Detección de intrusiones basada en modelado de red resistente a evasión por técnicas de imitación

Jorge Maestre-Vidal
jmaestre@ucm.es/ Universidad Complutense Madrid, España

Marco Antonio Sotelo-Monge
masotelo@ucm.es/ Universidad Complutense Madrid, España

Recepción: 11-6-2019 / Aceptación: 9-7-2019

Resumen. Los sistemas de red emergentes han traído consigo nuevas amenazas que han sofisticado sus modos de operación con el fin de pasar inadvertidos por los sistemas de seguridad, lo que ha motivado el desarrollo de sistemas de detección de intrusiones más eficaces y capaces de reconocer comportamientos anómalos. A pesar de la efectividad de estos sistemas, la investigación en este campo revela la necesidad de su adaptación constante a los cambios del entorno operativo como el principal desafío a afrontar. Esta adaptación supone mayores dificultades analíticas, en particular cuando se hace frente a amenazas de evasión mediante métodos de imitación. Dichas amenazas intentan ocultar las acciones maliciosas bajo un patrón estadístico que simula el uso normal de la red, por lo que adquieren una mayor probabilidad de evadir los sistemas defensivos. Con el fin de contribuir a su mitigación, este artículo presenta una estrategia de detección de intrusos resistente a imitación construida sobre la base de los sensores PAYL. La propuesta se basa en construir modelos de uso de la red y, a partir de ellos, analizar los contenidos binarios de la carga útil en busca de patrones atípicos que puedan evidenciar contenidos maliciosos. A diferencia de las propuestas anteriores, esta investigación supera el tradicional fortalecimiento mediante la aleatorización, aprovechando la similitud de paquetes sospechosos entre modelos legítimos y de evasión previamente construidos. Su eficacia fue evaluada en las muestras de tráfico DARPA'99 y UCM 2011, en los que se comprobó su efectividad para reconocer ataques de evasión por imitación.

Palabras clave: anomalías, ataques de evasión, detección de intrusiones, redes de comunicación

Intrusion Detection Based on Evasion-Resistant Network Modeling by Imitation Techniques

Abstract. Emerging network systems have brought new threats that have sophisticated their modes of operation in order to go unnoticed by security systems, which has led to the development of more effective intrusion detection systems capable of recognizing anomalous behaviors. Despite the effectiveness of these systems, research in this field reveals the need for their constant adaptation to changes in the operating environment as the main challenge to face. This adaptation involves greater analytical difficulties, particularly when dealing with threats of evasion through imitation methods. These threats try to hide malicious actions under a statistical pattern that simulates the normal use of the network, so they acquire a greater probability of evading defensive systems. In order to contribute to its mitigation, this article presents an imitation-resistant intrusion detection strategy built on the basis of PAYL sensors. The proposal is based on building network usage models and, from them, analyzing the binary contents of the payload in search of atypical patterns that can show malicious content. Unlike previous proposals, this research overcomes the traditional strengthening through randomization, taking advantage of the similarity of suspicious packages to previously constructed legitimate and evasion models. Its effectiveness was evaluated in 1999 DARPA and 2011 UCM traffic samples, in which it was proven effective in recognizing imitation evasion attacks.

Keywords: abnormalities, evasion attacks, intrusion detection, communication networks

