulletin. Overall Statistics for 2017”. *Tech. Report*, 2017. Disponible en https://kasperskycontenthub.com/securelist/files/2017/12/KSB_statistics_2017_EN_final.pdf.

[3] Wikipedia: ”WannaCry Ransomware Attack”, 2017. Disponible en https://en.wikipedia.org/wiki/WannaCry_ransomware_attack.

[4] Xataka: “Strava pone de nuevo, sobre la mesa, el debate sobre los riesgos en los servicios de tracking y localización”. Disponible en https://www.xatakamovil.com/movil-y-sociedad/strava-pone-de-nuevo-sobre-la-mesa-el-debate-sobre-los-riesgos-en-los-servicios-de-tracking-y-localizacion.

[5] ONU: “Directrices para la Regulación de los Archivos de Datos Personales Informatizados”. Resolución 45/95 ONU, 1990. Disponible en http://www.agpd.es/portalwebAGPD/canaldocumentacion/legislacion/organismos_internacionales/naciones_unidas/common/pdfs/D.3BIS-cp--Directrices-de-Protecci-oo-n-de-Datos-de-la-ONU.pdf.

[6] LOPD: “Ley Orgánica de Protección de Datos de Carácter Personal”. RD 15/1999, BOE núm. 298 de 14 diciembre 1999. Disponible en https://www.boe.es/boe/dias/1999/12/14/pdfs/A43088-43099.pdf.

[7] UE-Datos: “Protección de las Personas Físicas en lo que respecta al Tratamiento de Datos Personales y a la Libre Circulación de estos Datos”. Directiva (UE) 2016/679, 27 abril 2016. Disponible en https://www.boe.es/doue/2016/119/L00001-00088.pdf.

[8] ENS: “Esquema Nacional de Seguridad en el Ámbito de la Administración Electrónica”. RD 3/2010, BOE núm. 25 de 29 enero 2010. Disponible en https://www.boe.es/boe/dias/2010/01/29/pdfs/BOE-A-2010-1330.pdf.

[9] ENS: “Modificación del RD 3/2010, por el que se regula el Esquema Nacional de Seguridad”. RD 951/2015, BOE núm. 264 de 4 noviembre 2015. Disponible en https://www.boe.es/boe/dias/2015/11/04/pdfs/BOE-A-2015-11881.pdf.

[10] UE-Sistemas: “Medidas Destinadas a Garantizar un Elevado Nivel Común de Seguridad de las Redes y Sistemas de Información en la Unión”. Directiva (UE) 2016/1148, 6 junio 2016. Disponible en https://www.boe.es/doue/2016/194/L00001-00030.pdf.

[11] D. Dor, Y. Elovici, “A Model of the Information Security Investment Decision-Making Process”. *Computers & Security*, Vol. 63, pp. 1–13, 2016.

[12] E. Weishäupl, E. Yasasin, “Information Security Investments: An Exploratory Multiple Case Study on Decision-Making, Evaluation and Learning”. *Computers & Security*, En prensa, 2018.

[13] CCN-806, “Guía de Seguridad (CCN-STIC-806. Esquema Nacional de Seguridad. Plan de Adecuación”. Enero 2011. Disponible en https://www.ccn-cert.cni.es/publico/seriesCCN-STIC/series/800-Esquema_Nacional_de_Seguridad/806-Plan_adecuacion_ENS/806_ENS-adecuacion_ene-11.pdf.

[14] CCN-804, “Guía de Seguridad de las TIC (CCN-STIC 804. ENS. Guía de Implantación”. Junio 2017. Disponible en https://www.ccn-cert.cni.es/publico/seriesCCN-STIC/series/800-Esquema_Nacional_de_Seguridad/804-Medidas_de_implantacion_del_ENS/804_medidas_de_implantacion_del_ens.pdf.

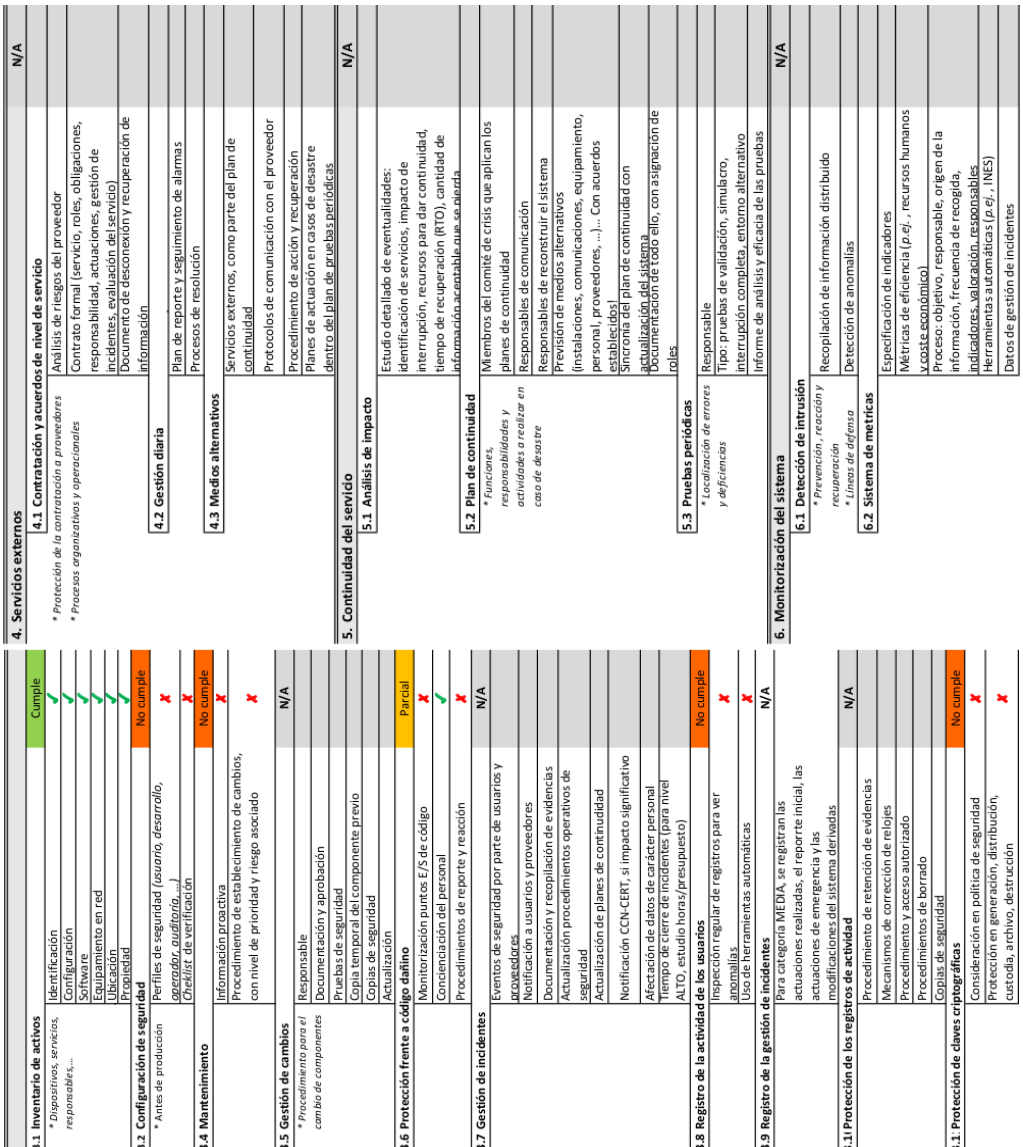


Figura 5: Plantilla de medidas de seguridad del *Marco operacional*.

Esquemas criptográficos post-cuánticos aplicados a la e-democracia: LWE-Helios Voting

David Yeregui Marcos del Blanco
Universidad de León
Campus de Vegazana, 24071 León
dmarcb01@estudiantes.unileon.es

Luis Panizo Alonso
Universidad de León
Campus de Vegazana, 24071 León
luis.panizo@unileon.es

José Ángel Hermida Alonso
Universidad de León
Campus de Vegazana, 24071 León
jahera@unileon.es

***Resumen*—La implantación de sistemas de voto electrónico remoto o voto por internet a procesos electorales se está produciendo a un ritmo menor al esperado. Uno de los factores que explican dicho fenómeno son los intentos de ataques derivados de las debilidades en varios de los protocolos existentes. En el presente trabajo se analizan los últimos avances en la aplicación de esquemas post-cuánticos en la materia, haciendo un especial énfasis en el sistema LWE-Helios. Con ello se pretende clarificar el estado actual del arte de cara a una introducción segura de esta nueva familia de esquemas.**

***Index Terms*—e-democracia, i-voting, criptografía post-cuántica**

I. Introducción

Las tecnologías de la información y la comunicación (TIC) han tenido un gran impacto en la vida de miles de millones de personas en ámbitos tales como la sanidad, las finanzas, la educación o el turismo. A principios del siglo XXI se preveía que los procesos electorales también se beneficiasen de la revolución tecnológica en lo que vino a denominarse e-democracia.

Han pasado más de 15 años desde entonces y el modo de ejercer el derecho a voto no ha cambiado de una manera sustancial. Es cierto que actualmente se puede utilizar el voto por internet o i-voting como canal adicional para determinadas elecciones y segmentos de población en países como Suiza, Australia, Canadá o Estonia. Pero es también cierto que muchos otros países han decidido cancelar su implantación tras un período de pruebas piloto. Tal es el caso de los Estados Unidos, Reino Unido, Alemania, Holanda, Noruega, España o Nueva Zelanda.

El voto por internet presenta unas características que lo convierte en una implementación de las TIC especialmente exigente [1]:

- La necesidad de preservar simultáneamente integridad y privacidad [2].
- El gran poder y acceso a recursos que obtiene el vencedor, atrayendo a atacantes con enormes medios.
- Los comicios se producen en una fecha definida, pública y comunicada con suficiente anterioridad, lo que facilita la preparación de ataques elaborados y coordinados a gran escala.
- Resultados difíciles de revertir en el caso de detectar el ataque una vez han concluido los comicios.

Adicionalmente, existen tres vectores de ataque que han sido utilizados repetidamente:

- El dispositivo del votante. Se estima que un 30-40 % de ellos están infectados por malware [2].
- La red, cuyos protocolos criptográficos han sido objeto de numerosos ataques [3,4].
- El propio sistema de i-voting, que en numerosas ocasiones presenta vulnerabilidades solo descubiertas cuando han sido explotadas durante un piloto real [5,6]

En las últimas elecciones generales de los Estados Unidos, la Oficina del Director de Inteligencia Nacional apuntó que "Putin y el gobierno ruso aspiraron a mejorar las posibilidades de Trump, tratando de desacreditar a la Secretaria Clinton. Tenemos una gran confianza en esta afirmación". El documento en el que se incluyen estas declaraciones es de acceso público *"Assessing Russian Activities and Intentions in Recent US Elections"* [7].

Para tratar de mejorar la seguridad de los actuales sistemas de i-voting, en fechas recientes se han presentado metodologías prácticas de evaluación basadas en los requisitos legales y técnicos del Consejo de Europa con el fin de establecer rangos seguros de utilización [8, 9, 10].

A pesar de ello, la implantación de soluciones de i-voting ha ido perdiendo intensidad, en parte por la ausencia de esquemas criptográficos seguros a largo plazo. Por suerte, en fechas muy recientes han ido apareciendo las primeras propuestas post-cuánticas en el campo. En el presente artículo se analiza una de las más prometedoras: LWE-Helios [11].

II. Definiciones

Debido a la restricción de longitud del presente artículo, se van a ofrecer únicamente las definiciones directamente relacionadas con el mismo.

En primer lugar y de manera formal, se define indistintamente voto por **internet, i-voting o voto electrónico remoto (VER)** como: "Aquel sistema de votación que se produce en un entorno remoto y no controlado, por medios electrónicos y en el que el voto es enviado total o parcialmente a través de una conexión a internet desde un ordenador personal o dispositivo móvil que no es una máquina construida al efecto de votar".

Por lo que respecta a los **requisitos de seguridad del i-voting**, estos son la **Verificabilidad Extremo a Extremo o E2Ev** [12] y la **privacidad** [13]. Para profundizar, en [9] se detalla la correlación entre la E2Ev, la privacidad y los cinco requisitos legales de una elección democrática según el Consejo de Europa (universal, igual, libre directa y secreta).

- **Verificabilidad Extremo a Extremo (E2Ev)**: No existe una definición formal y universal de la E2Ev debido a la ausencia de pruebas automatizadas. En el caso del i-voting, los operadores conmutativos y asociativos están fuera del alcance de herramientas de análisis simbólico tales como ProVerif [14]. No se pueden analizar pues propiedades homomórficas del tipo [15]:

$$enc\,(pk; v_1) * enc\,(pk; v_2) = enc\,(pk; v_1 + v_2) \qquad (1)$$

Por tanto, el análisis debe realizarse caso por caso de acuerdo a la definición más ampliamente aceptada [12] y que incluye las siguientes 3 propiedades:

  - Emitido según lo previsto: El votante puede obtener una prueba convincente de que su voto cifrado refleja fielmente su sentido de voto.
  - Guardado como se ha emitido: El votante puede comprobar que su voto cifrado se ha incluido correctamente en un listado público de votos cifrados o Bulletin Board (BB).
  - Recontado como se ha guardado: Cualquier parte puede comprobar que todos los votos cifrados publicados son correctos y se han incluido en el recuento si bien, no hay forma de saber el voto de ningún votante.

Las dos primeras se denominan *verificabilidad individual* y la última *verificabilidad universal*.

- **Privacidad** Dividida en tres niveles crecientes:

  - Privacidad del voto: El voto de un votante no es revelado a ninguna parte.
  - Ausencia de recibo: El votante no puede obtener ninguna información/recibo que pueda probar el sentido de su voto.
  - Resistencia a la coerción: El votante no puede demostrar a un coercionador cómo voto, incluso si ambos estuviesen colaborando activa y voluntariamente.

A principios de los 2000, Hirt and Sako [16] fueron los primeros en probar que la ausencia de recibo no es suficiente y que por tanto el nivel exigible al i-voting es el más alto, la resistencia a la coerción.

En la actualidad, ningún sistema de i-voting es resistente a la coerción [17, 18] de acuerdo a la definición original en [13]. Es necesario abordar este problema de cara a una mayor implantación de herramientas de VER. Una de las líneas de investigación más prometedoras en el campo viene de la mano de los **esquemas post-cuánticos** basados en **Cifrado Totalmente Homomórfico (CTH)** y **retículos**.

II-A. Cifrado Totalmente Homomórfico (CTH) y criptografía basada en retículos (lattices)

Peter Shor fue el primero en demostrar en 1994 que la seguridad de los sistemas criptográficos tradicionales se podría comprometer en tiempo polinomial con un ordenador cuántico; en concreto para el caso de RSA [19].

En realidad, el CTH fue definido por primera vez en 1978 sin saber si era resoluble o no [20], si bien no fue hasta 2010 cando Gentry presentó el primer esquema CTH utilizando criptografía basada en retículos [21].

El CTH permite computaciones arbitrarias sobre datos cifrados: si el usuario tiene una función f y quiere obtener $f(m_1, ..., m_n)$ para los inputs $m_1, ..., m_n$, es posible computar sobre cifrados de dichos inputs, $c_1, ..., c_n$ obteniendo un resultado que descifra $f(m_1, ..., m_n)$.

En el CTH, cada texto sin cifrar se enmascara con aleatoriedad interna y externa, denominándose a la primera ruido. A continuación, se pueden realizar adiciones y multiplicaciones, si bien en ocasiones debe introducirse también un operador NAND. Actualmente y pese a los avances, incluso las implementaciones más eficientes requieren unos 4 minutos para la evaluación homomórfica de una operación de cifrado AES-128, a una media de 2 segundos por bloque [22].

Operativamente, cada operación que se introduzca aumenta el nivel de ruido. Cuando éste llega al mismo tamaño que la aleatoriedad externa, el texto cifrado deja de ser descifrable. Las multiplicaciones añaden mucho más ruido por lo que deben ser utilizadas con cautela.

En cuanto a los **retículos** (***lattice*** en inglés), formalmente se refiere a un conjunto de puntos en un espacio de dimensión n con una estructura periódica. Un retículo de dimensión n denominado L es cualquier subconjunto de $\mathbb{R}^n$ que es:

- un **subgrupo aditivo**: $0 \in L$ y $-x, x + y \in L$ para todo $x, y \in L$; y
- **discreto**: todo $x \in L$ tiene un correspondiente en $\mathbb{R}^n$ en el cual x es el único punto del retículo.

Dados k vectores linearmente independientes $b_1, ...b_k \in \mathbb{R}^n$, el retículo de rango k generado por ellos es el siguiente conjunto de vectores [21]:

$$L(b_{1,...,b_n}) = \left\{ \sum_{i=1}^{k} z_i b_i : z_i \in \mathbb{Z} \right\} = \{Bz : z \in \mathbb{Z}^n\} \qquad (2)$$

La criptografía de retículos ha atraído importante atención académica desde que Regev [23] descubrió una reducción cuántica del problema del vector más corto (*Shortest Vector Problem* o SVP).

Para el caso concreto del presente artículo, presentaremos en más detalle el problema de aprendizaje con errores

(Learning with Errors o LWE), que se considera no resoluble en tiempo computacional, incluso para ordenadores cuánticos de gran capacidad. El motivo es que se trata de la base criptográfica sobre la que se construye el esquema LWE-Helios, motivo de estudio de este trabajo.

Problema del aprendizaje con errores o Learning with errors problem (LWE): Introducido por Regev [23] y mejorado por Lyubashevsky et al. para obtener variantes de anillos [24] y por Ducas et al. para obtener cifrado homomórfico [25]:

Sea $\alpha \in \mathbb{R}^+$ un parámetro de ruido, $(s_1, ..., s_n)$ un secreto binario uniformemente distribuido en $\mathbb{B}^n$, y $G \subseteq \mathbb{T}^n$ un grupo finito discretizado suficientemente denso. LWE (s, α, G) es la instancia invariante de escala.

Un ejemplo aleatorio de LWE para un mensaje $\mu \in \mathbb{T}$ se define como un elemento $(a, b) \in G \times \mathbb{T}$ donde $a = (a_1, ..., a_n) \in G \subset \mathbb{T}^n$ es un ejemplo uniforme de G, y b es igual a $\sum_{i=1}^{n} s_i a_i + \mu + e \in \mathbb{T}$, donde e es estadísticamente cercano a un ejemplo Gaussiano continuo y centrado en cero $\mathbb{T}$ sobre α.

La fase de un ejemplo de LWE $(a, b) \in \mathbb{T}^n \times \mathbb{T}$ se define como:

$$\varphi_s((a, b)) = b - \sum_{i=1}^{n} s_i a_i \in \mathbb{T} \qquad (3)$$

La seguridad de LWE se basa en los otros dos parámetros: el parámetro de error Gaussiano α y el número n de bits correspondientes a la entropía en el secreto. De acuerdo a las estimaciones de reducción de retículos estándar [26], LWE es seguro 128-bits para α igual a 2^{-10}, 2^{-30} o 2^{-50} si $n \geq$ 300, 800 y 1500 respectivamente. Además, para cualquier α and $n = \Omega(log(\frac{1}{\alpha}))$, LWE se beneficia asintóticamente del caso más desfavorable de reducción media, dependiendo de la forma de G [27].

