
XV Jornadas de Ingeniería Telemática.
JITEL 2021.
Universidad de A Coruña.

Actas de las XV Jornadas de Ingeniería Telemática (JITEL 2021), A Coruña (España), 27-29 de octubre de 2021.

Generación automática de firmas para detección de ciberataques basados en URI

R. Estepa Alonso*, J. Diaz-Verdejo†, A. Estepa Alonso*, G. Madinabeitia*, F. J. Muñoz*
* Dpt. Ingeniería Telemática, Escuela Superior de Ingenieros, Univ. de Sevilla
C/ Camino de los Descubrimientos s/n, 41092 Sevilla (Spain)
E-mail: {rafa,aestepa,german,javi }@trajano.us.es
† Dpt. Teoría de Señal, Telemática y Comunicaciones, CITIC, Univ. de Granada
C/ Periodista Daniel Saucedo Aranda, s/n, 18071 Granada (Spain)
E-mail: jedv@ugr.es

La mayor parte de los sistemas de detección de intrusiones (IDS) operativos se basan en el uso de firmas que permiten identificar ataques conocidos. La dependencia de estos IDS con la actualización de las bases de datos de firmas constituye una de sus mayores limitaciones, siendo de interés el desarrollo de sistemas que posibiliten la generación automática o supervisada de firmas.

En el presente trabajo se evalúa experimentalmente un sistema para la generación de firmas a partir de un IDS basado en anomalías propuesto en un trabajo previo. También se desarrolla y evalúa un sistema automatizado para la selección del punto de operación óptimo del generador de firmas. Los resultados preliminares de este trabajo en curso muestran que se pueden generar firmas nuevas que aumenten la capacidad de detección del IDS basados en firmas o patrones conocidos (SIDS) controlando el número de falsos positivos introducidos.

***Palabras Clave*—Cybersecurity, Intrusion Detection, Automatic signatures generation, Web-based attacks**

I. INTRODUCCIÓN

La necesidad de proteger los equipos y redes de ciberamenzas es cada vez más notoria y relevante. Uno de los elementos clave en la seguridad de los sistemas y redes son los denominados sistemas de detección de intrusiones (IDS, del inglés *Intrusion Detection Systems*) [1], que emiten alertas a partir de la observación de los diversos eventos que ocurren en la red o los sistemas a proteger. Los IDS generan alertas según dos modos de operación básicos: basado en *firmas* (SIDS, del inglés *Signature-based IDS*), que identifican un patrón malicioso preestablecido denominado firma, como por ejemplo una secuencia dentro de la URI de una petición HTTP; o basados en anomalías (AIDS, del inglés *Anomaly-based IDS*), que identificación de comportamientos anómalos, dando lugar a los IDS.

Los SIDS son sistemas muy extendidos en la actualidad, dado que permiten detectar ataques ya conocidos con una fiabilidad y coste computacional razonables. Como es lógico, el adecuado comportamiento de los SIDS depende fuertemente de la disponibilidad y calidad de las firmas, que deben ser generadas y actualizadas periódicamente. Por tanto, estos sistemas resultan inadecuados para detectar ataques novedosos, o de día cero (*0-day*), por no existir firmas para los mismos. Sin embargo, éstos representan un porcentaje importante del total de ataques y, sobre todo, generan un fuerte impacto. La solución pasaría por la generación de las firmas correspondientes, pero este problema es recursivo, ya que para poder generar la firma es necesario detectar previamente el ataque, por lo que debe utilizarse algún procedimiento alternativo. De ahí el interés de desarrollar sistemas que sean capaces de generar las firmas de forma automática o semiautomática.

Como hemos mencionado anteriormente, los AIDS [1] constituyen una aproximación diferente a la detección de ataques y son potencialmente capaces de detectar ataques *0-day*. Su rendimiento dependerá de su capacidad de aprender y discriminar el comportamiento normal/anómalo. En entornos IT, donde en ocasiones no hay un patrón claro de comportamiento del usuario, esta tarea se adivina compleja, lo que propicia la aparición de numerosos falsos positivos (FP), siendo ésta una de las mayores limitaciones de los AIDS en la actualidad.