1. INTRODUCCIÓN

Las soluciones iniciales para la detección de intrusos, basadas en el modelado y análisis de entornos de red, aprovecharon originalmente las estrategias de reconocimiento de patrones capaces de descubrir evidencias de ataques previamente conocidos (García-Teodoro, Díaz-Verdejo, Tapiador y Salazar-Hernández, 2015). Pero la rápida proliferación de las tecnologías dio lugar a una cantidad masiva de amenazas antes no vistas, fomentando así el desarrollo de soluciones alternativas capaces de hacer frente a comportamientos maliciosos. Debido a su eficacia en este contexto, el paradigma de detección de intrusos basado en anomalías se ha consolidado como la base de la mayoría de los sistemas de detección de intrusos en la red (NIDS) existentes (Karami, 2018). Este modo de operación típicamente se basa en la construcción de modelos de uso a partir de observaciones legítimas, para luego monitorizar el entorno operativo en busca de discordancias significativas, las cuales son etiquetadas como "sospechosas". Entre las diferentes publicaciones que han sentado las bases para su desarrollo, nuestra investigación se centra en la detección de intrusiones basada en PAYL (Wang, Cretu y Stolfo, 2005; Wang y Stolfo, 2004). PAYL se basa en el análisis de la carga útil del tráfico en busca de valores estadísticos atípicos dentro de cada contexto de paquetes. A pesar de la evolución de este método, la revisión de la bibliografía revela aun retos a la hora de operar en los escenarios de comunicación (Hadziosmanovic, Simionato, Bolzoni, Zambon y Etalle, 2012; Viswanathan, Tan y Neuman, 2013), como las dificultades al modelar datos extraídos de fuentes heterogéneas, el alto consumo de recursos computacionales, la escasa adaptabilidad a la no estacionariedad (*concept drift*) y la susceptibilidad a los métodos de evasión basados en aprendizaje automático (*machine learning*) (Pastrana, Orfila, Tapiador y Peris-López, 2014), siendo este último el principal objetivo de esta investigación. Con el fin de contribuir a su mitigación, este artículo presenta una nueva estrategia de detección de intrusos resistente a imitación, construida sobre la base de la familia de sensores PAYL, que intenta reforzar los avances del método APAP (Maestre Vidal, Sandoval Orozco y García Villalba, 2017a). De manera similar a sus predecesores, la propuesta construye modelos de uso de la red y, a partir de ellos, analiza los contenidos binarios de la carga útil de tráfico en busca de patrones discordantes que revelen contenidos maliciosos. A diferencia de las soluciones anteriores, esta investigación supera el tradicional fortalecimiento mediante la aleatorización, aprovechando la estimación de similitud de paquetes entre modelos legítimos y de evasión previamente construidos. Se ha llevado a cabo una amplia experimentación que demuestra la efectividad de esta solución en la detección de ataques de ofuscación basados en imitación. Este documento está dividido en cinco secciones, siendo la primera de ellas la presente introducción. En la sección 2 se revisan los trabajos relacionados y se describe el modelo de ataque de evasión considerado en esta propuesta. La sección 3 presenta un nuevo sistema de intrusiones en red (NIDS), descendiente de la familia de sensores PAYL y reforzados contra ataques de evasión. La sección 4 trata sobre la experimentación realizada y discute los resultados obtenidos. Por último, en la sección 5 se resumen las conclusiones y líneas de trabajo futuro.

2. ESTADO DEL ARTE

2.1 Detección de intrusiones basado en carga útil

El primer sensor de la familia PAYL fue elaborado por Wang y Stolfo en 2004 (Wang y Stolfo, 2004). Según sus autores, el objetivo inicial era detectar la presencia de un gusano (*worm*), ya sea a nivel de puerta de enlace o dentro de una red protegida, evitando así su propagación. Aunque el problema a resolver se basaba en el reconocimiento de gusanos, la propuesta también era válida para una amplia gama de intentos de intrusión, lo que inspiró nuevas líneas de investigación. PAYL se caracterizó por construir un modelo de uso legítimo a partir de 256 características interrelacionadas representadas como histogramas de 256 elementos. Estas se extrajeron según la metodología *n-gram* (Sidorov *et al.*, 2014), como *1-grams*, sobre la distribución de frecuencia de *bytes* presentes en la carga útil. En la etapa de detección, se comparó la similitud entre el modelo normal construido en la etapa de entrenamiento con el modelo generado por el tráfico entrante. Si su divergencia superaba un umbral predefinido, se comunicaba una alerta. Algunos problemas de PAYL inherentes a la construcción de modelos fueron resueltos por una solución impulsada por un mapa autoorganizado (SOM) (Bolzoni Etalle, Hartel y Zambon, 2006). Perdisci, Ariu, Fogla, Giacinto y Lee (2009) propusieron un conjunto de máquinas de vectores de soporte (SVM) para mejorar la precisión del NIDS, siendo abordado de forma similar por modelos ocultos de Markov en Ariu Tronci, y Giacinto (2011). Según esta investigación, las publicaciones anteriores no fueron capaces de reconocer con precisión ataques como *cross site scripting* y *SQL-injection*, donde las estadísticas de carga útil no eran significativamente diferentes del tráfico normal. Este problema fue el principal objeto de otro estudio (Swarnkar y Hubballi, 2016) en donde la efectividad del detector aumentó al implementar la clasificación bayesiana multinomial de una clase. Asimismo, estos sensores introdujeron el muestreo aleatorio con fines de mejora del rendimiento (Ariu *et al.*, 2011; Bolzoni *et al.*, 2006), que se convirtió en una estrategia habitual para reducir el impacto de la inspección profunda de paquetes (del inglés, *deep packet inspection* o DPI) en entornos de comunicación reales.

2.2 Detección de *malware* en la carga útil

En respuesta a los retos inherentes a los entornos de monitorización emergentes, el Advanced Payload Analyzer Preprocessor (APAP) (Maestre Vidal *et al.*, 2017a) introdujo una variante más compleja del detector ANAGRAM. APAP se desarrolló originalmente como un módulo de preprocesamiento de Snort, combinando así las capacidades basadas en reglas de dicho NIDS con una nueva capacidad de detección basada en anomalías. Este último se basó en *n-gram* para la extracción de las características de la carga útil (Sidorov *et al.*, 2014). Por otro lado, su almacenamiento/acceso fue gestionado mediante estructuras Counting Bloom Filters (CBF)