Para profundizar en cualquiera de las definiciones y problemas sobre retículos, se recomiendan las referencias bibliográficas [23-27].

III. LWE-Helios

Los criptosistemas basados en retículos son un área de investigación reciente, por lo que existen muy pocos esquemas en el campo del i-voting. No obstante, en 2016 Chillotti et al. propusieron un protocolo denominado LWE-Helios [11], basado en una de las herramientas más asentadas en el campo del voto por internet: Helios Voting [28].

III-A. Protocolo de i-voting LWE-Helios

Existen multitud de diferencias entre el esquema original de Helios Voting y la versión post-cuántica LWE-Helios que se irán explicando en el presente apartado III.

En primer lugar, respecto a las fases del protocolo de votación, son muy similares, si bien las definiciones difieren ampliamente. En el caso de LWE-Helios son las siguientes:

Setup: El responsable del tablón o *Bulletin Board* (BB) genera un par de claves $(pk_{BB}, sk_{BB}) = KeyGenE_{BB}(1^\lambda)$ y publica pk_{BB}. Los administradores reinician el esquema de LWE concatenados y cada uno genera su propia clave secreta LWE $s_i \in \mathbb{B}^n$, *master trapdoor* R_i y una clave pública correspondiente $pk_i \in (G \times \mathbb{T})^m$. También suministran 3 claves *bootstrapping* $BK_1 := BK_{\left[(s^{(f)}, \frac{1}{4}) \to (s^{(m)}, \frac{1}{4})\right]}$, $BK_2 := BK_{\left[(s^{(m)}, \frac{1}{4}) \to (s^{(m)}, \frac{1}{16})\right]}$ y una mayor para amplitud de ruido baja $BK_3 := BK_{\left[(s^{(m)}, \frac{1}{4}) \to (s, \frac{1}{L^{3/2}})\right]}$. El *bootstrapping* se puede realizar con BK_1 y BK_2 en menos de 700ms [25]. Para BK_3 se produce una ralentización con un factor constante de ≈ 16.

Registro del votante: $Register(1^\lambda, id)$. La autoridad ejecuta $(upk_{id}, usk_{id}) \leftarrow KeyGenS(1^\lambda)$, añade upk_{id} en L_U y devuelve (upk_{id}, usk_{id}).

Votación: $Vote(pk, usk, upk, v)$: Cada votante computa la descomposición binaria $(v_0, ..., v_{k-1}) \in \{0, 1\}^k$, de tal forma que $v = \sum_{j=0}^{k-1} v_j 2^j$. Cada bit se cifra como $c_j = LWEPubEncrypt_{pk^{(f)}}(\hat{v}_j)$. c'_j se define como $c'_j = Bootstrap_{BK_1}(c_j, \frac{1}{2}, 0)$ y devuelve el voto final $b = (content, \sigma)$, tal que $content = (aux, upk, (c'_0, ..., c'_{k-1}), num)$ y $\sigma = Sign(usk, content)$.

Procesado del voto: $ProcessBB(BB, b, sk_{BB})$: Tras recibir el voto b, se analiza como $(content, \sigma)$, donde $content = (aux, upk, (c'_0, ..., c'_{k-1}), num)$. El BB verifica que $upk \in L_U$ y comprueba si $VerifyS(upk_{id}, content)$. A diferencia del Helios tradicional, no hace falta una prueba de conocimiento cero o ZKP puesto que todo los mensajes binarios son respuestas válidas.

Posteriormente, el BB aplica una secuencia de operaciones homomórficas públicas que no requieren la participación del votante:

1. Pre-Bootstrapping:

$$Bootstrap_{BK_2}(c'_j, \frac{1}{4}, 0) \qquad (4)$$

se aplica a cada c'_j.

2. Expansión binaria homomórfica:

$$HomAND(c_1, c_2) = Bootstrap_{BK_2}((0, -\tfrac{1}{8}) + c_1 + c_2, \tfrac{1}{4}, 0) \qquad (5)$$

$$HomANDNot(c_1, c_2) = Bootstrap_{BK_2}((0, \tfrac{1}{8}) + c_1 - c_2, \tfrac{1}{4}, 0) \qquad (6)$$

3. Bootstrapping general:

$$Bootstrap_{BK_2}((0, -\frac{1}{8}) + c_1 + c_2, \frac{1}{4}, 0) \qquad (7)$$

4. Adición homomórfica:El BB añade homomórficamente todos los textos cifrados sin bootstrapping. Se obtiene el texto cifrado LWE definitivo $(C_0, ..., C_{l-1})$ de $(\frac{n_0}{L}, ..., \frac{n_{l-1}}{L})$ con ruido $\Omega(L^{-1})$.

Recuento y verificación:

1. $Tally(BB, sk = (sk_1, ..., sk_t))$: para cada C_j, los administradores realizan el cifrado distribuido como se ha definido anteriormente y se publica un texto cifrado *trapdoor* $\Pi_{i,j} \in \mathbb{Z}^m$, que se revela a cualquier parte.
2. $VerifyTally(BB, (\Pi_1, ..., \Pi_t))$. Si una *trapdoor* $\Pi_{i,j}$ no es válida, ello prueba que el administrador i-ésimo no es honesto y por tanto $VerifyTally$ devuelve $\perp$. Si todas las *trapdoors* son válidas, cualquiera puede usar $(\Pi_{1,j}, ..., \Pi_{t,j})$ para descifrar C_j y recuperar n_j para todo j, revelando por tanto el número de votos al candidato j . $VerifyTally$ devuelve finalmente el resultado $(n_o, ..., n_{l-1})$.

III-B. Propiedades y análisis

Comparado con el sistema Helios Voting tradicional, LWE-Helios presenta las siguientes diferencias:

- Diseñado y construido sobre primitivas post-cuánticas: firmas basadas en retículos, cifrado homomórfico basado en LWE y trapdoors para retículos.
- No implementa pruebas de conocimiento cero (ZKP). El en protocolo original se incluían dos: una para que el votante pueda demostar que el voto es válido y otra para que los delegados puedan descifrar el resultado final.
- Para garantizar la privacidad incluso en el caso de que algunas autoridades fuesen corruptas, se utilizan esquemas LWE concatenados en lugar del protocolo Fiat-Shamir de reparto de secretos [29].

Por lo que se refiere a la **privacidad**, se considera que LWE-Helios $Vote$ cumple con la propiedad si existe un simulador eficiente tal que, para cualquier adversario en tiempo polinomial probabilístico (PPT) A, se cumple lo siguiente [11]:

$$\left| Pr\left[Exp^{bpriv,\beta}_{A,Vote}(\lambda) = \beta \right] - \frac{1}{2} \right| \qquad (8)$$

En cuanto a la **verificabilidad**, el protocolo $Vote$ se considera verificable extremo a extremo si para cualquier adversario A :

$$Succ^{ver} = \left[Exp^{ver}_{A,Vote}(\lambda) = 1 \right] \qquad (9)$$

es despreciable en λ.

Finalmente, respecto a la **corrección**, asumiendo que el esquema de firma S es infalsificable, que ε es un esquema de cifrado no maleable, y que las operaciones homomórficas públicas realizadas por el BB son correctas, LWE-Helios es correcto.

IV. Conclusiones

En los últimos años, el estudio e implementación de sistemas de i-voting novedosos que permitiesen abordar los problemas de seguridad pendientes a largo plazo ha sido escasa.

Afortunadamente, en fechas recientes han aparecido las primeras aplicaciones de herramientas criptográficas post-cuánticas en la materia. Una de las más destacadas ha sido la actualización del sistema de i-voting Helios[28] mediante la introducción de criptografía de retículos, en su variante LWE.

Al esquema se le ha denominado LWE-Helios [11] y conforma el iniciativa más completa en el campo hasta la fecha. Entre los logros alcanzados destacan:

- La aplicación exitosa de primitivas post-cuánticas a un protocolo de i-voting: firmas infalsificables basadas en retículos, cifrado homomórfico basado en LWE y *trapdoors* para retículos.
- Lograr verificabilidad sin utilización de pruebas de conocimiento cero. Los autores lo consiguieron implementando cifrado totalmente homomórfico (CTH) [25] y firmas de retículos basadas en *trapdoors* [30].
- El protocolo es correcto, verificable y cumple con la privacidad, según [13,15].

Es importante destacar no obstante que LWE-Helios se encuentra en una fase de desarrollo teórico, sin haber llegado a una implementación en código. Aún así, se trata de un excelente punto de partida para un desarrollo completo posterior.

Adicionalmente, la versión actual de LWE-Helios deja los siguiente problemas abiertos:

- La definición formal y posterior adaptación de la *corrección fuerte* contra un tablón y/o autoridades de registro corruptas. Es decir, los autores asumen que el BB no es completamente deshonesto o corrupto. Dependiendo de la tipología de las elecciones y del país donde se celebren, dicho supuesto podría ser demasiado optimista.
- La prueba de privacidad se basa en una premisa exigente: la función aleatoria de *bootstrapping* se modeliza como un oráculo aleatorio perfecto. El problema equivalente en el modelo estándar todavía no se ha resuelto.

Por todo ello, LWE-Helios supone un prometedor primer paso hacia un nuevo paradigma de protocolos de i-voting llamados a aumentar la seguridad y apuntalar el uso de herramientas TIC en procesos electorales y en general a la e-democracia.

Referencias

[1] Springall D., Finkenauer T., Durumeric Z., Kitcat J., Hursti H., MacAlpine M., et al. Security Analysis of the Estonian Internet Voting System, Proc 21st ACM Conf Comput Commun Secur., 2014, 703-715. doi:10.1145/2660267.2660315

[2] Foundation USV. The Future of Voting. In: The Future of Voting [Internet]. Available: https://www.usvotefoundation.org/e2e-viv/summary , 2015
[3] The FREAK Attack. In: The FREAK Attack [Internet]. Available: https://censys.io/blog/freak , 2015
[4] Adrian D., Bhargavan K., Durumeric Z., Gaudry P., Green M., Halderman JA., et al. Imperfect Forward Secrecy: How Diffie-Hellman Fails in Practice. Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security, New York, NY, USA, 2015, 5-17 doi:10.1145/2810103.2813707
[5] Goldwasser S., Tauman Y., On the (In)security of the Fiat-Shamir Paradigm. Focs, 2003
[6] Achenbach D., Kempka C., Lowe B., Muller-Quade J., Improved Coercion-Resistant Electronic Elections through Deniable Re-Voting. USENIX J Elect Technol Syst., 2015, 26-45
[7] Background to Assessing Russian Activities and Intentions in Recent US Elections: The Analytic Process and Cyber Incident Attribution, 2017
[8] Neumann SR., Ph.D. Thesis Evaluation and Improvement of Internet Voting Schemes Based on Legally-Founded Security Requirements, Technische Universität Darmstadt, Germany, 2016
[9] Marcos del Blanco DY., Panizo Alonso L., and Hermida Alonso JA., The need for Harmonization in the online voting field: Towards an European Standard for edemocracy. The International Conference on Electronic Voting, E-Vote-ID 2016, October 18-21, Bregenz, Austria, 2016, 339-343
[10] T. Standards, "Guidelines on the implementation of the provisions of Recommendation CM/Rec(2017)5 on standards for e-voting," 1289 th Meet. , 14 June 2017 2 . 3 Ad hoc Comm. Expert. Leg. , Oper. Tech. Stand. e- voting (CAHVE), no. June, pp. 1–19, 2017
[11] Chillotti I., Gama N., Georgieva M., Izabachene M. An Homomorphic LWE based E-Voting Scheme In: 7th International Workshop, PQCrypto 2016, February 24-26, 2016, Fukuoka, Japan, 2016, 245-265
[12] Benaloh JDC., Rivest R., Ryan PYA., Stark P., Teague V., and Vora P., End-to-end verifiability arXiv e-prints, 2014
[13] Juels A., Catalano D. and Jakobsson M. Coercion-resistant electronic elections, In: Lecture Notes in Computer Science vol. LNCS 6000, 2010 37-63
[14] Blanchet B., An Automatic Security Protocol Verifier based on Resolution Theorem Proving (invited tutorial) In: Nieuwenhuis R. (Ed.) 20th International Conference on Automated Deduction, July 22-27, Tallinn, Estonia, 2005, 3-51
[15] Cortier V., Formal Verification of e-Voting: Solutions and Challenges, ACM SIGLOG News, vol. 2, 1, 2015, 25-34
[16] Hirt M., Sako K. Efficient Receipt-Free voting based on homomorphic encryption In: Preneel B., editor, EUROCRYPT'00, vol. 1807 LNCS Bruges, Belgium, 2000, 539-556
[17] Heiberg S., Parsovs A., Willemson J. Log Analysis of Estonian Internet Voting 2013 - 2015. Smartmatic – Cybernetica Centre of Excellence for Internet Voting, Software Technology and Applications Competence Centre, Tartu University, 2015
[18] Nore H., Implementing E-Voting in Norwegian Elections New Voting Technology Consulting AS, 2015
[19] Shor P. Algorithms for quantum computation: discrete logarithms and factoring. In: 35th Annual Symposium on Foundations of Computer Science November 20-22 Santa Fe, New Mexico, USA, 1994, 124-134
[20] Rivest R., Adleman L., and Dertouzos M. On data banks and privacy homomorphisms. Foundations of Secure Computation, Academia Press, 1978, 169-180
[21] Gentry C., Fully homomorphic encryption using ideal lattices., In: Proceedings of the forty-first annual ACM symposium on Theory of computing May 31 - June 02 2009, Bethesda, MD, USA, 2009, 169-178
[22] Gentry C., Halevi S., and Smart NP. Homomorphic Evaluation of the AES Circuit. In: Advances in Cryptology - CRYPTO 2012 - 32nd Annual Cryptology Conference, August 19-23, Santa Barbara, CA, USA, 2012, 850-867
[23] Regev O. On lattices, learning with errors, random linear codes, and cryptography. In: H N Gabow and R Fagin, (Eds.) Proceedings of the 37th Annual ACM Symposium on Theory of Computing, May 21-24, Maryland, USA,2005, 84-93
[24] Lyubashevsky V., Peikert C., and Regev O. On ideal lattices and learning with errors over rings. In: Proc. of EUROCRYPT, vol. 6110 of LNCS May 30 - June 3 Monaco and Nice, Monaco, France,2010, 1-23
[25] Ducas L. and Micciancio D. FHEW: Bootstrapping homomorphic encryption in less than a second. In: Eurocrypt 2015, April 26-30 Sofia, Bulgaria, 2015, 617-640
[26] Chen Y. and Nguyen PQ. BKZ 2.0: Better lattice security estimates. In: Wang X. and Lee DH. (Eds.) Asiacrypt 2011, Seoul (Korea,2011) 1-20
[27] Gama N., Izabachene M., Nguyen PQ., and Xie X. Structural lattice reduction: Generalized worst-case to average-case reductions. In: EUROCRYPT 2016, 2016, 528-558
[28] Adida B., Helios: Web-based Open-audit Voting, In: Proceedings of the 17th Conference on Security Symposium, July 28 - August 1 2008, San Jose, CA, USA, 2008, 335-348
[29] Fiat A., Shamir A., How to prove yourself: Practical solutions to identification and signature schemes, In: Advances in Cryptology Crypto'86 Springer-Verlag, 1986, 186-194
[30] Micciancio D. and Peikert C. Trapdoors for lattices: Simpler, tighter, faster, smaller. In: Pointcheval D. and Johansson T. (Eds.) Eurocrypt 2012, April 15-19, Cambridge, UK, 2012, 700-718

Notificaciones Certificadas sobre Blockchain

Macià Mut-Puigserver
Dpt. de C. Matemàtiques i Informàtica
Universitat de les Illes Balears,
Ctra. Valldemossa, km7,5 07122 Palma
macia.mut@uib.es

M. Magdalena Payeras-Capellà
Dpt. de C. Matemàtiques i Informàtica
Universitat de les Illes Balears,
Ctra. Valldemossa, km7,5 07122 Palma
mpayeras@uib.es

Miquel A. Cabot-Nadal
Dpt. de C. Matemàtiques i Informàtica
Universitat de les Illes Balears,
Ctra. Valldemossa, km7,5 07122 Palma
miquel.cabot@uib.es

***Resumen*—Muchas aplicaciones tradicionales se pueden redefinir gracias a los beneficios de la tecnología Blockchain. Uno de estos servicios es la provisión de notificaciones certificadas, una de las aplicaciones que requieren un intercambio equitativo de valores: un mensaje y una prueba de no repudio de origen a cambio de una evidencia de recepción. Este artículo presenta el primer sistema basado en blockchain para notificaciones certificadas. Proponemos dos soluciones que permiten enviar notificaciones certificadas cuando se requiere confidencialidad y cuando es necesario registrar el contenido de la notificación, respectivamente. Primero presentamos un protocolo para notificaciones certificadas no confidenciales que satisface las propiedades de equidad fuerte y transferibilidad de las pruebas gracias al uso de un smart contract y sin la necesidad de una tercera parte de confianza (TTP). A continuación presentamos un DApp para notificaciones certificadas confidenciales con un smart contract que permite un intercambio optimista de valores asíncrono y con una tercera parte de confianza que no registra información de estado.**

I. Introducción

La tecnología Blockchain proporciona un sistema inalterable de registro de datos que permite nuevas soluciones para una amplia gama de aplicaciones tradicionales. Uno de los servicios que puede beneficiarse de las características distintivas de blockchain es la provisión de notificaciones certificadas, es decir, un servicio que permite a un remitente demostrar que ha enviado un mensaje a un receptor o conjunto de receptores. Por lo tanto, los servicios de notificación certificada proporcionan evidencia de que un receptor tiene acceso a un mensaje desde una fecha/hora específica.

Las notificaciones certificadas, junto con otros servicios electrónicos, como la firma electrónica de contratos, la compra electrónica (pago a cambio de un recibo o producto digital) o correo certificado, requieren un intercambio equitativo de elementos entre dos o más usuarios. Con el fin de crear protocolos que permitan llevar a cabo estos intercambios y, al mismo tiempo, mantener la seguridad de las comunicaciones, existen soluciones que caen dentro del patrón genérico llamado *intercambio equitativo de valores*. Un intercambio equitativo siempre proporciona un trato igual a todos los usuarios, y, al final de cada ejecución, cada parte tiene el elemento que desea obtener de la otra parte, o el intercambio no se ha llevado a cabo con éxito para nadie (ninguna parte ha recibido el elemento esperado). En un caso típico de notificación, el elemento a intercambiar es el mensaje junto con las pruebas de no repudio de origen y recepción.

Los protocolos de intercambio equitativo propuestos hasta ahora generalmente usan TTPs [10], [11], [17], que son responsables de resolver cualquier conflicto que surja como resultado de intercambios interrumpidos o intentos de fraude. Además, estos protocolos normalmente usan mecanismos de no repudio para generar evidencia que demuestre el comportamiento de los actores del protocolo. Sin embargo, es difícil tener TTPs que sean realmente confiables para cualquier usuario de la red y que tengan un marco de acción definido. Además, las TTP también pueden causar problemas a nivel técnico (Ej: causar cuellos de botella en las comunicaciones), falta de eficacia en los protocolos (Ej: ralentizar la resolución de problemas) y aumentar el costo de la ejecución del protocolo (Ej: cobrar altas tarifas por la provisión de servicios).

Actualmente, con la llegada de la tecnología blockchain y los smart contracts, las TTP pueden ser reemplazadas o complementadas por este nuevo know-how, que abre una gama de nuevas posibilidades para encontrar soluciones efectivas a las versiones electrónicas de los protocolos que cumplen con el patrón genérico de intercambio equitativo de valores. Además, la blockchain Ethereum y su criptomoneda Ether ofrece un conjunto de funcionalidades aún más rico que las criptomonedas convencionales como bitcoin, ya que admiten smart contracts en un sistema completamente distribuido. Esta nueva tecnología nos permite definir transacciones con reglas predeterminadas (escritas en un contrato) en una lógica programable que puede garantizar un intercambio equitativo entre las partes con una desconfianza mutua inicial. El poder de la blockchain se potencia aún más por el hecho de que incorpora naturalmente una noción discreta de tiempo. La existencia de un reloj/registro confiable es crucial para lograr la propiedad de equidad en los protocolos. Esta característica puede hacer que el modelo de criptografía en la cadena de bloques sea aún más poderoso que el modelo tradicional sin una cadena de bloques donde la equidad es muy difícil de garantizar sin la intervención de una TTP.

Utilizando Blockchain podemos reducir o incluso eliminar el rol de las TTP dentro de los protocolos de intercambio equitativo. En este trabajo se presentan dos protocolos, uno para notificaciones que no requieren confidencialidad y otro para notificaciones confidenciales. Estas dos propuestas tienen un conjunto muy diferente de propiedades, así como diferentes necesidades relacionadas con la participación de la TTP. Por lo que sabemos, no hay trabajos anteriores que se ocupen de notificaciones certificadas con blockchain utilizando smart contracts. Los estudios previos sobre equidad utilizando blockchain solo se enfocan en operaciones de compra de un producto (o recibo) a cambio de criptomonedas (generalmente bitcoin) [2], [8], [9]. Hasta donde tenemos conocimiento, solo hay una propuesta especialmente adaptada para blockchain [13], que admite smart contracts con un lenguaje completo de Turing como Ethereum. [13] usa, por primera vez, un contrato inteligente para la resolución de una operación de compra.

II. Propiedades ideales de las notificaciones certificadas.

Algunos requisitos para el intercambio equitativo se establecieron en [1], y se volvieron a formular en [18]:

1. **Efectividad.** Si dos partes se comportan correctamente, recibirán los elementos esperados sin la participación de un tercero.
2. **Equidad.** Después de completar una ejecución del protocolo, cada parte recibe el elemento esperado o ninguna de las partes recibe información útil sobre el elemento de la otra parte. La equidad es *débil* si, al final de la ejecución, ambas partes han recibido los elementos esperados o si una entidad recibe el elemento esperado y otra entidad no, este último puede obtener evidencia de esta situación.
3. **Timeliness o Asincronía** En cualquier momento, durante una ejecución de protocolo, cada parte puede elegir unilateralmente terminar el protocolo sin perder equidad.
4. **No repudio.** Si un artículo ha sido enviado desde A a B, A no puede negar el origen del elemento y B no puede negar la recepción del elemento.
5. **Verificabilidad de la TTP.** Si la TTP actúa incorrectamente y provoca la pérdida de equidad para una parte, la víctima puede probar este hecho.

Podemos añadir, a este conjunto de propiedades, algunas otras propiedades interesantes para el caso de notificaciones certificadas.

6. **Confidencialidad**. Solo el emisor y el receptor de la notificación conocen el contenido del mensaje certificado.
7. **Eficiencia.** Un protocolo eficiente usa la cantidad mínima de pasos que permiten el intercambio efectivo o el costo mínimo.
8. **Transferibilidad de evidencias.** Las pruebas generadas por el sistema pueden transferirse a entidades externas para probar el resultado del intercambio.
9. **State Storage.** Si la TTP que puede estar involucrada en el intercambio no necesita mantener información de estado, entonces el sistema es *stateless*.

Algunas de las propiedades anteriores no se pueden lograr en el mismo protocolo. Los autores de [7] enumeran las incompatibilidades entre las características ideales. Algunos ejemplos son Equidad débil y Transferibilidad de evidencias, TTP sin estado y Asincronía y Verificabilidad de la TTP y Transparencia.

III. Estado del arte de las notificaciones certificadas

Las notificaciones certificadas siguen el patrón de intercambio equitativo de valores. Este tipo de intercambio no tiene una solución definitiva y estandarizada en su versión electrónica. Las notificaciones pueden hacerse mediante correo electrónico y hasta la fecha se han presentado varias propuestas para este servicio. Sin embargo, no se requiere que las notificaciones certificadas usen correo electrónico, como se discutirá en este documento. Una notificación certificada incluye un intercambio de elementos entre el emisor y el receptor; el emisor tiene que enviar un mensaje al receptor, el receptor puede leerlo y, a cambio, debe enviar una prueba de recepción al remitente.

Para superar las reticencias entre las partes y garantizar la equidad, casi todas las propuestas existentes utilizan una TTP. Este tercero de confianza puede jugar un papel importante, participando en cada intercambio o un rol más relajado en el cual solo está activa en caso de surgir una disputa entre las partes (protocolos optimistas).

Debido a la incompatibilidad entre algunas de las propiedades y la dificultad para lograr simultáneamente otras propiedades, podemos encontrar protocolos que resuelvan el intercambio de manera eficiente con una TTP optimista, aunque solo logran una equidad débil [5], otros sistemas enfocados al logro de características específicas como la transferibilidad de evidencias [12], la verificabilidad de la TTP [14], la prevención del rechazo selectivo basado en la identidad del remitente [16], la flexibilidad para permitir la entrega a múltiples receptores [6] o la reducción del volumen de información de estado que la TTP debe almacenar [15].

Este trabajo propone el uso de tecnologías basadas en blockchain y el ecosistema Ethereum para implementar una DApp para notificaciones certificadas de forma descentralizada. Esta propuesta no requiere el uso de correo electrónico y, dependiendo de la aplicación deseada, no requiere ni el uso de una TTP para garantizar la imparcialidad del intercambio. Por lo que sabemos, no hay propuestas para resolver el problema de la notificación certificada utilizando tecnologías basadas en blockchain. Solo algunos documentos sobre pagos equitativos, muy recientemente, se refieren a protocolos de intercambio equitativo sobre blockchain [4], [13].

IV. Desarrollo de los Smart Contracts

Para desarrollar nuestras propuestas no hemos utilizado un framework para DApps como Truffle porque manejan y ocultan algunos de los aspectos importantes de cualquier desarrollo de smart contract. Entonces, aunque estos entornos pueden ayudar a desarrollar y desplegar aplicaciones descentralizadas, hemos elegido, por razones de seguridad, interactuar con los paquetes de software para realizar un seguimiento de todos los detalles de las tareas de desarrollo y despliegue.

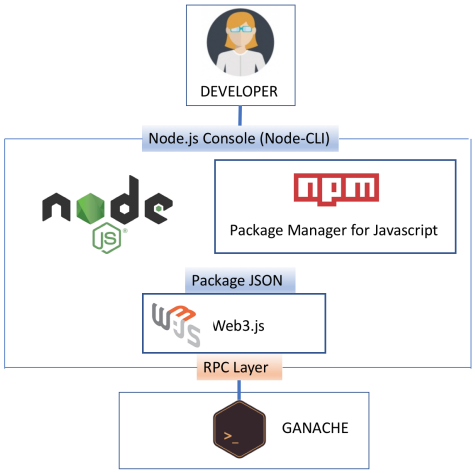


Figura 1. Arquitectura de Desarrollo de Smart Contracts

En la figura 1 hemos representado nuestra configuración. Como desarrolladores, hemos interactuado con la consola de la plataforma Node JS. Esta plataforma modular nos proporciona los componentes necesarios para desarrollar, probar y desplegar aplicaciones descentralizadas, como smart contracts. Las funcionalidades proporcionadas por Node JS se implementan mediante módulos y paquetes independientes. De esta forma, usamos NPM (Node Package Manager) para instalar/desinstalar, configurar y actualizar fácilmente los diferentes módulos y paquetes de software de la plataforma (llamados módulos de terceros).

Para implementar el smart contract usamos un archivo de texto plano llamado *package.json* en formato JSON que se

almacena en la carpeta raíz de la aplicación. Este archivo proporciona los aspectos específicos para administrar las dependencias del módulo que requiere la aplicación. Por ejemplo, el archivo indica información como el nombre de nuestra aplicación, las versiones de los módulos (nombre y versión juntos funcionan como un identificador que se supone único), licencia, directorios, depósito de control de versiones, etc. Teniendo en cuenta esta estructura, un smart contract listo para implementarse se almacenará dentro de una carpeta dentro del archivo de solidity que especifica el código de nuestra aplicación, la carpeta *package.json* y la carpeta *node_modules* con todos los paquetes necesarios. Para comunicar y desplegar el smart contract en una cadena de bloques de Ethereum utilizamos la biblioteca web3.js en la plataforma del nodo. Web3 es una API de Javascript compatible con Ethereum que implementa el protocolo JSON-RPC. JSON-RPC es una especificación de llamada de procedimiento remoto (RPC) que permite la ejecución remota de una función. El formato JSON (JavaScript Object Notation) de los datos intercambiados es simple y más concreta que XML, que se combina con el protocolo de comunicación RPC stateless y liviana. Por lo tanto, utilizamos los métodos de callback de un objeto web3 para comunicarnos con el nodo blockchain a través de llamadas RPC que nos permiten ejecutar programas en una computadora o red remota.

Finalmente, al final de nuestro entorno utilizamos Ganache. Ganache implementa una cadena de bloques personal creada para el desarrollo de smart contracts de Ethereum. Además, Ganache también simula una Máquina Virtual Ethereum (EVM). Después de iniciar Ganache, se muestra una lista con la clave privada y pública de diez cuentas recién creadas en la consola del nodo para ser usadas como banco de prueba (cada cuenta tiene 100 ether en su saldo). Por lo tanto, Ganache está listo para ser utilizado para implementar, ejecutar y probar smart contracts. Dado que Ganache se está ejecutando en local, con el fin de experimentar en un entorno real (por ejemplo, experimentar con retrasos en las confirmaciones), también hemos implementado nuestro contrato en una red de prueba de Ethereum: Rinkeby. Es análogo a la red Ethereum, excepto que el éther utilizado es solo para fines de prueba y no puede utilizarse para comprar objetos o servicios. Sin embargo, las redes de prueba son un lugar ideal para probar smart contracts. Al igual que en la red real de Ethereum, en Rinkeby también tenemos que pagarles gas (pago por las operaciones que permite limitar el cálculo) a los mineros para ejecutar las acciones de nuestros contratos inteligentes y poner nuestras transacciones en la cadena de bloques (existen éthers gratuitos para fines de prueba, disponibles en faucet). No necesitamos descargar toda la cadena de bloques de Ethereum para usar Rinkeby, hemos utilizado el servicio Infura que proporciona clientes de Ethereum que se ejecutan en la nube.

V. Protocolo de Notificaciones Certificadas No Confidenciales basado en Blockchain.

En la primera propuesta consideraremos que no se requiere confidencialidad o incluso no se desea. El protocolo para notificaciones certificadas no confidentiales funciona de la siguiente manera:

1. El remitente, creador del mensaje, usa el contrato inteligente para publicar en la blockchain el hash del mensaje de notificación. Otros parámetros de esta transacción son la dirección del receptor y la fecha límite para que se complete la notificación. Además, se puede requerir un depósito en este paso. Dicha cantidad será incluida en la transacción.
2. El receptor, si acepta la recepción de la notificación, publica un mensaje expresando su voluntad.
3. Finalmente, el remitente publica el mensaje. Como consecuencia, el contrato inteligente publica la prueba de no repudio. Si la ejecución del intercambio requiere un depósito, el smart contract devuelve el importe al remitente y se autodestruye.
   Si no se ejecutan los tres pasos, el smart contract no se destruye automáticamente y, después de la fecha límite, ambas partes pueden acceder a una función del smart contract para solicitar la cancelación de los elementos transferidos.
   a) Cancelación de recepción, solicitada por el receptor si el remitente no publica el mensaje cuando el receptor ha aceptado la notificación.
   b) Cancelación de emisión, solicitada por el remitente, si el receptor no ha aceptado la notificación.
   En ambos casos, el smart contract verifica la identidad del usuario y la fecha límite. El contrato inteligente genera una transacción para señalar que la notificación ha sido cancelada. En el primer caso, el remitente no recibirá el reembolso del depósito (de esta manera, el depósito es útil para motivar al remitente a finalizar el intercambio antes de la fecha límite).

Dado que el mensaje se incluye en una transacción, se registrará en la cadena de bloques, por lo que la notificación en este caso no es confidencial.

V-A. *Smart Contract*

Hemos diseñado e implementado el smart contract para notificaciones no confidenciales. Para esta propuesta, el remitente creará una nueva instancia del smart contract y administrará todos los pasos del intercambio. Ha sido programado usando Solidity y se ha desplegado a través de una red Ethereum. Las direcciones de Ethereum se han asignado tanto al emisor A como al receptor B. Tanto A como B interactuarán con la blockchain utilizando las interfaces de Web3.js.

La figura 2 muestra el smart contract que gestiona las notificaciones no confidenciales. El constructor del contrato inteligente incluye variables para almacenar las direcciones del emisor y el receptor, el hash del mensaje de notificación, el instante de ejecución del primer paso del intercambio y el valor del período de ejecución. El contrato incluye cinco funciones: iniciar la notificación (el constructor), aceptar una notificación, entregar el mensaje, cancelar el intercambio y consultar sobre el estado del intercambio. La notificación certificada es creada por el emisor, que especifica el receptor, el resumen del mensaje y la duración máxima del intercambio. El smart contract también tiene un atributo para almacenar el contenido del mensaje. El contrato administra un evento de Solidity para seguir el progreso del intercambio. Este evento *stateInfo* permite a las partes ver la evolución del estado del intercambio (*created*, *accepted*, *finished* o *cancelled*).

Las funciones *Accept* y *Finish* comprueban la identidad de la dirección que lanza la transacción. Las direcciones del receptor y del remitente se verifican antes de actualizar el estado. La función *Cancel* también verifica las direcciones. En este caso, tanto el emisor como el receptor pueden ejecutar la función, dependiendo del estado del intercambio. Además, todas las funciones que pueden causar cualquier cambio en el estado del intercambio verifican el valor de la variable *State*.

El smart contract administrará la publicación en la blockchain de todos los valores de las variables requeridas para mantener la equidad del intercambio.

```
pragma solidity ^0.4.11;

contract NonConfidentialNofication {
    // Parties involved
    address public sender;
    address public receiver;

    bytes32 public messageHash;
    string public message;
    uint public term;           // Time limit in seconds
    uint public start;          // Start time

    // Possible states
    enum State {created, cancelled, accepted, finished }
    State public state;

    event StateInfo( State state );

    function NonConfidentialNofication(address _receiver, bytes32 _messageHash,
                uint _term) public {
        sender = msg.sender;
        receiver = _receiver;
        messageHash = _messageHash;
        start = now;
        term = _term;
        state = State.created;
        StateInfo(state);
    }

    function accept() public {
        require (msg.sender==receiver && state==State.created);
        state = State.accepted;
        StateInfo(state);
    }

    function finish(string _message) public {
        require (msg.sender==sender && state==State.accepted);
        require (messageHash==keccak256(_message));
        message = _message;
        state = State.finished;
        StateInfo(state);
    }

    function cancel() public {
        require(now >= start+term); // Check term and now (block.timestamp)
        require((msg.sender==sender && state==State.created)
                || (msg.sender==receiver && state==State.accepted));
        state = State.cancelled;
        StateInfo(state);
    }

    function getState() public view returns (string) {
        if (state==State.created) return "created";
        if (state==State.cancelled) return "cancelled";
        if (state==State.accepted) return "accepted";
        if (state==State.finished) return "finished";
    }
}
```

Figura 2. Smart Contract del protocolo de Notificaciones No Confidenciales.

V-B. Propiedades

El protocolo para notificaciones certificadas no confidenciales permite el intercambio equitativo de un mensaje y pruebas de no repudio. Las principales propiedades logradas por el protocolo se analizan en esta sección.

- **Strong Fairness** A no recibirá la prueba de no repudio de recepción provista por el smart contract a menos que realice la transacción que registra el mensaje en la blockchain (caso *State* = *finished*). Por otro lado, B no tendrá acceso al mensaje a menos que realice la transacción aceptando la notificación (*State* = *Accepted*). En ningún momento, el smart contract no genera pruebas alternativas de cancelación o finalización que puedan crear una situación en la que una de las partes pueda tener pruebas contradictorias (lo que llevaría el intercambio a una equidad débil), como se puede ver en la figura 3 .

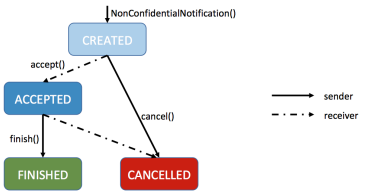


Figura 3. Posibles estados del intercambio en el Protocolo de notificaciones no confidenciales.

- **Ausencia total de TTP. Sustitución por un smart contract.** Esta propuesta no requiere que una parte externa actúe como TTP. Las partes ejecutan las funciones del smart contract creando las transacciones asociadas y no hay necesidad de resolución de disputas.
- **Transferibilidad de las pruebas.** Dado que las partes no pueden obtener pruebas contradictorias de ninguna manera, las pruebas generadas se pueden presentar como evidencia a una entidad externa. Además, su transferibilidad es sencilla, ya que los resultados del intercambio se almacenan en la blockchain. Debido a la inmutabilidad de la cadena de bloques, el contenido de la notificación no se puede modificar, por lo que el sistema proporciona integridad a la notificación. El momento en que se produce la notificación puede derivarse de la marca de tiempo del bloque donde se incluye la transacción.
- **Timeliness Débil.** El protocolo no es asíncrono. Si una de las partes retrasa su intervención en el intercambio, la otra parte no podrá resolverlo hasta la fecha límite. Sin embargo, después de la fecha límite, ambas partes pueden solicitar la finalización del intercambio. Además, el protocolo quiere motivar al emisor a concluir el intercambio antes del tiempo de espera bloqueando una cantidad de dinero en el contrato inteligente. Esta cantidad solo será reembolsada al remitente si concluye antes de la fecha límite.
- **No Repudio.** El protocolo logra el no repudio de origen junto con el no repudio de recepción después de la ejecución del intercambio. A no puede negar haber enviado el mensaje ya que hay una transacción en la blockchain de su dirección que contiene el mensaje y otra relacionada con el mismo mensaje, incluida la dirección del receptor y el hash del mensaje. B no puede negar haber recibido la notificación ya que hay una transacción desde su dirección en la blockchain aceptando la recepción del mensaje y el *State* del intercambio es *Finished*, por lo que el mensaje es accesible públicamente en la blockchain.