(Shana y Venkatachalam, 2014), lo que redujo el consumo de memoria del sistema y mejoró la aplicación de funciones de *hashing*. APAP consta de cinco etapas diferentes de procesamiento de datos, que se agrupan en un par de conjuntos de acciones: entrenamiento y detección. En el entrenamiento se distinguen cuatro fases: inicialización, entrenamiento base, entrenamiento de referencia y definición de valores *K*. Durante la inicialización, APAP procede a eliminar información de entrenamientos anteriores, inicializando los CBF y estableciendo la función de *hashing* adecuada. En el entrenamiento base, el CBF se llena con información extraída de muestras de carga útil de tráfico normal. Esta estructura de datos almacena la ocurrencia de cada posible *n-gram* en el contenido binario de la carga útil. En el entrenamiento de referencia, se calculan los valores *K* del sensor, que son métricas que resumen el contenido de la CBF y facilitan la generación de reglas de detección. Con este fin, se considera un conjunto de datos de muestras "maliciosas", que sirven como relleno (*padding*) de una réplica del CBF modificado en el entrenamiento base. Los valores *K* resultantes se traducen en reglas de detección en la definición de valores *K*, que se basa en contrastar los CBF de contenido "normal" y "malicioso". Finalmente, en el modo detección de APAP, estas reglas determinan la activación de alertas y/o contramedidas.

2.3 Evasión basada en imitación

Durante la última década, la comunidad investigadora ha variado su percepción de los ataques basados en imitación del entorno operativo para el que fueron diseñados. En Jonathon, Somesh y Miller (2006) se presentó una de las visiones preliminares de la imitación para evadir los IDS basados en la modelización y el análisis de secuencias de llamadas de sistemas mediante la intercalación de acciones típicamente legítimas entre acciones maliciosas. Investigaciones posteriores (Maestre Vidal, Sandoval Orozco y García Villalba, 2016; Tapiador y Clark, 2010) se enfocaron en cómo fortalecer los sistemas de detección interna contra estas amenazas. Los ataques de imitación también fueron revisados en profundidad en el campo del reconocimiento de intrusiones mediante el análisis del contenido de la carga útil de los paquetes de una red (Pastrana *et al*., 2014; Wang, Parekh y Stolfo, 2006), lo que supone hoy un reto a considerar. Teniendo en cuenta la representación de ataques de imitación presentada en Jonathon *et al*. (2006), estas amenazas pueden ser entendidas como acciones que intentan explotar la situación descrita en la figura 1; donde $\sum$ es el conjunto de *n-grams* extraíbles de una carga útil de paquetes, y $\sum^*$ es el conjunto infinito de todas las cargas útiles posibles. Un modelo legítimo acepta los *n-grams* de M(L) como normales, mientras que los *n-grams* de M(A) son etiquetados como potencialmente dañinos. Por consiguiente, las observaciones que caen en la intersección entre M(L) y M(A) conducirán a etiquetados no deterministas. Cuanto mayor es la cercanía a M(L), mayor es la probabilidad de ser etiquetado como normal.

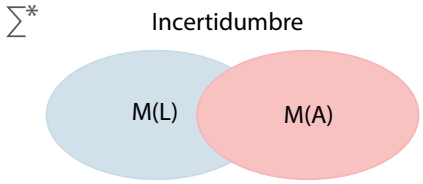


Figura 1. Observaciones potencialmente inadvertidas por el NIDS
Elaboración propia

En el contexto de PAYL, los ataques de imitación intentan explotar la incertidumbre inherente a la región M(L) ∩ M(A) ocultando el contenido de *malware* dentro de un envoltorio de relleno extraído de M(L). Para evitar las técnicas de imitación se asumió que el atacante tiene un conocimiento adecuado de M(L), el modelado del uso de la red objetivo y las estrategias de detección implementadas. En respuesta a las capacidades de PAYL, los atacantes adoptaron una amplia variedad de métodos de evasión, siendo la imitación una de sus tácticas más destacadas. Como se demostró en Fogla, Sharif, Perdisci, Kolesnikov y Lee (2006). Frente a ello, la propuesta de ANAGRAM (Wang *et al.*, 2006) introdujo un clasificador binario que implementó filtros de Bloom (Rottenstreich y Keslassy, 2015) para registrar la distribución de los datos de la carga útil, permitiendo así operar sobre *n-grams* de mayor tamaño. ANAGRAM propuso un modelo de *n-gram* aleatorizado para dificultar la generación de muestras de evasión, el cual fue probado como una solución muy precisa, pero la evolución de los métodos de evasión basados en aprendizaje automático mostró que dicho enfoque resultaría insuficiente (Pastrana *et al.*, 2014). Con ello, el fortalecimiento de nuestra propuesta supera la aleatorización al estimar la similitud de paquetes sospechosos entre M(L) y los modelos de evasión.