VI. Protocolo de Notificaciones Certificadas Confidenciales basado en Blockchain.

La segunda propuesta se ha diseñado teniendo en cuenta las notificaciones que requieren confidencialidad. Es decir, la blockchain tiene que preservar la equidad del intercambio, pero el mensaje no puede almacenarse en un bloque de acceso público. La principal diferencia con la primera propuesta es que en este caso el protocolo permite un intercambio optimista, es decir, el intercambio se puede ejecutar completamente sin la intervención de la TTP ni de la cadena de bloques. Otra característica importante es que esta propuesta no requiere un plazo y puede finalizar en cualquier momento. Una TTP que no almacene información de estado puede resolver las disputas que surjan entre las partes.

El protocolo para notificaciones certificadas confidenciales funciona de la siguiente manera: Las partes, A (emisor) y B (receptor) ejecutarán un intercambio directo en tres pasos:

1. A envía un mensaje cifrado a B con una clave de sesión. Además, A también envía un elemento a B que podría ser útil en caso de disputa, es decir, si A no sigue los pasos del protocolo (es decir, la clave de sesión cifrada con la clave pública de la TTP). La TTP es la responsable del despliegue del smart contract que gestionará el intercambio.
2. B envía la prueba de no repudio.
3. A envía la clave para descifrar el mensaje.

Si alguna de las partes no sigue el protocolo, el intercambio se puede resolver de la siguiente manera:

a) Si A no recibe el elemento del paso 2., puede enviar una solicitud al smart contract con el identificador de la notificación. Si el estado de esta notificación es 'Created' (ni 'Canceled' ni 'Finished'), el estado cambiará a 'Canceled', lo que indica que la notificación no se ha realizado correctamente.

b) Si B no recibe el mensaje en el paso 3., B se pondrá en contacto con la TTP que proporciona los elementos recibidos en el paso 1. junto con la prueba de no repudio de recepción. La TTP accederá al contrato inteligente para verificar el estado de la notificación. Si la notificación no se ha cancelado, la TTP publicará en la blockchain una prueba de no repudio de recepción y los elementos necesarios para que B obtenga el mensaje confidencial.

El protocolo propuesto para notificaciones certificadas confidenciales se basa en el protocolo descrito en [5], un protocolo optimista en tres pasos (figura 4 y figura 5) con una TTP que participa solo en caso de disputas entre las partes. En [5], ambas partes pueden contactar con la TTP que mantiene la información del estado.

La notación y los elementos utilizados en la descripción del protocolo son los siguientes:

1. $A \rightarrow B$: c, k_T, h_A
2. $B \rightarrow A$: h_B
3. $A \rightarrow B$: k_A

Figura 4. Intercambio Optimista en tres pasos [5].

M	message to be sent certified from *A*(lice) to *B*(ob)
$c = E_K(M)$	symmetric encryption of message *M* with key *K*, producing the ciphertext *c*
$k_T = PU_T(K)$	key *K* enciphered with the public key of the TTP (it means that only the TTP, who knows the correspondent private key, can decrypt it)
$h_A = PR_A\{H[H(c), k_T]\}$	signature of *A* on the concatenation of the *hash* of ciphertext *c* and k_T, part of the evidence of NRO (non-repudiation of origin) for *B*
$h_B = PR_B\{H[H(c), k_T]\}$	signature of *B* on the concatenation of the *hash* of ciphertext *c* and k_T, evidence of NRR (non-repudiation of receipt) for *A*
$k_A = PR_A("key=", K)$	key *K* enciphered with the private key of *A* (and so, signed), second part of NRO evidence for *B*

Figura 5. Descripción de los elementos intercambiados [5].

En esta solución basada en blockchain, el emisor A y el receptor B intercambiarán directamente evidencias de no repudio y mensajes. Solo como último recurso, en caso de que no puedan obtener los elementos esperados de la otra parte, se invocará el smart contract o la TTP, iniciando las funciones *cancel* o *finish*. Hemos diseñado e implementado un DApp que permite el intercambio optimista entre las partes y un smart contract para la resolución de disputas. El contrato inteligente se ha programado en Solidity y se ha implementado a través de la red Ethereum (ver § IV). Las direcciones de Ethereum se han asignado al emisor A, al receptor B y a la TTP. En comparación con el protocolo descrito en [5], el papel de la TTP se ha reducido. El remitente nunca se pondrá en contacto con la TTP. La TTP responderá solo las solicitudes del receptor B accediendo al contrato inteligente que se ha implementado. La TTP es totalmente stateless y en ningún caso almacena información sobre el estado del intercambio. Tanto A como B pueden interactuar con la blockchain si es necesario a través de la interfaz Web3.js.

VI-A. Smart Contract

La figura 6 muestra el smart contract que gestionará las posibles disputas entre las partes tras el intercambio descrito en la figura 4 para notificaciones confidenciales. Este smart contract es desplegado por la TTP, quien define las identidades de emisor y receptor. El smart contract gestiona la variable *state* para mantener el estado del intercambio.

El evento *stateInfo* permite el seguimiento del estado. La función *Cancel* comprueba la identidad de la dirección que lanza la transacción, que debe ser la dirección del remitente, junto con el valor de la variable *state*. Esta función solo puede ser ejecutada por el remitente. La función *Finish* comprueba la identidad de la parte que envía la transacción, es decir, la TTP, junto con el valor de la variable *state*. La TTP ejecutará esta función si recibe una solicitud del receptor. El smart contract es responsable de la publicación en la blockchain

```
pragma solidity ^0.4.11;

contract ConfidentialNotifications {

   //Parties involved
    address sender;
    address receiver;
    address ttp;

    string hB; //NRR proof
    string hBt; //Intervention proof

    //Possible states
    enum State { created, cancelled, finished }
    State public state;

    function ConfidentialNotification (address _sender, address _receiver){
        ttp = msg.sender;
        sender = _sender;
        receiver = _receiver;
        state = State.created;
        }


    event stateInfo(
        State state
        );

    function Cancel() returns (string) {
         if(msg.sender==sender){
             if(state==State.created){
                 state=State.cancelled;
                 //return abort token
                 stateInfo(state);
             }else if (state == State.finished){
                 return hB;
             }
         }
    }

    function Finish(string _hB, string _hBt) returns (State) {
        if (msg.sender==ttp){
            if(state==State.cancelled){
                return state;
            }else{
                hB=_hB;
                hBt=_hBt;
                state=State.finished;
            }
        }
    }

    function getState() returns (string){
        if(state==State.cancelled) return "cancelled";
        if(state==State.created) return "created";
        if(state==State.finished) return "finished";
    }
}
```

Figura 6. Smart Contract para Notificaciones Confidentiales.

de los valores de los elementos utilizados para mantener la equidad del intercambio. En la función *Finish*, si se cumplen las condiciones, el smart contract publica en la blockchain tanto la prueba de no repudio de recepción (hB) como la clave de sesión cifrada con la clave pública de B (hBt).

VI-B. Propiedades

A continuación se evaluan las principales propiedades del sistema de notificaciones confidenciales.

- **Equidad Débil.** El protocolo no permite que ninguna de las partes reciba el elemento esperado si la otra parte no lo recibe. Sin embargo, la intervención de la TTP puede llevar a una situación en la que una de las partes posea evidencias contradictorias. Un A malicioso puede tener la prueba de no repudio recibida directamente de B y también la prueba de cancelación generada por el smart contract después de una solicitud de cancelación de A. Por esta razón, la equidad será débil y las pruebas generadas serán intransferibles. Pero al comparar esta característica con la versión del protocolo sin blockchain, el nuevo sistema no requiere que el árbitro consulte a ambas partes para resolver el estado final del intercambio. Es suficiente verificar una de las partes y luego confrontar esta versión con los contenidos de la blockchain.
- **Optimista.** Las partes pueden finalizar el intercambio sin la necesidad de contactar con una TTP ni ejecutar ninguna función del smart contract. Si las partes no siguen el protocolo y se requiere la ejecución del smart contract, el gas necesario para su funcionamiento se reduciría en comparación con el protocolo para notificaciones no confidenciales.
- **Stateless TTP.** Cuando la TTP se involucra en el intercambio, lo resuelve usando el smart contract. La TTP no necesita almacenar ningún tipo de información de estado sobre el intercambio.
- **Timeliness.** Las partes pueden finalizar el intercambio en cualquier momento accediendo al smart contract (emisor A) o contactando con la TTP (receptor B). La duración de la resolución dependerá del tratamiento de las validaciones de bloques. El protocolo puede suponer que las transacciones son válidas inmediatamente o esperar hasta que se confirme el bloque en la cadena.

- **No repudio.** El protocolo proporciona no repudio de origen junto con no repudio de recepción tras la ejecución del intercambio de tres pasos o la finalización mediante el smart contract. A no puede negar haber enviado el mensaje ya que B tiene el elemento recibido en el tercer paso o el estado del smart contract es *Finished*. B no puede negar haber recibido la notificación ya que A tiene los elementos enviados por B en el tercer paso del protocolo.
- **Confidencialidad.** Si el intercambio se completa mediante la ejecución del protocolo de intercambio de tres pasos, entonces no hay otra entidad involucrada en el intercambio, y el mensaje permanece confidencial. Si la TTP participa o se ejecutan las funciones del smart contract, entonces la TTP procesará los elementos recibidos y realizará una transacción que incluye el elemento que permitirá que B descifre el mensaje, pero el mensaje simple no se incluye en la transacción por lo que no se incluirá en un bloque de la blockchain para preservar la confidencialidad.

VII. Conclusiones y Trabajo Futuro

Las soluciones anteriores para certificaciones notificadas se basan principalmente en la intervención de una TTP que actúa como intermediario entre emisor y receptor. Este trabajo presenta dos alternativas para el envío de notificaciones certificadas con equidad basada en blockchain. Las propuestas presentan diferentes conjuntos de propiedades, por lo que la elección de uno u otro sistema dependerá basicamente de las necesidades de cada intercambio. La primera solución (§ V) permite a los usuarios el envío de notificaciones no confidenciales mediante una DApp que soporta el envío y la recepción de mensajes certificados y garantiza la equidad del intercambio sin requerir la intervecnión de ninguna TTP para garantizar las propiedades de seguridad del intercambio ya que las acciones de los diferentes actores se registran en la blockchain y en caso de que alguno de los actores no siga las especificaciones del protocolo, el smart contract generará las evidencias correspondientes para preservar la equidad. Esta propuesta también garantiza las propiedades de Timeliness limitado (las partes saben con certeza que el intercambio se completará en un tiempo finito [13]), Transferibilidad de las pruebas y No repudio, tal como se demuestra en § V-B.

La segunda solución (§ VI) es un protocolo de intercambio equitativo que permite a los usuarios enviar notificaciones confidenciales con una TTP optimista (su intervención solo es necesaria si una de las partes no cumple con el protocolo) para garantizar la equidad. Gracias al uso de blockchain y un smart contract, la TTP no necesita almacenar información del estado del intercambio de la ejecución de protocolo porque toda la información de cada intercambio se almacena en el blockchain usando el smart contract). Esta solución asegura el intercambio equitativo (equidad, ver § VI-B) y también preserva las propiedades de asincronía y no repudio. Sin embargo, estrictamente, no se consigue la propiedad de transferibilidad de las pruebas porque cualquier entidad externa que quiera verificar la corrección del intercambio no solo tiene que verificar la evidencia provista por una de las partes sino también tiene que verificar el contenido de la blockchain. La tabla I compara las propiedades de ambas soluciones.

Como trabajos adicionales vamos a probar los contratos inteligentes en redes reales, verificando el tiempo de retardo de confirmación de transacciones y posibles efectos no deseados causados por estos retrasos. A continuación modificaremos el smart contract para notificaciones no confidenciales para exigir un depósito cuando el remitente crea un nuevo mensaje y para permitir administrar más de una notificación. Otra modificación permitirá, en el smart contract de notificaciones confidenciales, que la TTP pueda crear un contrato principal, y a partir de esto, múltiples notificaciones puedan ser creadas por los remitentes. También mejoraremos el uso de eventos. Los usuarios pueden recibir notificaciones, pero necesitan la dirección del smart contract para suscribirse a ellos. Finalmente, estudiaremos la cantidad necesaria de gas para ejecutar estos contratos y las posibles mejoras al código para reducir las comisiones pagadas por el uso de la blockchain.

Tabla I
Comparación de Propiedades

Propiedad	Notificaciones No Confidenciales	Notificaciones Confidenciales
No-repudio	SI	SI
Equidad	Fuerte	Débil
Timeliness	Limitado	SI
Efectividad	SI	SI
TTP	NO	OPTIMISTA/STATELESS
Transferabilidad de evidencias	SI	NO
Confidencialidad	NO	SI

Agradecimientos

Este trabajo ha sido paricalmente financiado por los proyectos AccessTur TIN2014-54945-R y ARES TIN2015-70054-REDC.

Referencias

[1] Asokan, N., Shoup, V. y Waidner, M.: “Asynchronous Protocols for Optimistic Fair Exchange”; Proceedings of the IEEE Symposium on Research in Security and Privacy, pages 86-99, Oakland, CA, May 1998.

[2] Barber, S., Boyen, X. Shi, E. y Uzun, E., ”Bitter to Better — How to Make Bitcoin a Better Currency”. Financial Cryptography and Data Security, 2012. Springer Berlin Heidelberg. p. 399-414.

[3] Chaudhry, S., Farash, M., Naqvi, H. y Sher, M. “A secure and efficient authenticated encryption for electronic payment systems using elliptic curve cryptography”. Journal of Electronic Commerce Research, 2016. Vol. 16, Num. 1, p. 113-139.

[4] Delgado-Segura, S., Perez-Sola, C., Navarro-Arribas G. y Herrera-Joancomarti, J. “A fair protocol for data trading based on Bitcoin transactions”, Future Generation Computer Systems, 2017.

[5] Ferrer-Gomila, J., Payeras-Capellà, M. y Huguet i Rotger, L.: “An Efficient Protocol for Certified Electronic Mail”. ISW 2000: 237-248

[6] Ferrer-Gomila, J., Payeras-Capellà, M. y Huguet i Rotger. L.: “A Realistic Protocol for Multi-party Certified Electronic Mail”. ISC 2002

[7] Ferrer-Gomilla, J., Onieva, J. Payeras-Capellà, M. y Lopez. J: “Certified electronic mail: Properties revisited”, “Computers & S ecurity”, vol. 29, nº 2, p. 167 - 179, 2010.

[8] Heilman, E., Baldimtsi, F. y Goldberg, S. ”Blindly Signed Contracts: Anonymous On-Blockchain and Off-Blockchain Bitcoin Transactions”.Financial Cryptography and Data Security, 2016. p. 43–60.

[9] Ethan H., Baldimtsi, F., Alshenibr, L., Scafuro, A. y Goldberg, S.: ”TumbleBit: An Untrusted Tumbler for Bitcoin-Compatible Anonymous Payments”. Network and Distributed System Security Symposium 2017.

[10] Huang, Q., Yang, G. Wong, D. y Susilo, W. .A new efficient optimistic fair exchange protocol without random oracles”. International Journal of Information Security, 2012. Vol. 11, Num. 1, p.53-63. ISSN 1615-5270.

[11] Huang, Q. Wong, D. y Susilo, W. “P2OFE: Privacy-Preserving Optimistic Fair Exchange of Digital Signatures”. Topics in Cryptology – CT-RSA 2014, 2014. Springer International Publishing. p. 367-384.

[12] Kremer S, Markowitch O. Selective receipt in certified e-mail. LNCS 2247. In: Advances in cryptology: proceedings of indocrypt. Springer-Verlag; 2001. p. 136–48.

[13] Jian Liu, Wenting Li, Ghassan Karame, N. Asokan: “Towards Fairness of Cryptocurrency Payments”, en IEEE Security and Privacy, 2017.

[14] Mut-Puigserver, M., Ferrer-Gomila, J. y Huguet-Rotger, L.: “Certified e-Mail Protocol with Verifiable Third Party”. 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service. p. 548-551.

[15] Onieva, J., Zhou, J. y Lopez J.: “Enhancing certified email service for timeliness and multicast”. Fourth international network conference; 2004. pp. 327–35.

[16] Payeras-Capellà, M., Mut-Puigserver, M., Ferrer-Gomila, J. y Huguet-Rotger, L.: ”No Author Based Selective Receipt in an Efficient Certified E-mail Protocol”. PDP 2009: 387-392.

[17] Shao, Z. ”Fair exchange protocol of Schnorr signatures with semi-trusted adjudicator. Computers & Electrical Engineering, 2010.

[18] Zhou, J., Deng, R. y Bao, F.: “Some Remarks on a Fair Exchange Protocol”; Third International Workshop on Practice and Theory in Public Key Cryptosystems, LNCS 1751, Springer, p. 46-57, 2000.

Uso y retos de *blockchain* en plataformas de votación electrónica

Victor Garcia-Font
Departament d'Enginyeria Informàtica i Matemàtiques,
Universitat Rovira i Virgili (URV)
CYBERCAT-Center for Cybersecurity Research of Catalonia
victor.garcia@urv.cat

Helena Rifà-Pous
Internet Interdisciplinary Institute (IN3),
Universitat Oberta de Catalunya(UOC)
CYBERCAT-Center for Cybersecurity Research of Catalonia
hrifa@uoc.edu

***Resumen*—Aunque los proyectos de votación electrónica son una realidad desde hace ya algunos años, muchos países se resisten a adoptar esta tecnología en sus procesos electorales fundamentalmente por falta de confianza en las soluciones propuestas y por motivos de coste. En 2009, la aparición de Bitcoin y su tecnología subyacente *blockchain* supuso una revolución en los sistemas TI descentralizados, y su influencia también se produjo en el contexto de la votación electrónica para superar los problemas de seguridad y coste. La adopción de *blockchain* en los sistemas de votación se ha realizado de dos maneras: utilizando criptomonedas o *tokens* criptográficos para representar votos o utilizando la *blockchain* como registro inmutable. En este artículo revisamos sistemas representativos de ambos usos, estudiamos que aporta *blockchain* a cada proyecto, y discutimos los beneficios y problemáticas de esta tecnología en el escenario de la votación electrónica.**

***Index Terms*—votación electrónica, blockchain, seguridad información**

I. Introducción

La transición de un sistema de voto convencional a electrónico es un paso que muchos países y juntas electorales se resisten a tomar. De hecho, innovaciones tecnológicas como DRE (*Direct Recording Electronic*) para facilitar el voto o el recuento ya han probado ser vulnerables y tener bugs difícilmente detectables [1] y por lo tanto, la falta de confianza en este tipo de sistemas es notable. En sistemas que funcionan online la preocupación es aún mayor ya que las transacciones están expuestas a ciberataques, puede haber vulnerabilidades en los dispositivos de votación personal, hay un elevado riesgo de coacciones, etc.