3. ANÁLISIS DE CARGA ÚTIL DE TRÁFICO PARA LA DETECCIÓN DE INTRUSIONES ROBUSTO A EVASIÓN

Como mejora de los métodos APAP/APACS (Maestre Vidal *et al.*, 2017a, 2017b), el método de fortalecimiento presentado en este trabajo aborda el siguiente modo de operación: en el entrenamiento, tanto los modelos normales como los de evasión se construyen de acuerdo con las características extraídas por la metodología *n-gram* y se almacenan como CBF. En la etapa de detección, las cargas útiles a analizar se recogen del entorno protegido y se comparan con los modelos de uso. Las medidas de similitud entre las observaciones y los modelos de uso de la red previamente construidos en la etapa de entrenamiento permiten estimar su naturaleza (normal o sospechosa) y la coherencia del etiquetado (véase la figura 2). En esta investigación se asume que a mayor diferencia, mayor es la probabilidad de que las muestras fueran elaboradas por los intrusos. Esta sección describe cada etapa del procesamiento de datos y los criterios de decisión adoptados.

3.1 Entrenamiento base

En la etapa de entrenamiento se construye el modelo que resume las principales características de las cargas útiles legítimas extraídas de la red. Para su elaboración es necesaria una colección de muestras representativas del tráfico legítimo (habitual) de la red. El modelo de uso normal utiliza CBF para almacenar la frecuencia de ocurrencia de cada *n-gram* dentro de la carga útil, alimentando así el CBF(L) con la información extraída de las muestras de referencia hasta que sea posible concluir que su contenido es suficientemente representativo. Se supone que aquello ocurre cuando se añade nueva información y no hay variaciones representativas en la distribución de datos dentro del CBF. Este grado de saturación se evalúa implementando y adaptando el método elbow (Green, Staffell y Vasilakos, 2014), donde el punto de inflexión se calcula observando la suma de errores cuadráticos (SSE) entre las posiciones de los CBF(L) ocupadas por el tráfico normal.

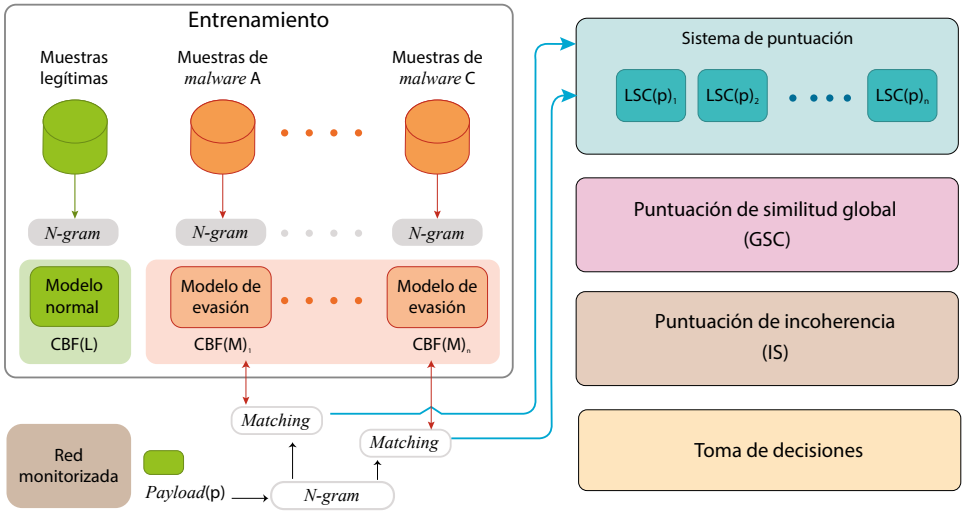


Figura 2. Esquema de procesamiento de datos
Elaboración propia

3.2 Refinamiento por modelos de evasión

Con el fin de reducir la tasa de falsos positivos, en la etapa de entrenamiento refinado, la propuesta genera modelos de evasión capaces de representar los contenidos binarios de diferentes rasgos de *malware*. Siguiendo el mismo procedimiento aplicado para la construcción de modelos de uso normal, se rellena correctamente un CBF por cada grupo de muestras

maliciosas. Sean *n* las caracterizaciones de tráfico malicioso tomadas en cuenta, se obtienen los modelos de evasión provisionales correctamente alimentados:

$$CBF'(M) \coloneqq CBF'(M)_1, \dots, CBF'(M)_n \qquad (1)$$

Téngase en cuenta que las trazas maliciosas originales pueden contener elementos de tráfico normal sobre los que se transfiere el *malware*. Para intentar minimizar el número de errores de etiquetado de la carga útil, la operación de sustracción se aplica entre cada uno de ellos y el modelo de uso normal de CBF(L):