Por otro lado, la votación electrónica mejora problemas de los sistemas tradicionales, como las dificultades para ejercer el derecho a voto de residentes en el extranjero [2], la dificultad de participar que tienen personas enfermas o con problemas de movilidad, el desincentivo para la participación que suponen largas colas en frente de una mesa electoral (en las elecciones de presidenciales americanas de 2012 se registraron colas de hasta 7 horas [3]), etc. También abren un abanico de nuevas posibilidades, permitiendo, además de las típicas votaciones a candidatos cada varios años, otro tipo de consultas sobre temas específicos que podrían producirse más a menudo, o incluso permitirían reformar más profundamente el proceso electoral introduciendo conceptos de democracia líquida [4], donde por ejemplo, se pudiera delegar el voto a otro participante en consultas relacionadas con un tema concreto.

Todo esto hace que empresas tecnológicas y organismos públicos sigan trabajando en sistemas de votación electrónica e investiguen la incorporación de nuevas tecnologías para mejorar la seguridad y la transparencia de sus sistemas. En este sentido *blockchain* es una tecnología que recientemente ha ganado popularidad en varios sectores, y podría contribuir a mejorar el campo de las votaciones electrónicas gracias a propiedades como la inmutabilidad, la integridad y la auditabilidad. En este artículo, revisamos algunos de los principales proyectos de votación electrónica y vemos que las principales contribuciones de *blockchain* en este campo pueden dividirse entre los proyectos que usan criptomonedas o *tokens* criptográficos como representación de votos, por lo que *blockchain* representa una parte fundamental del sistema de votación, y por otro lado, los proyectos que usan *blockchain* como registro inmutable para mejorar o sustituir parte de un sistema de votación electrónico implementado con otras tecnologías. En este artículo estudiamos sistemas representativos de ambos usos, vemos que aporta *blockchain* a cada proyecto en términos de seguridad, y de forma más general discutimos las ventajas e inconvenientes de la tecnología *blockchain* en este contexto. El resto del paper está organizado en las siguientes secciones: la Sección II introduce las tecnologías básicas para contextualizar este artículo; en la Sección III se describen los diferentes usos de *blockchain* en diferentes iniciativas de votación electrónica; en la Sección IV se discuten las contribuciones y los problemas que trae *blockchain* en este contexto; y finalmente en la Sección V se concluye el paper.

II. Contexto tecnológico

Esta sección revisa los conceptos y las tecnologías básicas para contextualizar el resto del paper. Para ello, la Sección II-A revisa los conceptos más importantes y las propiedades de seguridad que deben cumplir los sistemas de votación electrónica. Posteriormente, la Sección II-B introduce la tecnología *blockchain*.

II-A. Votación electrónica

Entre los sistemas que permiten votaciones electrónica, se puede encontrar tanto sistemas que han estado concebidos sin alterar el espíritu de las votaciones tradicionales en papel (*i.e.* votaciones grandes y muy espaciadas en el tiempo, con registros de votantes centralizados), como sistemas que pretenden añadir funciones al proceso (democracia líquida, consultas populares, ...). En esta sección revisamos las propiedades de seguridad que deben cumplir estos sistemas y describimos los componentes y servicios más destacados para este artículo que

integran un sistema de votación electrónica. Información más exhaustiva sobre este tema puede encontrarse en [5].

A continuación listamos las características relacionadas con la seguridad y la privacidad que los sistemas de votación electrónica buscan conseguir:

- **Elegibilidad**: sólo los votantes legítimos (forman parte del censo o de un registro de votantes) pueden votar.
- **Autenticidad** e **integridad**: el voto debe ser emitido por un votante y no debe haberse modificado entre su emisión y su recuento.
- **Privacidad de los votantes**: ningún ente, tanto interno como externo de la votación, debe poder conocer el voto de un votante.
- **Secreto de los resultados intermedios**: hasta que no se finaliza la votación, no deben poder conocerse resultados parciales.
- **Incoercibilidad**: los votantes deben poderse proteger ante intentos de coerción.
- **Verificabilidad E2E** (extremo a extremo): las partes implicadas deben poder verificar que el proceso en el que están implicados finaliza satisfactoriamente, *i.e.*, los votantes comprueban que su elección se ha registrado al sistema, y se puede verificar que los votos escrutados corresponden a los votos emitidos y están bien contabilizados.

A continuación, veremos los componentes y servicios relacionados con este artículo más destacados de un sistema de votación electrónica:

- **Aplicación cliente**: utilizada por el votante para seleccionar y enviar la opción que desea votar. Esta aplicación puede estar en un dispositivo del votante (*e.g.* teléfono móvil, *PC*), y por lo tanto está en un entorno que debe ser considerado como hostil, no controlado por el proveedor del sistema de votación.
- **Servicio de filtrado *(cleansing)***: elimina votos duplicados o inválidos, y los datos personales (p.e., firma del votante) asociados a cada voto.
- **Servicio de mezclado**: recifra los votos cifrados y los mezcla, de forma que los votos se reordenan completamente en la urna y no sea posible correlarlos con ningún votante.
- **Sistema multi autoridad**: impide que una sola parte acceda al contenido de los votos. Con estos sistemas se suelen cifrar los votos utilizando una clave dividida en N partes, requiriendo M de las partes para poder recuperar la clave para descifrar los votos.
- **Servicio de recuento de votos *(talling)***: descifra los votos y hace el recuento.
- **Tablón de anuncios *(Bulletin board)***: panel público donde se registran datos importantes sobre la votación.
- **Servicio de *logs***: registra los eventos destacados del sistema de votación.

II-B. Blockchain

En 2008, el *white paper* titulado *Bitcoin: A peer-to-peer electronic cash system* [6] no sólo actúa de punto de partida para Bitcoin, sino que también introduce la tecnología *blockchain* (cadena de bloques en español), un sistema descentralizado que, mediante un protocolo de consenso, permite a los participantes del sistema registrar datos de forma inmutable sin utilizar una tercera parte de confianza. Posteriormente, utilizando la misma filosofía de la *blockchain* propuesta en Bitcoin, se han creado otros sistemas que se adaptan a otros propósitos o requerimientos. Por ejemplo, algunas propuestas interesantes que intentan mejorar el diseño original de Nakamoto plantean la implementación de técnicas orientadas a preservar la privacidad de los usuarios, como el uso de *ring signatures* en CryptoNote [7] o de las pruebas de conocimiento nulo zk-SNARKs en ZeroCash [8].

Otra característica destacada de la *blockchain* de Bitcoin, y de la de la mayoría de criptomonedas, es que se trata de un sistema público, en el cual no sólo los registros pueden ser consultados libremente, sino que también, cualquier nodo que siga las normas del protocolo puede participar en el protocolo de consenso e incluir nuevos datos sin necesidad de contar con permisos especiales. Por esto a estas *blockchain* se las denomina sin permisos (*permissionless*). Por otro lado, siguiendo otros requerimientos de seguridad y accesibilidad, posteriormente se propusieron las *blockchains* denominadas con permisos (*permissioned*), que podían tener contenido privado y requerir permisos especiales para poder participar en el consenso. Ripple[1] es uno de los proyectos más destacados usando este tipo de *blockchain*.

La tecnología *blockchain* ha evolucionado bastante desde su diseño original. Aun así, las propiedades básicas de seguridad que la distinguen como un sistema de registro descentralizado e inmutable están presentes de forma transversal entre todas las nuevas propuestas. En este artículo, nos centramos en el uso que hacen los diferentes proyectos de votación electrónica de las características de seguridad y privacidad vinculadas a las *blockchain* utilizadas. Para más información sobre la tecnología, el lector puede consultar otros artículos como [9].

III. Tipología de uso de blockchain

Aprovechando las características de seguridad en sistemas descentralizados que aporta la tecnología *blockchain*, algunos proyectos de votación electrónica han incluído esta tecnología ya sea para mejorar la seguridad de alguno de sus componentes, o incluso para que forme parte del corazón del sistema. De entre los usos que hemos observado, se pueden diferenciar claramente dos maneras de emplear la *blockchain*: por un lado, algunos proyectos utilizan una criptomoneda o *token* (i.e una representación de un recurso o utilidad que reside en la parte superior de una *blockchain*) y representan un voto a través de una transacción; y por otro lado, otros proyectos usan *blockchain* tan solo por sus propiedades de registro inmutable, básicamente para implementar el tablón de anuncios o para enviar pruebas de integridad de *logs*, votos, o algún otro evento. En las siguientes secciones veremos ejemplos representativos de estas dos tipologías de uso.

III-A. Criptomonedas para representar votos

De entre los proyectos revisados que usan criptomonedas para representar votos, queremos destacar la propuesta hecha por *P. Tarasov et al.* [10], donde los autores proponen el uso de la criptomoneda Zcash. Para entender el sistema propuesto, primero es necesario conocer que las criptomonedas utilizan

[1]Ripple, https://ripple.com/

unos identificadores llamados direcciones que tienen un uso parecido al número de una cuenta bancaria, se utilizan para enviar y recibir criptomonedas. Una dirección está asociada a una clave pública y cuando un usuario recibe una transacción en ella, puede utilizar el contenido recibido si dispone de la clave privada correspondiente. Las direcciones y claves privadas de cada usuario se almacenan en un monedero. Zcash dispone de dos tipos de direcciones: las direcciones transparentes (*t-address*) que son direcciones asimilables a las direcciones de Bitcoin en las que el envío de las criptomonedas es transparente, y las direcciones blindadas (*z-address*) que son direcciones para preservar la anonimidad en las transacciones. Así, dependiendo de la combinación de direcciones que se usen para el envío y recepción de las criptomonedas podemos diferenciar entre 4 tipos de transferencias: (1) pública cuando se usan dos *t-address*; en este caso se conoce el importe y se permite la trazabilidad de la moneda, (2) *shielding* cuando se usa una *t-address* de envío y una *z-address* de recepción; en este caso se revela el saldo enviado pero se quiebra la enlazabilidad con una futura dirección transparente, (3) *deshielding* cuando se usa una *z-address* de envío y una *t-address* de recepción; en este caso, la dirección transparente de recibo anula el blindaje de las monedas y revela en el *blockchain* el valor recibido, y (4) privada cuando se usan dos *z-address* y se mantiene secreto el saldo y las direcciones de la transacción.

De esta forma los autores proponen un sistema de votación basado en 4 etapas básicas en las que usando movimientos de ZEC (criptomoneda de Zcash) entre diferentes tipos de direcciones de los votantes, de los candidatos y de la junta electoral, se busca mantener las propiedades básicas de un sistema de votación descritas en II-A y a la vez tener un sistema seguro, flexible y más económico de implementar y desplegar que un sistema de votación electrónica convencional. Las 4 etapas que propone el sistema son las etapas que en líneas generales proponen todos los sistemas de *e-voting*: registro, notificación, votación y recuento/auditoría. A continuación explicaremos de forma resumida estas etapas.

Para empezar, poco antes de iniciar el periodo electoral, los usuarios registrados como votantes en el sistema reciben un *email* con un enlace que se activa con el inicio oficial de los comicios. Al acceder al enlace, al votante se le presenta un formulario donde puede escoger a su candidato y donde introduce una *t-address* controlada por él. Una vez que el usuario rellena el formulario y autoriza el voto, se transfiere cierta cantidad de ZEC desde una dirección de un monedero de la junta electoral a la *t-address* del votante. Estos ZEC son reenviados automáticamente utilizando una *z-address* del votante para preservar su privacidad a una dirección de un monedero del candidato que se haya escogido. Una vez finaliza la elección, los *tokens* de las direcciones de los candidatos se transfieren otra vez al monedero de la junta electoral para su recuento. En la Figura 1 se puede ver un esquema simplificado de la etapa de votación de esta propuesta. Existen dos variantes del sistema dependiendo de si los candidatos reciben los *tokens* en direcciones del tipo *z-address* o *t-address*. Con el uso de *z-addresses* se incrementa la privacidad del sistema, pero se rompe la enlazabilidad de las transacciones, con lo que un votante no podría tracear que

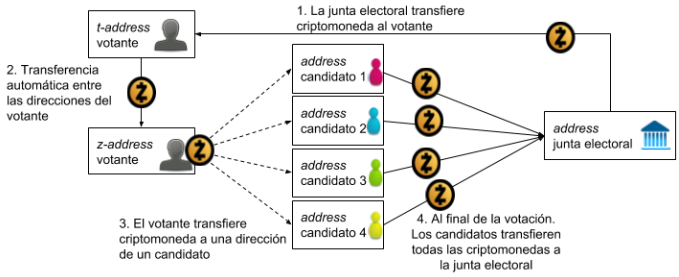


Figura 1. Esquema simplificado de la etapa de votación propuesto en [10]

su voto ha sido vaciado del monedero del candidato al de la junta electoral para su recuento. Por contra, si los candidatos usan *t-addresses* para recibir los *tokens*, entonces los votantes pueden comprobar que su voto ha sido enviado a la junta electoral para su recuento, facilitando la verificabilidad E2E. Sin embargo, el uso de estas direcciones también comporta que las direcciones de los candidatos puedan ser observadas en tiempo real, vulnerando el principio de no poder conocer resultados parciales antes de que finalice el proceso electoral.

El uso de *tokens* criptográficos como representación directa de votos se ha propuesto también en otros proyectos que no utilizan una criptomoneda, sino que optan por implementar su propio *token*. Por ejemplo, Democracy.earth[2] propone desplegar su propio *vote token*. Al desplegar su propio *token*, estos proyectos no sólo pueden implementar más fácilmente las funciones de votación, sino que también evitan algunos problemas asociados al uso de criptomonedas como *tokens*, como por ejemplo el riesgo asociado al valor de mercado que tiene la criptomoneda. Además, al utilizar su propio *token*, estos proyectos pueden customizar más los proyectos y ofrecer sistemas de votación que van mucho más allá de las características democráticas ofrecidas por las votaciones en papel, ofreciendo mecanismos de democracia líquida, o permitiendo la creación y el mantenimiento de organizaciones descentralizadas en una jurisdicción digital sin fronteras, como es el caso de Aragon[3].

III-B. Blockchain como registro inmutable

Uno de los proyectos más avanzados de votación electrónica que utiliza *blockchain* como registro inmutable como pieza central de su tecnología es Agora[4]. La arquitectura de este proyecto [11] está dividida en 4 capas: un tablón de anuncios, Cotena, la *blockchain* de Bitcoin y Votapp. A continuación describiremos los componentes relacionados con la *blockchain*.

Por un lado, el tablón de anuncios es una *blockchain* con permisos con una arquitectura de *skipchain* [12], donde a ciertos nodos operados por Agora o por terceros se les concede permisos o bien de escritura o de solo lectura. Esta *blockchain* se utiliza de registro inmutable de datos generados durante el proceso electoral (p.ej. votos, pruebas de conocimiento nulo evidenciando la corrección de un subproceso del sistema de votación). Así, existen un conjunto de nodos con permisos suficientes de escritura conocidos como Cothority (Collective authority), que confirman de forma colectiva la inserción de

[2]Democracy.earth, https://www.democracy.earth/
[3]Aragon, https://aragon.one/
[4]Agora, https://agora.vote/

transacciones en el tablón de anuncios. La forma de funcionar de la Cothority se basa en la elección rotativa de un nodo especial al que se denomina oráculo y que es el encargado de realizar la mayor parte de la gestión, dejando a los otros nodos básicamente tareas de colaboración y de monitorización del oráculo. De esta manera, todos los nodos reciben las transacciones con los votos emitidos por los votantes, entre otros datos de la votación, pero es el oráculo el encargado de proponer nuevos bloques para ser incluídos en la *blockchain* con las transacciones recibidas. Además, el tablón de anuncios actúa de canal central de comunicación y de memoria, siendo el oráculo el encargado de escribir bloques en el *log* de Cotena (explicado más adelante) y enviar pruebas de integridad a la *blockchain* de Bitcoin. Así, los nodos tienen la responsabilidad de: (1) mantener una copia del tablón de anuncios; (2) recibir los votos encriptados de los votantes, autenticar sus datos y asegurar que los votos son enviados por un votante autorizado; (3) confirmar los bloques propuestos por el oráculo; (4) descifrar los votos una vez que se ha acabado la elección y crear votos en texto en claro para que puedan ser recontados; y (5) mantener una copia del *log* de Cotena y monitorizar que es correcto.

La segunda capa de la arquitectura de Agora es conocida como Cotena. Este componente es básicamente un mecanismo de *logging tamper-resistant* construido encima de la *blockchain* de Bitcoin, donde se escriben pruebas de integridad del tablón de anuncios. De esta forma, Agora usa el poder computacional y la transparencia que ofrece la red de Bitcoin para proteger la integridad de su *blockchain* con permisos. La *blockchain* de Agora permite que haya una constancia pública de cada paso de la votación y que las etapas sean fácilmente auditables. El sistema es verificable extremo-a-extremo ya que un auditor podría comprobar que: los parámetros de configuración son correctos (no lo hemos mencionado pero otros muchos datos son publicados por los nodos de la Cothority en el tablón de anuncios, como los ficheros de configuración); que los votos cifrados corresponden a votantes autenticados especificados en la lista de configuración y reflejan sus elecciones (se utiliza una validación *cast-or-challenge* en la que votantes descifran de forma aleatoria votos de tipo test para comprobar que el voto emitido es correctamente registrado en el sistema); que la red de mezclado funciona correctamente y que las pruebas de conocimiento nulo que se generan corresponden a pasos correctos para hacer una verificación aleatoria y anonimizar correctamente los votos y que los votos descifrados parcialmente por cada nodo son correctos y el texto en claro ha sido reconstruido correctamente; y que el recuento final está bien computado. Finalmente, auditores oficiales pueden firmar un atestado y registrarlo también en el tablón de anuncios. De esta manera, Agora utiliza una arquitectura híbrida de *blockchain* con permisos y sin permisos, donde se usa el tablón de anuncios para escribir todos los datos del proceso de votación (*e.g.* votos recibidos en cada nodo) que requieren de un sistema con poca latencia, y posteriormente se hacen capturas del tablón de anuncios en la *blockchain* de Bitcoin, asegurando de esta manera la integridad y la inmutabilidad en una *blockchain* más segura pero que requiere de una latencia más alta para registrar transacciones. Con este diseño, los creadores de Agora aseguran que el sistema dispone de una arquitectura escalable, de alta disponibilidad, usable, de bajo coste, de baja latencia, sin puntos únicos de falla, que permite una verificabilidad fuera de línea y que permite ser adaptada a las regulaciones de diferentes tipos de elección. Aunque todos los pasos son publicados en la *blockchain* y por lo tanto son accesibles por todas las partes, cabe destacar que esta propuesta también respeta otras propiedades de privacidad mencionadas anteriormente requeridas en un sistema de votación electrónica. Por un lado no permite conocer el resultado parcial de la elección, ya que los votos se publican cifrados con un sistema de multi autoridad. Por otro lado, también protege la privacidad del votante, ya que la red de mezclado aporta anonimidad, el voto es secreto, y además, a diferencia de otros sistemas, no se provee al votante con un recibo de votación que pueda después enseñar a un tercero, evitando posibles maniobras de coerción. El uso de *blockchain* como registro inmutable también ha sido propuesto en otros proyectos. Por ejemplo, *J.Cucurull et al.* [13] de Scytl[5] proponen registrar pruebas de integridad de su sistema de *logs* en la *blockchain* de Bitcoin, con una propuesta parecida a lo comentado en Agora. Por otro lado, TIVI[6] proponen en [14] y [15] usar *blockchain* en un servicio de sellado de tiempo digital. Así, con este sistema pueden asegurar que un dato ha existido en un cierto momento en el tiempo. Por ejemplo, se envía una huella dactilar de cada voto en el servicio externo de sellado de tiempo. El sello de tiempo se guarda junto al voto y se le pasa al votante, quien puede verificar que el voto se registró correctamente, y no fue alterado ni eliminado del sistema.