$$C(M): = CBF'(M)_1 - CBF(L), \dots, CBF'(M)_n - CBF(L) \quad (2)$$

3.3 Sistema de puntuación y detección de intrusiones

Sea la carga útil de tráfico *p* y de longitud *m* a analizar, p: b_1, b_2, b_3, b_4, b_5, b_6, b_7, $\cdots$ bm, en donde cada bi, $0 < i \leq m$, representa un *bit* [0,1] de su contenido, en la que se asume una ventana deslizante *K-gram* para su extracción, las posiciones del CBF indicadas por las secuencias binarias b_1, b_2, ... b_{k+1}, son accedidas en CBF(L) y cada modelo de evasión $CB(M)_1$,..., $CBF(M)_n$. A partir de los resultados obtenidos, se calcula la puntuación del paquete, que tiene en cuenta tanto los aspectos binarios como los espectrales. En el primer caso, se obtiene el número de coincidencias (posiciones CBF superiores a 0) respecto a los modelos normales (γ) y de evasión ($\delta_1, \dots, \delta_n$). En los enfoques basados en espectro se tienen en cuenta las frecuencias de aparición almacenadas en los CBF, siendo α respecto a γ, y $\beta_1, \dots, \beta_n$ respecto a $\delta_1, \dots, \delta_n$. Sea modelo legítimo CBF(L) y su correspondiente modelo de evasión $CBF(M)_i$, $0 < i \leq n$, la puntuación de similitud local (LSC) de la carga útil *p* se define mediante la siguiente expresión:

$$\mathrm{LSC}(p) = \frac{\alpha - \beta_i}{\alpha + \beta_i + \mu}, \ \mathrm{LSC} \in [-1,1] \qquad (3)$$

donde α es la suma de las ocurrencias de los *n-grams* extraídos de *p* y emparejados exitosamente en CBF(L) (cuando la posición del CBF es mayor que 0), βi es la suma de ocurrencias de los *n-grams* extraídos de *p* emparejados exitosamente en $CBF(M)i$, y μ es el número de *n-grams* con valor 0 en los CBF. Por otra parte, la puntuación de similitud global (GSC) de *p* se define como el mínimo LSC calculado $(GSC(p) \in [-1,1])$. En la etapa de detección, el NIDS monitorea la carga útil de los paquetes en busca de patrones de contenido malicioso. En particular, el NIDS calcula las puntuaciones local y global (LSC y GSC, respectivamente) con respecto al modelo de uso normal de la red CBF(L) y los modelos de evasión $CBF(M)_1$ ..., $CBF(M)_n$ previamente construidos en la etapa de entrenamiento. Cuando se emiten alertas $(p) < \tau$, siendo $\tau \in [-1,1]$ un intervalo de confianza previamente definido que actúa como parámetro de ajuste del sensor.

3.4 Fortalecimiento contra imitación

En esta propuesta, es posible separar el contenido normal de las muestras maliciosas una vez que se realizan las operaciones $CBF'(M)_i = CBF(M)_i - CBF(L)$, donde cada patrón común entre $CBF(L)$ y $CBF(M)_i$ es reducido o eliminado. Consecuentemente, es altamente improbable que una carga útil de p a analizar muestre una gran similitud con (L) y algunos de los $CBF'(M)_1, ..., CBF(M)_n$ en la etapa de detección. Se planteó como hipótesis que de esta situación se pueden deducir situaciones en las que el NIDS no etiquetó la muestra con suficiente confiabilidad. En consecuencia, se asumió que cuando el sensor opera sobre modelos de evasión limpios, esto plantea una evidencia potencial de ofuscación por imitación, o un rasgo de entrenamiento deficiente. Por consiguiente, durante la experimentación se considera la ocurrencia de un etiquetado incoherente (IL) cuando se cumple la siguiente expresión:

$$\frac{\alpha}{\alpha+\mu_\alpha} > \phi \qquad and \qquad \exists_i, \frac{\beta_i}{\beta_i + \mu_\beta} > \phi \longrightarrow \text{IL} \qquad (4)$$

donde α es la suma de ocurrencias de los *n-grams* extraídos de p coincidentes en CBF(L) (cuando la posición del CBF es mayor que 0); $\mu\alpha$ es el número de *n-grams* en CBF(L) establecidos en 0; β_i es la suma de las ocurrencias de los *n-grams* extraídos de p y emparejados exitosamente en $CBF(M)_i$; y μ es el número de *n-grams* con valor 0. El intervalo de confianza de incoherencia φ actúa como parámetro de ajuste para el reforzamiento del sensor contra imitación. La puntuación de incoherencia (IS) se calcula de la siguiente manera:

$$IS(p) = 1 - \left| \frac{\alpha}{\alpha + \mu_\alpha} - Max\left\{ \frac{\beta_i}{\beta_i + \mu_i} \right\} \right| \qquad (5)$$