IV. Discusión

Como hemos visto, los sistemas de votación electrónica que incluyen tecnología *blockchain* pueden dividirse claramente entre los que usan la *blockchain* para registrar transacciones que representan directamente los votos (de forma anàloga a las transacciones de una criptomoneda), o sistemas que usan la *blockchain* tan solo para publicar ciertos datos del proceso de voto en un registro inmutable. Además, también hemos visto que la importancia de la *blockchain* en el conjunto de la arquitectura puede ser muy dispar. Sistemas como Agora o [10] tienen la *blockchain* como pieza central, en cambio *J.Cucurull et al.* [13] tan solo proponen mejorar algún elemento dentro de la arquitectura típica de un sistema de *e-voting*.

En cualquier caso, la *blockchain* se incluye en los diferentes proyectos para contribuir en los requerimientos de seguridad y verificabilidad que debe cumplir un sistema de votación electrónica. En la Tabla I se puede ver un resumen de cómo la *blockchain* puede contribuir a estas propiedades.

Además de las propiedades resumidas en la tabla, también merece la pena destacar para los proyectos que usan transacciones con criptomonedas como sistema de votación, que por la naturaleza pública y sin permisos de la mayoría de criptomonedas, éstas están siendo atacadas continuamente y numerosos estudios analizan las vulnerabilidades de estas *blockchains* y describen los ataques descubiertos [16], lo que

[5]Scytl, https://www.scytl.com

[6]TIVI, http://www.smartmatic.com/voting/online-voting-tivi/

resulta en sistemas altamente probados y mucho más seguros que sistemas privados y más cerrados.

Sobre todos los proyectos estudiados, también cabe destacar la voluntad de hacer pública y accesible gran parte de la información del proyecto, licenciar el código como *open source*, en muchos casos haciéndolo fácilmente accesible en repositorios como Github[7]. Sin embargo, en la mayoría de proyectos, a la hora de la verdad, se echa en falta documentación clara, revisada por fuentes externas (*e.g.* artículos revisados por pares o auditorías) y certificada especificando claramente las funciones de los elementos tecnológicos utilizados. En cambio, generalmente lo único que se ofrece son *white papers* o *brochures* donde se explica de forma muy superficial la arquitectura de los sistemas. Esto requiere que cualquier evaluación del sistema tenga que hacerse revisando el código fuente directamente, o con poca asistencia documental, dificultando el estudio de la seguridad de estos proyectos. En este sentido, los proyectos que utilizan criptomonedas como base tecnológica tienen la ventaja que, en el caso de usar criptomonedas populares, ya se cuenta con multitud de expertos que conocen sus propiedades de seguridad y no tienen que ser auditadas desde zero.

Sin embargo, el uso de *blockchains* de las criptomonedas más populares, ya sea como *token* o simplemente para registrar datos como pruebas de integridad, también trae consigo algunos inconvenientes. Para empezar, el valor monetario de los *tokens* supone un riesgo de que los votantes intenten robar el *token* en vez de utilizarlo para votar. También la volatilidad de los precios del *token* puede suponer un riesgo para la viabilidad de algunos comicios, con el posible agravio de que ciertos actores puedan manipular los precios de los criptoactivos [17]. Posibles soluciones a estos inconvenientes pasan por el uso de *smart contracts* que impidan a los votantes o cualquier otra parte implicada en el proceso poder controlar las transferencias con las criptomonedas, sino que las criptomonedas deben ser transferidas directamente de monederos controlados por la junta electoral a direcciones representando al votante y de allí automáticamente a otras direcciones controladas por los candidatos o por la misma junta electoral. También estos problemas pueden mitigarse con el uso de las unidades más pequeñas de una criptomoneda (*i.e. satoshi* en el caso de bitcoin o *zatoshi* en el caso de Zcash) que suelen tener precios muy bajos, con lo que se mitiga el riesgo de robo de estos *tokens*. Sin embargo, en muchos casos no es menospreciable ni el precio de las comisiones por transacción, ni la posible demora entre la emisión de una transacción y su inclusión en un bloque (sin contar el tiempo requerido para que se minen el suficiente número de bloques para poder considerar una transacción como confirmada, *i.e.* 6 bloques en Bitcoin que equivalen a un mínimo de 1 hora de tiempo). En el caso de Bitcoin, las comisiones llegaron a situarse alrededor de los 55 USD de media el 22 de diciembre de 2017[8], y el 12 de noviembre del 2017, de mediana las transacciones tardaban 27 minutos en ser aceptadas en un bloque minado[9].

Además, estos proyectos deben demostrar que los ataques conocidos sobre las *blockchain* no afectarán a un proceso electoral. Entre ellos destaca el ataque del 51 % [16], donde un atacante que reuniera la suficiente fuerza de computación podría manipular la cadena de bloques, por ejemplo eliminando transacciones (lo que supondría la eliminación de votos en este contexto). Por esto es necesario estudiar la resistencia de estas *blockchains* para su utilización en un proceso electoral, donde los adversarios pueden llegar a ser estados poderosos con un poder computacional muy elevado. Esto hace que a nivel de seguridad, las *blockchain* con más capacidad de computación (medido por el *hashing rate*, número de *hashes* por segundo que computa de forma global una red de una criptomoneda) sean más seguras, y por lo tanto, el uso de criptomonedas minoritarias sea totalmente desaconsejable. A día de hoy (abril de 2018), el *hashrate* de Bitcoin está alrededor de 27 EH/s[10] y el de Ethereum alrededor de 250 TH/s[11]. Bastante inferior es el *hashrate* de Zcash, que dispone alrededor de 475 MH/s[12].

De los proyectos analizados, ninguno publica un estudio

[7]Github, https://github.com/

Tabla I
RESUMEN DE LAS CONTRIBUCIONES EN LAS PROPIEDADES DE SEGURIDAD DE LOS SISTEMAS DE VOTACIÓN ELECTRÓNICA CON EL USO DE TECNOLOGÍA BLOCKCHAIN EN FORMA DE (1) TRANSACCIONES CON MONEDAS, Y (2) REGISTRO INMUTABLE.

Elegibilidad	
(1)	No mejora los mecanismos existentes. Por otro lado, algunos de los proyectos proponen el uso de mecanismos descentralizados también utilizando *blockchain* para poder identificar a los usuarios, como por ejemplo el propuesto en [4].
(2)	Permite la inclusión de la lista de votantes autorizados en un tablón de anuncios público.
Autenticidad e integridad	
(1)	*Blockchain* asegura la autenticidad y la integridad de sus transacciones.
(2)	El registro de pruebas de integridad en una *blockchain* asegura su inmutabilidad y protege a los sistemas de ataques de modificación y truncado en caso de exposición de claves privadas.
Seguridad en general	
(1)	El sistema evita inherentemente que un votante vote dos veces gracias a que *blockchain* evita *double spending*.
(2)	El sistema facilita el registro de pruebas de integridad como capturas del estado del sistema cada cierto tiempo.
Privacidad de los votantes	
(1)	Algunos incorporan mecanismos para anonimizar a los usuarios. Esto hace que servicios adicionales como las redes de mezclado no sean necesarios.
(2)	No mejora los mecanismos existentes.
Secreto de los resultados intermedios	
(1)	Algunos incorporan mecanismos multi autoridad que pueden contribuir a no realizar transacciones hasta que la votación haya concluído.
(2)	No mejora los mecanismos existentes.
Incoercibilidad	
(1)	No mejora los mecanismos existentes.
(2)	No mejora los mecanismos existentes.
Verificabilidad	
(1)	Las *blockchain* aportan transparencia. Las transacciones en la *blockchain* son públicas y por lo tanto, el rastreo de los *tokens* puede ayudar a conseguir verificabilidad E2E.
(2)	La *blockchain* sirve de registro donde publicar las pruebas que aseguran que el proceso ha funcionado correctamente.

[8]Bitcoininfocharts. Media del coste de las comisiones por transacción., https://bitinfocharts.com/comparison/bitcoin-transactionfees.html

[9]Blockchain.info. Mediana del tiempo de confirmación de una transacción., https://blockchain.info/charts/median-confirmation-time

[10]Bitcoininfocharts. *Hashrate* de Bitcoin., https://bitinfocharts.com/comparison/bitcoin-hashrate.html

[11]Bitcoininfocharts. *Hashrate* de Ethereum., https://bitinfocharts.com/comparison/ethereum-hashrate.html

[12]Bitcoininfocharts. *Hashrate* de Zcash., https://bitinfocharts.com/comparison/zcash-hashrate.html

en profundidad de cómo afectarían los principales ataques conocidos a los ecosistemas *blockchain*. De hecho, tan solo los autores de [10] mencionan que algunos de los ataques podrían suponer un problema.

A todo esto hay que añadirle los problemas de escalabilidad de las principales criptomonedas [19]. Aunque gracias al despliegue de SegWit[13] en la red principal de Bitcoin, o al incremento del tamaño del bloque en Bitcoin Cash[14], se haya incrementado ligeramente el máximo de 7 transacciones por segundo que tenía Bitcoin [19], actualmente las principales *blockchains* todavía tienen un límite que se sitúa lejos del que debería poder soportar una plataforma de votación electrónica. En los proyectos analizados, esta limitación es o bien ignorada, o se utiliza un sistema de menor latencia para el registro de las acciones que requieran un *throughput* alto, como por ejemplo en Agora, donde se usa una *blockchain* con permisos. Otra solución pasaría por desarrollos de capas por encima de las *blockchain*, como las Lightining Network[15], las cuales prometen incrementar de forma exponencial la capacidad de las *blockchain*.

Además, estos proyectos también tienen que demostrar que en un futuro los datos almacenados en las *blockchain* públicas no podrán usarse para revelar la identidad de los votantes. Aunque en estos momentos parece un riesgo lejano, es necesario evaluar los riesgos que puede tener la computación cuántica en un futuro, ya que es conocido que la seguridad de algunas técnicas criptográficas podrían romperse [18]. También es necesario considerar que pueden aparecer fallos no descubiertos que lleven a identificar a los usuarios. Por ejemplo, los autores de [20] recientemente han comprobado empíricamente que es posible vincular varias transacciones de Monero[16], una moneda que tiene como principal característica ser anónima y no linkable.

V. Conclusiones

En este artículo hemos estudiado varios proyectos que proponen diferentes usos de *blockchain* dentro de un contexto de votación electrónica. Hemos visto que en general, estos proyectos pueden dividirse entre los que pretenden utilizar un *token* criptográfico como *token* de votación, y por otro lado los que básicamente utilizan *blockchain* tan sólo como un registro inmutable. Los primeros son proyectos ambiciosos y todavía tienen que solventar muchos problemas tecnológicos (*e.g.* escalabilidad de *blockchain*) para que se pueda plantear seriamente su uso en elecciones oficiales. Por otro lado, los segundos plantean mejorar alguno o varios de los elementos dentro de un sistema de votación electrónica convencional, basando la mayor parte de la seguridad del sistema en los principios ya conocidos, como sistemas multi autoridad, redes de mezclado, etc. implementados completamente a parte de la *blockchain*, por lo que parece que el posible uso de estos sistemas de votación electrónica esté más cerca. Sin embargo, la utilización de un *token* de votación plantea un cambio de paradigma mayor, que podría suponer la creación de plataformas de votación electrónica más rentables, de más fácil despliegue, soportando nuevas formas de participación democrática.

[13]Segregated Witness, https://github.com/bitcoin/bips/blob/master/bip-0141.mediawiki

[14]Bitcoin Cash, https://www.bitcoincash.org/

[15]Lightining Network, https://lightning.network/

[16]Monero, https://getmonero.org/

Agradecimientos

Este trabajo está financiado parcialmente por el Ministerio de Economía y Competitividad a través del proyecto SMART-GLACIS (TIN2014-57364-C2-2-R).

Referencias

[1] J. Bannet, D. Price, W. Rudys, J. Singer y D.S. Wallach: "Hack-a-vote: Security issues with electronic voting systems", en *IEEE Security & Privacy*, vol. 2, n. 1, pp. 32-37, 2004.

[2] M.P. López: "El desdichado voto exterior" [Online], 2017. Disponible: LaVanguardia.com, http://www.lavanguardia.com/politica/20171221/433787580695/voto-extranjero-expatriados-documentacion-electoral-elecciones-cataluna-21d.html. [Accedido 9 de Abril, 2018].

[3] C. Famighetti: "Long voting lines: explained" [Online], 2016. Disponible: BrennanCenter.org, https://www.brennancenter.org/analysis/long-voting-lines-explained. [Accedido 9 de Abril, 2018].

[4] Democracy.earth: "The Social Smart Contract" [Online], 2018. Disponible: Democracy.earth, https://www.democracy.earth/#paper. [Accedido 9 de Abril, 2018].

[5] D.Y. Marcos del Blanco, L. Panizo y J.A. Hermida: "Review of cryptographic schemes applied to remote electronic voting systems: remaining challenges and the upcoming post-quantum paradigm", en *Open Mathematics*, vol. 16, n. 1, pp. 95-112, 2018.

[6] S. Nakamoto: "Bitcoin: A peer-to-peer electronic cash system" [Online], 2008. Disponible: Bitcoin.org, https://bitcoin.org/bitcoin.pdf. [Accedido 9 de Abril, 2018].

[7] N. van Saberhagen: "Cryptonote v 2. 0" [Online], 2013. Disponible: Cryptonote.org, https://cryptonote.org/whitepaper.pdf. [Accedido 9 de Abril, 2018].

[8] E.B. Sasson, A. Chiesa, C. Garman, M. Green, I. Miers, E. Tromer y M. Virza: "Zerocash: Decentralized anonymous payments from bitcoin", en *IEEE Symposium on Security and Privacy (SP)*, pp. 459-474, 2014.

[9] M. Crosby, P. Pattanayak y S. Verma y V. Kalyanaraman: "Blockchain technology: Beyond bitcoin", en *Applied Innovation*, vol. 2, pp. 6-10, 2016.

[10] P. Tarasov y H. Tewari: "The future of e-voting", en *Int. Journal on C.S. & I.S.*, vol. 12, n. 2, pp. 148-165, 2017.

[11] Agora: "Agora 0.2" [Online]. Disponible: Agora.vote, https://agora.vote/Agora_Whitepaper_v0.2.pdf. [Accedido 9 de Abril, 2018].

[12] K. Nikitin, E. Kokoris-Kogias, P. Jovanovic, N. Gailly, L. Gasser, I. Khoffi, J. Cappos y B. Ford: "CHAINIAC: Proactive software-update transparency via collectively signed skipchains and verified builds", en *26th USENIX Security Symposium*, pp. 1271-1287, 2017.

[13] J. Cucurull y J. Puiggalí : "Distributed immutabilization of secure logs", en *Int. Work. on Security and Trust Management*, pp. 122–137, 2016.

[14] Smartmatic y Cibernetica: "TIVI. Successfully solving the challenges." [Online]. Disponible: Smartmatic.com, http://www.smartmatic.com/fileadmin/user_upload/Whitepaper_Online_Voting_Challenge_Considerations_TIVI.pdf. [Accedido 9 de Abril, 2018].

[15] Smartmatic y Cibernetica: "TIVI. Verifiable voting. Accessible, anytime, anywhere." [Online]. Disponible: Smartmatic.com, http://www.smartmatic.com/fileadmin/user_upload/Smartmatic_Cybernetica_RemoteVoting.pdf. [Accedido 9 de Abril, 2018].

[16] M. Conti, C. Lal y S. Ruj: "A survey on security and privacy issues of bitcoin", en *arXiv preprint arXiv:1706.00916*, 2017.

[17] N. Gandal, J.T. Hamrick, T. Moore, y T. Oberman: "Price manipulation in the Bitcoin ecosystem", en *Journal of Monetary Economics*, 2018.

[18] J. Proos y C. Zalka: "Shor's discrete logarithm quantum algorithm for elliptic curves", en *Quantum Information & Computation*, vol. 3, n. 4, pp. 317-344 , 2003.

[19] K. Croman, C. Decker, I. Eyal, A.E. Gencer, A. Juels, A. Kosba, A. Miller, P. Saxena, E. Shi, E.G. Sirer y otros: "On scaling decentralized blockchains", en *Int. Conf. on Financial Cryptography and Data Security*, pp. 106-125, 2016.

[20] A. Miller, M. Möser, K. Lee, K. y A.Narayanan: "An empirical analysis of linkability in the Monero blockchain", en *arXiv preprint arXiv:1704.04299*, 2017.

Fidelización Mediante CryptoTokens

M. Magdalena Payeras-Capellà
Dpt. de C. Matemàtiques i Informàtica
Universitat de les Illes Balears,
Ctra. Valldemossa, km7,5 07122 Palma
mpayeras@uib.es

Macià Mut-Puigserver
Dpt. de C. Matemàtiques i Informàtica
Universitat de les Illes Balears,
Ctra. Valldemossa, km7,5 07122 Palma
macia.mut@uib.es

Llorenç Huguet-Rotger
Dpt. de C. Matemàtiques i Informàtica
Universitat de les Illes Balears,
Ctra. Valldemossa, km7,5 07122 Palma
l.huguet@uib.es

***Resumen*—Blockchain permite redefinir los programas de fidelización resolviendo algunos de los principales problemas de los sistemas actuales. En los últimos meses han ido apareciendo diversas propuestas de tokens de fidelización. Este artículo presenta un estudio sobre las aportaciones de la blockchain a los programas de fidelización así como sobre las técnicas utilizadas para el desarrollo de los cryptotokens. Se han evaluado las alternativas en diferentes aspectos como los mecanismos de consenso, la definición de los términos contractuales o el formato de los tokens, permitiendo obtener conclusiones sobre los criterios de diseño adecuados.**

I. Introducción

Un programa de fidelización es una estrategia establecida por una empresa con el propósito de premiar el comportamiento de compra de sus clientes, lo que produce en ellos un sentido de lealtad y fidelidad hacia la empresa. Fueron introducidos por American Airlines y Marriot en los años 80 [1]. Aunque tradicionalmente se han vinculado con sectores como las compañías aéreas u otros sectores relacionados con el turismo, como hoteles o compañías de alquiler de vehículos, actualmente se encuentran muy extendidos y ejemplos de ello son los programas de puntos asociados a las tarjetas de crédito.