4. EXPERIMENTOS Y RESULTADOS

4.1 *Datasets* y metodología de evaluación

La propuesta fue evaluada sobre los *datasets* DARPA'99 y UCM 2011. DARPA'...99 (Lippmann, Haines, Fried, Korba y Das, 2000) proporciona colecciones de muestras reales y sintéticas monitorizadas en un entorno experimental y plantea un estándar funcional que permite establecer comparativas con trabajos anteriores. Como es habitual en la bibliografía, la validación de la propuesta consideró la segunda versión (Hadziosmanovic *et al*., 2012) entrenada en base a capturas de tráfico recogidas durante siete días y separadas en sesiones etiquetadas como "normales" o "ataques". Nótese que las muestras "normales"' se utilizaron para construir el modelo CBF(L), y los segundos para construir los modelos de evasión $CBF(M)_1, \ldots, CBF(M)_m$. Con el fin de evaluar las mejoras de la propuesta respecto a APAP y las contribuciones a la familia PAYL, se consideró el conjunto de datos UCM 2011. Este recoge trazas reales de

tráfico monitorizadas en la red de la Facultad de Informática de la Universidad Complutense de Madrid. El registro de tráfico se realizó a lo largo del año 2011 en diferentes intervalos de tiempo y meses. Las muestras de tráfico normal incluyen actividades de uso habitual de la red, entre ellas: intercambios P2P, transferencias de archivos de varios formatos (.doc, .pdf, .mp3, .jpg, etc.) vía SMTP, navegación HTTP/HTTPS, etc. Los contenidos maliciosos se clasificaron en dieciséis grupos diferentes de amenazas, incluyendo varias familias de *malware*, ataques DoS o procedimientos de obtención de privilegios. La experimentación utilizó el generador de ataques de imitación descrito en Maestre Vidal *et al.* (2016). Con el propósito de hacer que la carga útil del tráfico malicioso incremente su similitud con las características del tráfico normal, la herramienta de ofuscación se basó en el CBF(L) construido en la etapa de entrenamiento. La selección de las secuencias de relleno consideró la estrategia de muestreo estocástico *probability proportional-to-size* (PPS).

4.2 Efectividad con DARPA'99

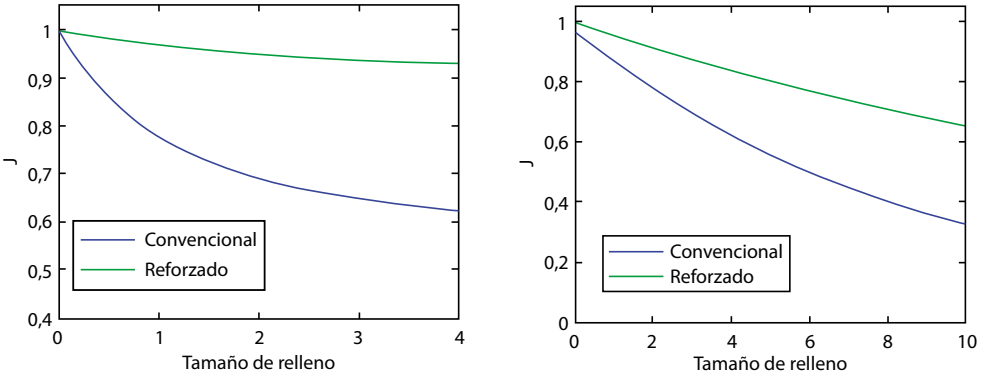
Tal y como se señala en Wang *et al.* (2006) y se ratifica en Hadziosmanovic *et al*. (2012) y Maestre Vidal *et al.* (2017a, 2017b), dada la naturaleza del conjunto de datos DARPA'99, una ventana deslizante de 3 *grams* resultó ser la configuración más adecuada para los sensores inspirados en PAYL, por lo que se mantuvo dicha configuración en esta propuesta. En esta experimentación, el parámetro de ajuste del sensor fue el intervalo de confianza $\tau \in [-1,1]$, por lo que cada carga útil p se etiquetó como anómala cuando la expresión $(p) < \tau$ fue satisfecha. Tal y como se muestra en las comparativas anteriores, los resultados obtenidos se evaluaron en función del mejor ajuste entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) estimada por el índice óptimo de Youden (J). La tabla 1 muestra la eficacia de las publicaciones relacionadas y resume los resultados obtenidos en esta propuesta. Como se indica, la calibración óptima resultó en una tasa de aciertos del 100 % y FPR = 0,01 % cuando J = 0,9998. Estos resultados son cercanos a los del mejor sensor cuando opera en circunstancias similares (ANAGRAM). Pero a pesar de estos hallazgos, el uso funcional de DARPA'99 carece de una manera de demostrar la robustez de los sensores frente a amenazas de evasión, por lo cual fue necesario realizar experimentos adicionales para evaluar el resto de las mejoras de esta propuesta. Con el fin de evaluar el fortalecimiento de la propuesta frente a las amenazas de evasión, esta prueba consideró un intervalo de confianza fijo de τ, en particular, el que mostraba el equilibrio óptimo entre el TPR y el FPR en los experimentos anteriores. El parámetro de calibración fue la variación del índice de incoherencia en el etiquetado φ, hipotetizando que cuanto mayor sea φ, mejor será la TPR. El experimento suponía que cada incoherencia en el etiquetado ocultaba una carga útil maliciosamente ofuscada.