I-A. Tipos de Programas de Fidelización

Podemos dividir los programas en tipos según criterios diferentes. Una primera distinción podría ser según las recompensas que ofrecen, a continuación en función del tipo de soporte, según la cantidad de establecimientos donde se puede utilizar y, finalmente, según el tipo de comercio emisor.

Según las recompensas ofrecidas, los programas de fidelización se podrían clasificar en las siguientes categorías:

1. **Programas monetarios.** Son programas que ofrecen a sus clientes recompensas monetarias en forma de descuentos o de un determinado porcentaje del dinero invertido en sus compras.
2. **Programas de puntos.** Por su estrategia de fidelización, son los más comunes. Se basan en la acumulación de puntos a lo largo del tiempo. Su funcionamiento es sencillo: un miembro del programa acumula puntos por realizar determinadas acciones en la plataforma y los canjea por opciones o recompensas. En algunos casos, los puntos acumulados tienen una caducidad, por lo que si una vez superado este tiempo no se han gastado, el cliente pierde el derecho a utilizarlos.
3. **Programas de ventajas.** Este tipo de programas ofrecen a sus clientes ventajas que sólo podrán conseguir a través de ellos. Algunas de las ventajas que se suelen ofrecer son: financiaciones muy beneficiosas de las compras realizadas, cajas preferentes a la hora de comprar, aumento del período de devolución de los productos, posibilidad de realizar devoluciones sin disponer del ticket de compra, protección de compra o aumento de garantía, invitaciones a eventos de la marca, etc.

Por su formato o tipo de soporte podemos diferenciar entre programas de fidelización que se desarrollan a través de una tarjeta física; otros que se basan en una plataforma online o aplicaciones móviles mientras que otros presentan características mixtas.

Según número de establecimientos de uso encontramos:

1. Uso únicamente en tiendas y comercios de la marca. El cliente sólo podrá hacer uso al realizar compras en un establecimiento.
2. Uso en más tiendas y comercios. En este caso, además de poder participar en el programa en comercios de la marca emisora también se permite hacerlo en establecimientos asociados, incrementando así las posibilidades de uso.

I-B. Características deseables

Los diferentes actores que intervienen en los programas de fidelización serían:

- **Usuario** Obtiene la tarjeta física o las credenciales de fidelización, las cuales presenta para validación y obtención de los beneficios concretos (descuentos, promociones, puntos, etc.).
- **Emisor.** Emite la tarjeta de fidelización o las credenciales. Pueden ser emitidas por el proveedor o por un intermediario. Se incluye una descripción del servicio ofrecido y los términos y condiciones de utilización, que pueden tener forma de contrato.
- **Proveedor del servicio.** Se encarga de validar la tarjeta de fidelización o las credenciales y, en caso de validación satisfactoria, proporciona al usuario los beneficios concretos.

A continuación se enumeran los requisitos de seguridad y privacidad de las interacciones entre estos actores.

- **Integridad.** Permite determinar si la membresía a un programa de fidelización ha sido manipulada o modificada.
- **Autenticidad.** Los usuarios deben poder verificar que el emisor es un emisor autorizado.
- **No repudio en origen.** Las emisiones de recompensas no pueden ser negadas por su emisor.
- **No repudio de recepción.** Un usuario no puede negar la recepción de las recompensas que realmente ha recibido.
- **Infalsificabilidad.** Sólo los emisores autorizados podrán emitir membresias de fidelización y recompensas válidas.
- **Equidad.** El intercambio de recompensas entre usuarios y proveedores de servicios ha de ser equitativo (cada parte recibe lo estipulado según las condiciones establecidas).
- **No sobreutilización.** Las recompensas recibidas por el usuario solo se deben poder utilizar de acuerdo con las condiciones del programa. Por ejemplo, pueden indicar que sólo se puede utilizar una vez por compra o bien una vez cada cierto tiempo.
- **No Transferibilidad.** Un usuario no debe poder transferir sus recompensas a otro para que éste las utilice en su beneficio (a menos que se especifique lo contrario en la condiciones del programa). Tampoco un usuario puede obtener recompensas de un vendedor a través de una factura hecha a otro usuario.
- **Identificación.** Si el programa de fidelización no es anónimo, la identidad del usuario debe poder ser verificada.
- **Anonimato.** Un proveedor no puede enlazar las recompensas con una compra realizada por un usuario.
- **Enlazabilidad Controlada.** Un usuario debe poder decidir si una de sus recompensas puede ser o no asociada a otras de sus recompensas.

II. Problemas de los Programas de Fidelización Actuales.

Una vez analizados los programas de fidelización se ha determinado que los principales problemas que plantean son [3]:

- **Participación**: La participación activa en los programas de fidelización se halla por debajo del 50 % y reduciéndose. En los últimos años se ha visto incrementado sustancialmente el volumen de usuarios enrolados en programas de fidelización (3,32 billones de usuarios en 2015, según COLLOQUY Loyalty census)[2]. Sin embargo, no están desarrollando todo su potencial.
- **Entusiasmo**: Más del 50 % de las recompensas acumuladas no se liquidan nunca. La falta de liquidación de recompensas genera un problema importante para las compañías ya que éstas pasan a formar parte de la deuda o pasivo de la compañía. Además, las generaciones jóvenes parecen menos interesadas en los programas de fidelización.
- **Compatibilidad**: Aunque algunos proveedores han alcanzado alianzas entre ellos ampliando las posibilidades de liquidación, existe una escasez de uniformidad de los sistemas de gestión de diferentes programas.
- **Usabilidad**: Reglas complicadas y políticas de liquidación de puntos complejas.

III. Posibles aportaciones de Blockchain a los programas de fidelización

La clasificación presentada en § I-A puede completarse con un modelo de programa de fidelización basado en blockchain. El nuevo modelo permitiría tanto la creación de programas monetarios como programas de puntos utilizando criptomonedas y criptotokens, respectivamente. Ambos modelos encajarían en la categoría de aplicaciones on-line, sin necesidad de soporte físico. Un programa de fidelización Blockchain permite la participación de los diferentes actores del sistema de fidelización (proveedores, gestores del sistema y usuarios) interactuar con el sistema sin intermediarios y sin comprometer la privacidad y la competitividad. Además, los programas permiten su uso a tiempo real y de forma transparente. Los programas basados en blockchain utilizaran Smart Contracts y Wallets, conectando un ecosistema ampliamente desconectado reduciendo costes y eliminando fricciones y permitiendo la liquidación de las recompensas casi a tiempo real y de forma segura. Para determinar las aportaciones blockchain se ha realizado un análisis de diferentes propuestas de programas de fidelización [3-24]. A través del análisis descubrimos que esta nueva tecnología aporta:

- **Tokens de Fidelidad**. Los tokens se generan mediante un algoritmo y son base para cualquier tipo de recompensa. Un conjunto de reglas y restricciones gobiernan la utilización de los tokens.
- **Conexión de Ecosistemas Distribuidos.** Blockchain permite la creación, la liquidación y el intercambio de recompensas de forma segura e instantánea entre diferentes programas, proveedores e industrias utilizando pruebas criptográficas en lugar de terceras partes de confianza gracias a los Smart Contracts. La plataforma o Red de Fidelización puede acomodar diferentes organizaciones que alcanzan el consenso sobre las transacciones sin necesidad de TTP.
- **Aplicaciones de Gestión de Recompensas.** La aplicación se halla en la wallet, punto de entrada a la plataforma. La wallet contiene la identidad en forma de clave y guarda el valor de los tokens de fidelidad. Posible plataforma única en forma de wallet para gestionar multiples programas de fidelización.
- **Reducción de Costes.** Un programa de fidelización basado en blockchain reduce los costes de gestión con Smart Contracts que registran las transacciones de forma segura reduciendo los costes asociados con errores o fraude. Además, los puntos sin utilizar dejaran de contabilizarse en los balances de los proveedores ya que residen en la red compartida. La infraestructura de recompensas queda fuera del sistema del vendedor, quien únicamente debe mantener su wallet para el envío y recepción de tokens cuando se cumplan determinadas condiciones.
- **Automatización.** Puede conseguirse mediante la programación de smart contracts que automaticen el envío de un determinado número de tokens cuando se produzca un evento determinado, como una compra.
- **Temporalidad.** Por razones técnicas y logísticas los programas tradicionales no proporcionan las recompensas a tiempo real. Blockchain permite que las transacciones se registren casi a tiempo real, mejorando la experiencia del usuario.
- **Seguridad.** Blockchain crea una base de datos inmutable y con registros de tiempo por cada transacción y este registro es fácilmente trazable por lo que previene la doble utilización, el fraude, el abuso y otros tipos de manipulaciones en las transacciones. Además, al tener los puntos en forma de token poseen identificadores únicos extremadamente difíciles de falsificar.
- **Oportunidad.** Con este nuevo escenario los proveedores que previamente no participaban en programas de fidelización podrán aprovechar las ventajas de la escala y ofrecer a sus clientes puntos de fidelización que pueden ser liquidados en una red más amplia.
- **Facilidad de formación de coaliciones.** Se puede vincular a los usuarios a más proveedores para proporcionarles una experiencia plena. Los usuarios tendrán más opciones de liquidación gracias a la temporalización y la posibilidad de programas vinculados.
- **Transparencia y Control sobre los puntos.** Mayor control del usuario sobre sus puntos. La conversión de los puntos en tokens los hace únicos, trazables y a prueba de fraude. Facilidad de localización de los puntos no liquidados. Analisis de datos más transparente que proporciona más detalles del comportamiento de los usuarios. Gestión transparente que reduce la perdida de puntos por caducidad o fraude. La funcionalidad de los smart contracts juega un papel fundamental en la habilidad de operar según lo establecido.

IV. Técnicas para el desarrollo de Cryptotokens de Fidelización

Actualmente encontramos diversas propuestas de tokens basados en blockcahin que encajan en este nuevo escenario, todas ellas de reciente creación e incluso algunas en fase de desarrollo [4], [5], [6], [7], [8], [9], [10], [16], [17], [18], [19], [20], [21], [22], [23], [24]. En ellas los cryptotokens representaran los puntos o recompensas del programa de fidelización. Algunas soluciones operan con una única empresa (Ej.: los Tokens WhopperCoin de Burger King [11]) y otras son plataformas que permiten la participación de varias empresas. Ejemplos de estas últimas son Loyyal [7], Incent Loyalty [9], Chain of Points [4], RWRD [5], Loyalty Coin [6], Blockpoint [8], Elements [10], TokenLoyalty [16], LoyalCoin [17], TopCoin [18], SandBlock [19] o Bitrewards [21].

En esta sección se describe el desarrollo de un programa de fidelización mediante blockchain. Se describen algunas de las tendencias y técnicas utilizadas para el desarrollo de programas de fidelzación mediante cryptotokens más utilizadas, novedosas y/o interesantes al tiempo que se analiza su conveniencia.

IV-A. Elección de la plataforma

El primer paso del desarrollo del sistema es la elección de una plataforma adecuada al programa que se va a desarrollar.

WAVES es una plataforma de blockchain descentralizada que se centra en los tokens personalizados de blockchain que pueden ser creados por los usuarios. La plataforma de WAVES permite almacenar, comercializar, gestionar y emitir activos

digitales de forma fácil y segura. Las características principales y ventajas de WAVES son: alta capacidad de procesamiento de red, baja latencia y bajas tarifas de transacción. WAVES tiene una alta velocidad de red con cientos de transacciones por segundo y comisiones pequeñas, lo que hace que sea muy adecuada para transferir criptomonedas entre los miembros de la red al instante. La plataforma de WAVES funciona para crowdfunding, programas de fidelización, votación,...

WAVES utiliza el Leased Proof-of-Stake (LPoS) como algoritmo de consenso. LPoS es una versión mejorada de Proof-of-Stake que posee las siguientes especificaciones: tiempos de bloque de un minuto, gastos de transacción de 0.001 WAVES y velocidades de hasta 1,000 txs/s

Por ejemplo, TokenLoyalty [16] utiliza la blockchain Waves. En su white paper se enumeran las razones:

1. Se puede establecer un lanzamiento de tokens máximo de 0.28 % en la fase ICO. Proporciona la oportunidad de informar a la comunidad Waves sobre el proyecto.
2. Las wallets de Waves son fáciles de usar.
3. Permite DEX sobre Wave. Es decir, es posible realizar intercambios descentralizados a coste cero o con tasas bajas. Es interensante si el modelo permite un intercambio *social* de tokens entre usuarios.
4. Hay usuarios estables que participan activamente en las ICOs.
5. Existe una gran comunidad emergente.
6. Tarifas internas bajas que son cruciales para el proyecto.

Otros programas de fidelización, como LoyalCoin [17] se han decantado por NEM. La principal contribución de NEM al panorama de criptomonedas es un nuevo mecanismo de consenso llamado Prueba de Importancia (PoI). A diferencia de la Prueba de trabajo (PoW), PoI es sostenible desde el punto de vista medioambiental y no requiere recursos informáticos a gran escala a perpetuidad. PoI es similar al proof-of-Stake (PoS), excepto que no se deriva únicamente del tamaño del saldo de una cuenta. De esta forma, intenta recompensar a los participantes economicamente activos a expensas de los inactivos y amortigua los efectos negativos inherentes a PoS.
ETHEREUM es la plataforma escogida por la mayoria de las propuestas de programas de fidelización, entre ellos TopCoin [18], SandBlock [19], EZtoken [23] y la aplicación RWRD, del programa RewardsToken [5]. RWRD destaca de Ethereum:

- Plataforma probada. Ethereum ha sido sometido a un tremendo uso y pruebas, y ha demostrado ser seguro y robusto.
- Herramientas de desarrollo. Ethereum cuenta con excelentes herramientas de desarrollo que la comunidad actualiza activamente. Uno de estos frameworks es Truffle, que se ha utilizado para el desarrollo del proyecto RWRD.
- Existencia de smart contracts reutilizables y bien desarrollados. Hay muchos repositorios de smart contracts de Ethereum que ya han sido desarrollados y probados. El token RWRD reutiliza los contratos Smart OpenZeppelin, probado y auditado por ingenieros experimentados.
- Turing Complete Smart Contracts. La aplicación de token RWRD consistirá en contratos de votación que tendrán funcionalidad personalizada que no podría desarrollarse en otras plataformas de contratos inteligentes

Estas no serían las únicas posibilidades, por ejemplo BioCoin [22] se basa en Peercoin. Mediante los ejemplos expuestos puede observarse que existe una gran diversidad a la hora de realizar la elección de la plataforma blockchain sobre la que desarrollar el programa, aunque priman las blockchain que permiten la ejecución de smart contracts.

IV-B. Generación y Liquidación de los Cryptotokens

El uso de cryptotokens para representar los puntos de fidelidad en lugar de utilizar cryptomonedas se fundamenta en el hecho de que estos puntos pueden no representar un valor monetario sino ser canjeados por la recompensa indicada en el programa de fidelización. La mayoria de propuestas presentan mecanismos de generación y liquidación similares, basados en la creación, transferencia y liquidación de cryptotokens. Sin embargo podemos encontrar propuestas que difieren del modelo general. Por ejemplo, para favorecer la funcionalidad del programa de fidelización entre diferentes empresas se puede optar por un mecanismo de recompensas en modo indirecto, como es el caso de Sandblock [19]. Por otra parte, la liquidación mediante tarjeta de crédito es una particularidad presente en [20]. Ambas propuestas se describen a continuación.

IV-B1. Modo Indirecto: El aspecto más innovador de creación de tokens de recompensa es el modo indirecto y lo encontramos en programas de fidelización como SandBlock [19]. El protocolo central del proyecto Sandblock se denomina Protocolo de Satisfacción. Este protocolo está descentralizado y se basa en smart contracts de Ethereum. El Protocolo de Satisfacción maneja las atribuciones de recompensas y las diversas reglas de comerciante. Los tokens de satisfacción (SAT) son una criptomoneda negociable ERC20 en la blockchian de Ethereum. El SAT se creará durante un crowdsale mientras que los tokens de marca (Branded Tokens) describen un token creado por un comerciante utilizando el Protocolo de Satisfacción. Este es un nombre genérico ya que los comerciantes tendrán que elegir un nombre para su moneda. Los tokens de marca están vinculados al SAT en una proporción definida en su creación. El contrato de tokens de marca solo permite que se envíen a direcciones predefinidas (wallets del comerciante), y a través de la transferencia de igual a igual manejada por el protocolo de satisfacción por la misma cantidad de SAT. Para crear sus tokens, el comerciante deberá bloquear los SAT. Los tokens de marca, por lo tanto, no son ERC20 y no son intercambiables en los exchanges.

IV-B2. Liquidación mediante tarjeta bancaria: Otro punto a tratar es como se puede integrar la liquidación de los tokens de fidelidad en el entorno actual de pagos y compras, especialmente en los establecimientos tradicioanales. [20] evalua esta situación y propone un sistema de liquidación compatible con los pagos mediante tarjeta de crédito, ya que no considera realista la instalación de un nuevo terminal de pago en una tienda físca. Por lo tanto, este estudio propone que la implementación y el mantenimiento de la plataforma blockchain deben establecer alianzas estratégicas con proveedores de redes, como gestores de tarjetas de crédito.
El proceso de liquidación de los tokens seria el siguiente:

1. El usuario de la tarjeta decide hacer uso del token de fidelidad como medio de pago a través de la realización de un pago mediante tarjeta de crédito, y el vendedeor carga a la compañía de la tarjeta de crédito.
2. La compañía de la tarjeta paga el efectivo al vendedor.
3. La compañía de tarjeta reclama y recibe el token del cliente.
4. El token recibido por la compañía de la tarjeta de crédito es vendido en el mercado de intercambio.
5. La compañía de la tarjeta de crédito carga y recibe efectivo de la cuenta bancaria del cliente en la fecha de pago de la tarjeta de crédito del cliente.

Los vendedores proporcionan a los clientes tokens que son parte de la cantidad de dinero que han usado de acuerdo con las políticas de su programa de fidelización. Este modelo integrado de tokens permitirá a los vendedores eliminar la inconveniencia de tener que asociar un programa de fidelización entre diversos vendedores como lo están haciendo ahora o la inconveniencia de tener que usar tokens de fidelización solo en ciertos vendedores.

IV-C. Formato Estandarizado de Tokens

En esta apartado se trataran los formatos estandarizados para token de fidelización sobre Ethereum, al tratarse del grupo más numeroso.

Teóricamente, los smart contracts en la blockchain de Ethereum pueden diseñarse arbitrariamente si su programación en Solidity es correcta. Sin embargo, para que el software de terceros tenga fácil acceso al smart contract, se han decidido un conjunto de convenciones para la creación de tokens que incluyen variables, funciones y sus nombres. Se recomienda encarecidamente a los desarrolladores de tokens de Ethereum que sigan este conjunto de convenciones, de lo contrario sus tokens podrían no ser reconocidos por wallets Ethereum.
El estándar de tokens ERC20 (Ethereum Request for Comments #20) describe las funciones y eventos que un contrato de token Ethereum debe implementar. En la generación del token se utilizan formatos estándares que incluyen las principales funciones asociadas al token. ERC20 garantiza que el token creado tendrá una transferibilidad simple entre wallets y contratos. Hay tres funciones: 'transfer', 'transferFrom' y 'allow' que deben existir en cada contrato ERC20. Se puede usar 'transfer' para enviar sus tokens y si se desea enviar tokens a un contrato, el contrato primero debe permitir al pagador realizar la operación 'transferFrom' desde el token de ese contrato. El estándar ERC-20 es utilizado por diversos tokens de fidelización, como el caso de RWRD [5] y del token BIT del programa BitRewards [21]. En la figura 1 aparecen

```
// ----------------------------------------------------------------------------
// ERC Token Standard #20 Interface
// https://github.com/ethereum/EIPs/blob/master/EIPS/eip-20-token-standard.md
// ----------------------------------------------------------------------------
contract ERC20Interface {
    function totalSupply() public constant returns (uint);
    function balanceOf(address tokenOwner) public constant returns (uint balance);
    function allowance(address tokenOwner, address spender) public constant returns (uint remaining);
    function transfer(address to, uint tokens) public returns (bool success);
    function approve(address spender, uint tokens) public returns (bool success);
    function transferFrom(address from, address to, uint tokens) public returns (bool success);

    event Transfer(address indexed from, address indexed to, uint tokens);
    event Approval(address indexed tokenOwner, address indexed spender, uint tokens);
}
```

Figura 1. Estándard ERC-20

sus funciones y eventos, pudiendose destacar también, además de la función 'transfer', que envia una cantidad de tokens a una dirección de destino, la función 'balanceOf' que proporciona el balance de la dirección indicada y el evento 'Transfer' que se activa cuando se transfieren tokens.

Sin embargo, ERC20 presenta un problema. Cuando un usuario envía un token ERC20 a un contrato de Ethereum que no reconoce los tokens ERC20, el usuario pierde el acceso a sus fondos para siempre. Estos contratos no esperaban recibir ningún token ERC20, por lo que cuando los usuarios envían tokens a estas direcciones, la red confirma las transacciones; sin embargo, el contrato receptor no reconoce los tokens. No sabe qué hacer con estos tokens, lo que hace que los fondos se bloqueen para siempre. Los tokens no son rechazados, simplemente son ignorados completamente por el contrato receptor. A fecha de 31 diciembre de 2017 se han producido pérdidas por valor de 3 millones de dólares debido a éste problema. La mayoría de estas transacciones son cometidas involuntariamente por usuarios finales que llaman a la función de transferencia. ERC20 usa 'Transfer' y 'TransferFrom', y resulta que algunos usuarios están usando 'Transfer' para enviar directamente tokens ERC20 a contratos que no los están esperando y, por tanto, no reconocen los tokens entrantes.

Finalmente, algunos miembros de la comunidad de Ethereum decidieron abordar este problema solicitando un nuevo estándar de token de ERC. El número de emisión de este nuevo estándar de tokens en GitHub es el número 223. Los tokens ERC223 son retrocompatibles con los tokens ERC20, por lo que es compatible con todas las funciones de ERC20 y los contratos o servicios que funcionan con tokens ERC20 funcionarán correctamente con los tokens ERC223.

Una transacción con token ERC20 entre una dirección regular no asociada a un contrato y un contrato son dos transacciones diferentes: debe hacer una llamada de aprobación en el contrato de token y luego invocar 'transferFrom' en el otro contrato cuando desee depositar sus tokens en él. ERC223 simplifica este requisito y permite el uso de la misma función de transferencia. Los tokens ERC223 se pueden enviar llamando a la función de transferencia en el contrato de token sin diferencias si el receptor es un contrato o una dirección de wallet, ya que hay una nueva forma de notificar al contrato de recepción de la transferencia. Si el receptor es una dirección regular no asociada a un contrato, una transferencia de token ERC223 será lo mismo que una transferencia ERC20. Por otro lado, si el receptor es un contrato, entonces el contrato del token ERC223 intentará llamar a la función 'tokenFallback' en el contrato del receptor. Si no hay una función 'tokenFallback' en el contrato del receptor, la transacción fallará. Ya no se obliga a los usuarios a llamar a la aprobación en el contrato. La transacción del token se manejará automáticamente dentro del contrato de intercambio, a través de la función 'tokenFallback'.

Ventajas de ERC223:

- Elimina el problema de los tokens perdidos durante la transferencia de tokens ERC20 a un contrato. ERC223 permite a los usuarios enviar sus tokens a una wallet o contrato con la misma función de transferencia, eliminando así la posibilidad de confusión y pérdida de tokens.
- Permite a los usuarios enviar sus tokens a cualquier lugar con una transferencia. No hay diferencia si el receptor es un contrato o no. No es necesario aprender cómo funciona el contrato para que los usuarios habituales envíen tokens.
- Permite a los desarrolladores manejar transacciones de token entrantes y rechazar tokens no admitidos (no es posible con ERC20).
- Ahorros de energía. La transferencia de tokens ERC223 a un contrato es un proceso de un paso en lugar de un proceso de 2 pasos (para ERC20), y esto significa 2 veces menos gas y menor sobrecarga de la cadena de bloques.
- Permite depositar tokens en contrato con una sola transacción.
- Evita la carga adicional de la cadena de bloques.
- Realiza transacciones con tokens similares a las transacciones Ether.

El nuevo estándar ERC223 ya se implementa en programas de fidelización como es el caso del token TopCoin [18]. Una vez vistas las ventajas es preferible utilizar ERC223 para los tokens de fidelización sobretodo en entornos complejos como los programas de fidelización multimarca.

IV-D. Términos y Condiciones

Todo programa de fidelización contiene un conjunto de términos y condiciones que regulan el uso de las recompensas. Las recompensas pueden expresarse mediante puntos representados por los tokens con unas determinadas reglas de uso, pero también se pueden definir recompensas en forma de promesas que deberan explicitarse con detalle (fechas de caducidad, políticas de degradación de los puntos, reglas de liquidación y uso, cesiones o transferencias,...). En los programas sobre blockchain se pueden usar Smart Contracts y Ricardian Contracts para hacer referencia a esta regulación.

IV-D1. Smart Contracts: Los contratos inteligentes, al igual que los contratos tradicionales, representan un acuerdo formalizado entre dos o más partes para proporcionar bienes o servicios a cambio de alguna forma de consideración. Mientras que los contratos tradicionales requieren un tercero para garantizar la correcta ejecución de los términos y condiciones, los contratos inteligentes están diseñados para ser ejecutados automáticamente y aplicar programáticamente los términos y condiciones acordados por las diferentes partes utilizando la tecnología blockchain. En otras palabras, usando Smart Contracts, los términos contractuales se pueden traducir en funciones lógicas que se activan automáticamente cuando se cumplen las condiciones definidas. Los smart contracts son ampliamente utilizados para garantizar el correcto funcionamiento de los programas de fidelización, desde la creación de los tokens, su liquidación e incluso auditoria. Un ejemplo de uso lo encontramos en RWRD [5], que utiliza hasta cinco smart contracts diferentes: *Token Contract, Token Distribution Contract, Community Voting Contract, Bonus and Promotions Contract y IPFS Contract*. El *Contract Token* es un contrato tipo ERC20 que define el token RWRD. *Token Distribution Contract* se usará para vender los tokens RWRD al público en diferentes etapas. Los propietarios de tokens podrán votar propuestas a través del *Community Voting Contract*, mientras que el *IPFS Contract* se utiliza para auditar la información del ledger a través de la publicación en IPFS. El hash se envía al contrato y permite verificar la integridad de los datos.

IV-D2. Ricardian Contracts: Un Ricardian contract puede definirse [13] como un acuerdo en forma de un documento único que es presentado por un emisor a un propietario, donde el derecho de propiedad del propietario es gestionado por el emisor del documento. El documento debe ser claramente legible, al igual que los contratos tradicionales en papel. Además, el acuerdo debe ser legible por programas y software y debe estar firmado digitalmente, usando claves e información del servidor, en combinación con un método de identificación único y seguro. Un Ricardian contract incluye toda la información del documento legal en un formato que puede ser ejecutado por software. De esta manera, es tanto un acuerdo legal como un protocolo que integra el acuerdo de forma segura dentro de una infraestructura digital, mientras tanto ofrece un alto nivel de seguridad debido a la identificación criptográfica. Son legibles tanto por máquinas como por personas. Las máquinas leen el contrato dentro de las infraestructuras digitales y las personas las leen como un documento de texto sin formato. Esto permite que todas las partes de un contrato procesen adecuadamente la información allí establecida. Este tipo de contrato entrelaza los lenguajes, lo que hace que los costos de transacción sean mucho más bajos: puede contribuir a una resolución más rápida de las disputas y los acuerdos se pueden ejecutar más fácilmente. Una característica especial de un contrato ricardiano y su tecnología blockchain es su transparencia. Los riesgos de fraude pueden reducirse al mínimo porque las publicaciones de, y las referencias a, esos datos específicos se realizan usando hash criptográficos. El programa de fidelización Chain-of-points [4] utiliza contratos ricardianos. Las recompensas de este sistema, POINTS, pueden contener no solo valor sino también promesas. Inicialmente el proveedor posee en token con el que quiere recompensar al usuario. En lugar de enviar únicamente el token, se procede a recompensarlo de un modo más estimulante. El proveedor bloquea el token adjuntándole una condición para que el usuario solo pueda utilizarlo en el servicio definido. Cuándo el usuario solicita el servicio prometido firma una transacción confirmando que ha recibido la recompensa prometida. En ese momento el token se desbloquea y vuelve al proveedor, quien puede utilizarlo de nuevo. El algoritmo de firma utilizada es la firma elíptica ECDSA [12]. Por otra parte, si el usuario no utiliza el token reclamando el servicio que el proveedor le ofrece todavía puede ganar algo por su fidelidad. En el momento del bloqueo del token se determina un intervalo de tiempo en el que el token será válido. Después de su caducidad sin reclamar, el valor del token se distribuye según un porcentaje acordado, Ej: 50 %-50 %. El proveedor determinara en que servicios puede usarse este remanente de tokens.

IV-E. Mecanismos de consenso

Dentro del conjunto de plataformas blockchain existen y conviven diversos mecanismos de consenso. Además de los más conocidos, como PoW (Proof-of-Work, utilizado por ejemplo en Bitcoin) o PoS (Proof-of-Stake), utilizado en los programas de fidelización como BioCoin [22] y EZToken [23], o los ya mencionados LPoS o PoI, describiremos a continuación algunas alternativas que se presentan en programas de fidelización.

IV-E1. DPoS: Delegated Proof of Stake [14] es un método para lograr consenso distribuido en un sistema de criptomonedas. Es una variación de la PoS, que se desarrolló como una alternativa al sistema de PoW. En Proof of Stake, la capacidad de obtener recompensas a través del procesamiento de transacciones recae en los propietarios de monedas en lugar de los mineros. En la DPoS, los propietarios de monedas votan a los delegados que son responsables de validar las transacciones, por lo que DPoS mitiga los efectos negativos de la centralización.
La blockchain de LoyaltyCoin [6] trabaja con el algoritmo de consenso DPoS. Los usuarios que posean tokens de

LoyaltyCoin (LYC) pueden descargar un nodo LoyaltyCoin y participar en la validación de bloques. Estos usuarios seran recompensados por la comunidad asegurando que los dueños de tokens se vean motivados a contribuir en la seguridad y estabilidad de la blockchain.
Chain-of-Points [4] es otro ejemplo de programa de fidelización que utiliza la DPoS.

IV-E2. Proof-of-Loyalty: La Prueba de lealtad (PoL) presentada por Katalyst [24] permite a una plataforma blockchain definir una actividad económica que quiera incentivarse. La PoL combina principalmente 3 mecanismos:

1. Prueba de trabajo (PoW)
2. Prueba de participación (PoS)
3. Prueba de tiempo/edad (PoT)

Tendrá un enfoque especial e innovador sobre lo que puede ser considerado como "trabajo". En el contexto de la plataforma blockchain, ejemplos de "trabajo"podrían ser los siguientes: 1) Aumentar la conciencia de la plataforma blockchain 2) Eliminar el juego sucio 3) Votar/contribuir a la plataforma blockchain. 4) Hacer que más usuarios usen la plataforma 5) Obtener comerciantes como proveedores de servicios 6) Servir como explorador fuera de la cadena. Estas actividades económicas son inherentemente útiles para la plataforma blockchain más que solo proporcionar capacidades de hash. El primer caso de uso de Katalyst estaría en OToken. Una plataforma de blockchain centrada en el mercado del espectáculo. Los OTokens son generados para recompensar la realización de actividades positivas para el ecosistema blockchain, como los siguientes;

- Participación en la proof-of-loyalty, asegurando así la seguridad de la cadena de bloques.
- Recompensar a los usuarios por compartir noticias de entretenimiento.
- Recompensar a los usuarios por la ayuda de otros instalando una nueva instancia de la wallet.
- Recompensar a los usuarios por votar o comentar noticias.
- Premiar a los productores por hacer pequeños videos.
- Premiar a los usuarios por usar tokens para pagar películas,...

El "trabajo" realizado en PoL permite a las partes acumular recompensas en forma de tokens.

V. Conclusiones

En este estudio se han analizado los puntos débiles de los programas de fidelización tradicionales y se han relacionado con las potenciales aportaciones de blockchain. Como primera conclusión puede afirmarse que los programas de fidelización basados en blockchain permitirán aumentar la participación de los usuarios a través de sistemas sencillos e interoperables y utilizar las recompensas casi a tiempo real satisfaciendo las propiedades deseables, con lo que son potencialmente sustitutos de los programas de fidelización tradicionales. Son diversos los sistemas propuestos que se hallan en diferentes fases de desarrollo y/o explotación. En este artículo hemos realizado una revisión y análisis de sus principales características. Es un hecho común en todos ellos que las recompensas se implementen en forma de tokens. Las plataforma donde desplegar el programa presentan costes por comisión y tiempo de validación de bloques diferentes y junto con el formato del token de fidelización son criterios de diseño fundamentales. Se ha mostrado como una mala elección el formato ERC20 ya que puede llevar a la pérdida de tokens. El estándard ERC223 puede considerarse un formato adecuado aunque apenas es utilizado en las propuestas estudiadas. La descripción detallada de usos del token, con sus términos y condiciones también condiciona el diseño del sistema. Un token puede representar una amplia gama de recompensas, por los que el buen uso de los mismos debe garantizarse con herramientas como los smart contracts y los ricardian contracts. En los programas de fidelización pueden convivir multitud de mecanismos de consenso. El mecanismo Proof-of-Loyalty permite que el trabajo realizado para el funcionamiento de la blockchain se vincule a actividades productivas. De este estudio que presentamos aquí se deriva que son diversos los criterios de diseño que deben evaluarse para obtener programas seguros y eficientes.

Agradecimientos

Este trabajo ha sido paricalmente financiado por los proyectos AccessTur TIN2014-54945-R y ARES TIN2015-70054-REDC.

Referencias

[1] "Impact Of Blockchain Technology In Enhancing Customer Loyalty Programs In Airline Business" Udegbe, Scholastica Ebarefimia, International Journal of Innovative Research and Advanced Studies (IJIRAS) Volume 4 Issue 6, June 2017.
[2] 2017 Colloquy Loyalty Census: Discover The New Missing Link To Successful Loyalty. https://www.loyalty.com/channel/
[3] "Making blockchain real for customer loyalty rewards programs", https://www2.deloitte.com/content/dam/Deloitte/us/Documents/financial-services/us-fsi-making-blockchain-real-for-loyalty-rewards-programs.pdf
[4] "Chain if Points: Transforming Loyalty into Rewards" Chain of Points Whitepaper, Stefan Crnojevic, Irene Katzela. January 2017.
[5] "RWRD – Loyalty Rewards Token" Technical Whitepaper, Draft v.05. Todd Rowan, Linda Butcher, Derek Chapin, James Hester, Ali Agha, Nathan Hayes, Adrian Armaselu. https://rewardstoken.io/wp-content/uploads/2017/12/Rewards-Technical-Whitepaper.pdf
[6] Loyalty Coin WhitePaper. https://www.loyaltycoin.ch/publication /LoyaltyCoin-Whitepaper/V1.0.pdf
[7] LOYYAL. www.loyyal.com
[8] Blockpoint Blockchain Loyalty. http://blockpoint.io
[9] Incent Loyalty. https://www.incentloyalty.com
[10] Elements. https://theelements.io
[11] Wopper Coin. https://steemit.com/whoppercoin/@cryptocoiners/whopper-coin-made-by-burger-king
[12] The elliptic curve digital signature algorithm (ECDSA)". D. Johnson, A. Menezes, and S. Vanstone. International Journal of Information Security 1.1 (2001), pp. 36-63.
[13] The Ricardian Contract. I. Grigg, First IEEE International Workshop on Electronic Contracting. IEEE. 2004, pp. 25-31.
[14] Delegated Proof-of-Stake (DPOS). D. Larimer. http://107.170.30.182/security/delegated-proof-of-stake.php. April 3, 2014.
[15] Token Standard ERC223. https://github.com/Dexaran/ERC223-token-standard.
[16] TokenLoyalty White Paper Version 2.0 March 13th 2018. LoyaltyCoin Ltd https://www.loyaltycoin.ch/publication/LoyaltyCoin-Whitepaper_V1.0.pdf
[17] LoyalCoin, The new Loyalty Economy Whitepaper. Version 1.3, 2018. https://loyalcoin.io/whitepaper.pdf
[18] TopCoin whitepaper: blockchain loyalty platform for digital products & services. George Minardos, Steve Peterson, Ryan Chiu, Vujadin Milinovich, Junmo Lee. October 17. http://topcoin.build/wp-content/uploads/2017/11/TPC-17000-02-D3-Whitepaper-Condensed.pdf
[19] SandBlock Satisfaction Protocol White Paper. Revision 6. January 2018. F. Bascoulengue, S. Eck, J.Grenier, Kairos SAS. https://sandblock.io/whitepaper.pdf
[20] Modeling the Intergrated Customer Loyalty Program on Blockchain Technology by Using Credit Card, Jaewon Choi, International Journal on Future Revolution in Computer Science & Communication Engineering ISSN: 2454-4248. Volume: 4 Issue: 2 388 – 391. 2018.
[21] BitRewards. BitRewards White Paper Bringing blockchain to customers' loyalty. December 2017. https://bitrewards.network/wp.pdf
[22] BioCoin. LavkaLavla Ltd. https://biocoin.bio/files/Whitepaper_en.pdf
[23] EZToken Whitepaper. December 2017. https://ico.ezpos.io/files/EZToken_White_Paper_en.pdf
[24] KatalystCoin White Paper V 2.0. October 2017. https://katalystcoin.com/whitepaper/KatalystWhitePaper.pdf