Tabla 1
Comparación con propuestas anteriores (DARPA'99)

Propuesta	Porcentaje FPR	Porcentaje TPR
PAYL (Wang y Stolfo, 2004)	0,00	90,76
Poseidon (Bolzoni *et al.*, 2006)	0,00	92,00
AnPDPP (Thorat Khandelwal, Bruhadeshwar y Kishore , 2009)	0,06	100,00
Anagram (Wang *et al.*, 2006)	0,00	100,00
McPAD (Perdisci *et al.*., 2009)	0,33	87,80
RePIDS (Jamdagni Tan, He, Nanda y Liu, 2013)	0,67	99,33
APAP (Maestre Vidal *et al.*, 2017a)	0,15	100,00
Esta propuesta	0,01	100,00

Elaboración propia

En la figura 3 se ilustra la variación del mejor ajuste de ϕ según el índice de Youden, donde se observa que, tanto en los modos de funcionamiento convencionales como en los reforzados, la propuesta presenta sensibilidad a las variaciones en la longitud del relleno. Sin embargo, el enfoque reforzado se redujo significativamente en el impacto, siendo mínima la precisión observada de aproximadamente $AUC \approx 0{,}95$.



a. DARPA'99 b. UCM 2011

Figura 3. Impacto del tamaño de relleno
Elaboración propia

4.3 Efectividad con UCM 2011

Como se muestra en la tabla 2, la mejor configuración (AUC = 0,9900, J = 0,9574) superó al mejor ajuste APAP (AUC = 0,9136, J = 0,8442), pero fue similar a APACS (AUC = 0,9902, J = 0,9339). Sin embargo, dado que esta propuesta no se basa en correlación de las alertas (como es el caso de APACS), supone una solución más eficaz. Por otro lado, APACS requiere una base de conocimiento correctamente etiquetada, mientras que la presente propuesta proporciona tolerancia a los errores de etiquetado en los datos de entrenamiento. El reforzamiento propuesto contra los intentos de evasión basados en imitación volvió a demostrar su eficacia al aplicarse sobre las muestras de tráfico UCM 2011. Esto se ilustra en la figura 3b, donde se reflejan las variaciones del índice de Youden basadas en la distribución del contenido de relleno. En este experimento se fijó el intervalo de confianza τ al ajuste óptimo de la prueba anterior, implementándose una ventana deslizante de 5 *grams*, y el parámetro de sensibilidad fue el índice de incoherencia en el etiquetado ϕ. La versión reforzada no mitigó por completo el impacto de las amenazas de imitación, pero la redujo considerablemente, registrando las peores AUC con valores cercanos a 0,9, lo que mejoró representativamente el 0,7 alcanzado por un despliegue no robusto. El punto de saturación fue de aproximadamente $AUC \approx 0{,}85$, mientras que en la versión original era cercana a 0,6.

Tabla 2
Resumen de resultados UCM 2011

NIDS	FPR	TPR	AUC	J
Esta propuesta	0,021	0,967	0,9900	0,9574
APAP (Maestre Vidal *et al*., 2017a)	0,080	0,947	0,9136	0,8442
APACS (Maestre Vidal *et al*., 2017b)	0,034	0,995	0,9902	0,9339

Elaboración propia

5. CONCLUSIONES

Este artículo presentó un nuevo enfoque para la detección estadística de *malware* en entornos de comunicación emergentes. La propuesta se basó en el análisis de la carga útil de tráfico en busca de discordancias con respecto a modelos legítimos construidos previamente en la etapa de entrenamiento. La propuesta adoptó las bases de la familia de sensores PAYL, y amplió las soluciones APAP y APACS, aprovechando así la metodología *n-gram* y las estructuras de datos CBF. A diferencia de las soluciones anteriores, esta investigación superó el tradicional fortalecimiento mediante la aleatorización, valiéndose de la similitud de paquetes sospechosos entre modelos legítimos y de evasión construidos previamente. La propuesta ha sido evaluada con las muestras de tráfico DARPA'99 y UCM 2011, demostrando alta precisión y similitud

con las mejores propuestas. Sin embargo, su eficacia demostró ser superior a la de sus antecesores en el procesamiento de carga útil maliciosa ofuscadas por los métodos de imitación. Pero a pesar de los detalles provistos en este artículo, la discusión de algunos otros aspectos fue pospuesta para el trabajo futuro, así como la evaluación de la propuesta sobre entornos de monitorización alternativos.

REFERENCIAS

Ariu, D., Tronci, R., y Giacinto, G. (2011). HMMPayl: An intrusion detection system based on hidden Markov models. *Computers y Security*, *30*(4), 221-241.

Bolzoni, D., Etalle, S., Hartel, P., y Zambon, E. (2006). Poseidon: a 2-tier anomaly-based network intrusion detection system. *Proceedings of the 4th IEEE International Workshop on Information Assurance (IWIA)*, 144-156.

Fogla, P., Sharif, M., Perdisci, R., Kolesnikov, O., y Lee, W. (2006). Polymorphic blending attacks. *Proceedings of the 15th USENIX Security Symposium*, 241-256.

García-Teodoro, P., Díaz-Verdejo, J. E., Tapiador, J. E., y Salazar-Hernández, R. (2015). Automatic generation of HTTP intrusion signatures by selective identification of anomalies. *Computers and Security*, *55*, 159-174.

Green, R., Staffell, I., y Vasilakos, N. (2014). Divide and Conquer? k-Means Clustering of Demand Data Allows Rapid and Accurate Simulations of the British Electricity System. *IEEE Transactions on Engineering Management*, *61*(2), 251-260.

Hadziosmanovic, D., Simionato, L., Bolzoni, D., Zambon, E., y Etalle, S. (2012). N-gram against the machine: On the feasibility of the n-gram network analysis for binary protocols. *Proceedings of the 15th International Symposium on Recent Advances in Intrusion Detection (RAID)*, 59-81.

Jamdagni, A., Tan, Z., He, X., Nanda, P., y Liu, R. P. (2013). RePIDS: A multi tier realtime payload-based intrusion detection system. *Computer Networks*, *57*, 511-824.

Jonathon, T., Somesh, J., y Miller, B. P. (2006). Automated discovery of mimicry attacks. *Proceedings of the 9th International Symposium on Recent Advances in Intrusion Detection (RAID),* 41-60.

Karami, A. (2018). An anomaly-based intrusion detection system in presence of bening outliers with visualization capabilities. *Expert Systems with Applications*, 108, 36-60.

Lippmann, R., Haines, J. W., Fried, D. J., Korba, J., y Das, K. (2000). The 1999 DARPA off-line intrusion detection evaluation. *Computer Networks*, *34*(4), 579-595.

Maestre Vidal, J., Sandoval Orozco, A. L., y García Villalba, L. J. (2016). Online masquerade detection resistant to mimicry. *Expert Systems with Applications: An International Journal*, **61**, 162-180.

Maestre Vidal, J., Sandoval Orozco, A. L., y García Villalba, L. J. (2017a). Advanced payload analyzer preprocessor. *Future Generation Computer Systems*, *76*, 474-485.

Maestre Vidal, J., Sandoval Orozco, A. L., y García Villalba, L. J. (2017b). Alert correlation framework for malware detection by anomaly-based packet payload analysis. *Journal of Network and Computer Applications*, *97*, 11-22.

Pastrana, S., Orfila, A., Tapiador, J. E., y Peris-López, P. (2014). Randomized anagram revisited. *Journal of Network and Computer Applications*, *21*, 182-186.

Perdisci, R., Ariu, D., Fogla, P., Giacinto, G., y Lee, W. (2009). McPAD: A multiple classifier system for accurate payload-based anomaly detection. *Computer Networks*, *53*(6), 864-881.

Rottenstreich, O., y Keslassy, I. (2015). The Bloom paradox: when not to use a Bloom filter. *IEEE/ACM Transactions on Networking*, *23*(3), 703-716.

Shana, J., y Venkatachalam, T. (2014). An Improved Method for Counting Frequent Itemsets Using Bloom Filter. *Procedia Computer Science*, *47*, 84-91.

Sidorov, G., Castillo, F., Stamatatos, E., Gelbukh, A., y Chanona-Hernández, L. (2014). Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications: An International Journal*, 41, 853-860

Swarnkar, M., y Hubballi, N. (2016). OCPAD: One class Naive Bayes classifier for payload based anomaly detection. *Expert Systems with Applications: An International Journal*, *64*, 330-339.

Tapiador, J. E., y Clark, J. A. (2010). Information-theoretic detection of masquerade mimicry attacks. *2010 Fourth International Conference on Network and System Security*, 183-190.

Thorat, S. A., Khandelwal, A. K., Bruhadeshwar, B., y Kishore, K. (2009). Anomalous packet detection using partitioned payload. *Journal of Information Assurance and Security*, *3*(3), 195-220.

Viswanathan, A., Tan, K., y Neuman, C. (2013). Deconstructing the Assessment of Anomaly-based Intrusion Detectors. *Proceedings of the 16th International Symposium on Recent Advances in Intrusion Detection (RAID)*, 286-306.

Wang, K., Cretu, G., y Stolfo, S. J. (2005). Anomalous Payload-based Worm Detection and Signature Generation. *Proceedings of the 8th International Symposium on Recent Advances in Intrusion Detection (RAID)*, 227-246.

Wang, K., Parekh, J. J., y Stolfo, S. J. (2006). Anagram: A Content Anomaly Detector Resistant to Mimicry Attack. *Proceedings of the 9th International Symposium on Recent Advances in Intrusion Detection (RAID)*, 226-248.

Wang, K., y Stolfo, S. J. (2004). Anomalous Payload-based Network Intrusion Detection. *Proceedings of the 7th International Symposium on Recent Advances in Intrusion Detection (RAID)*, 203-222.