Son múltiples los trabajos en los que se ha propuesto el uso de AIDS para identificar ataques y, a partir de ellos, generar las firmas correspondientes para los SIDS [2]. Para ello, se necesita no sólo determinar si se está desarrollando un ataque, sino también identificar los elementos significativos del mismo, que serán los asociados a la firma. El interés de esta aproximación reside en la mayor facilidad de uso e implementación de los SIDS, y en la posible capacidad de generalización de las firmas

así obtenidas, eliminando o reduciendo significativamente la intervención de los expertos. Su utilidad, no obstante, vendría limitada por las tasas de FP a las que podrían dar lugar estas nuevas firmas.

En un trabajo previo [3] se ha propuesto un sistema automático para la generación de firmas en el contexto de ataques basados en URI (véase Sección II). El AIDS subyacente se basa en [4], que modela las URI en base a una aproximación markoviana que permite identificar los elementos asociados en mayor medida a la clasificación como ataque y, consecuentemente, proponer firmas para los mismos. Los resultados obtenidos evidencian la posibilidad de conseguir una generación de firmas adecuada, pero son fuertemente dependientes del punto de operación del sistema, que es ajustado de forma manual en un procedimiento que puede resultar complejo.

En el presente trabajo en curso pretendemos explorar las capacidades de dicha propuesta en un escenario operativo real que incluye varios servidores que cooperan para establecer las nuevas firmas. Para ello se abordan propuestas y mejoras en tres aspectos relevantes. En primer lugar, se plantea un sistema automático de selección del punto de operación óptimo para la generación de las firmas, analizando el impacto de los FP sobre las reglas generadas y, consecuentemente, sobre el uso de las mismas en el escenario real. Por otra parte, se plantean diversas técnicas para la selección y agrupación de las firmas a partir de los segmentos identificados como asociados a ataques. Finalmente, se analizará la capacidad de generalización de las firmas a partir de su distribución a otros servicios diferentes a aquel en el que se ha inferido. El objetivo final es el desarrollo de un sistema global de generación y distribución de firmas para ataques basados en URI. Este trabajo se está llevando a cabo en el ámbito de un proyecto de colaboración con una empresa andaluza del sector de SmarCities, que proporcionará datos reales obtenidos durante operación.

El presente artículo se estructura como sigue. En primer lugar, en el Apartado II se presentará brevemente la técnica SSM y el trabajo previo en el que se basa la presente propuesta. El Apartado III describe la arquitectura general del sistema propuesto y aborda el problema del ajuste automático del punto de operación, presentándose el escenario utilizado para estas pruebas y los resultados experimentales obtenidos en el Apartado IV. Finalmente, en el Apartado V se presentan las conclusiones y se esbozan los desarrollos y resultados preliminares relativos a la agrupación de firmas y su distribución.

II. Generación de firmas

A continuación, describiremos brevemente los fundamentos de la técnica utilizada y su aplicación a la generación de firmas de ataques [3].

A. Detección de anomalías en URI

La técnica utiliza un autómata de estados finitos probabilístico para representar las instancias de un protocolo con estructura sintáctica en sus cargas útiles (en nuestro caso las URI de HTTP) mediante su segmentación en *palabras*. De acuerdo al estándar RFC 3986, un URI, U_k, debe presentar una estructura sintáctica de la forma:

"http://"host[":"port][abs_path["?"query]]

siendo posible su segmentación, a partir de los delimitadores estándar, en un conjunto de L palabras, $w_k = \left\{w_1^k, w_2^k, \cdots, w_L^k\right\}$, asociadas a cada uno de los campos (en nuestro caso sólo son de interés los campos *abs_path* y *query*, formada por los pares *atributo, valor*).

A partir de un conjunto de URI, es posible establecer un *diccionario*, $D = \{(w_i, f_i)\}$, compuesto por todas las palabras observadas, w_i y su frecuencia relativa de observación, f_i. De esta forma, dado un URI de entrada U_k compuesto por una secuencia de palabras, w^k y un diccionario previamente estimado, es posible asignar un *índice de anomalía*, $A_s(U_k)$, a partir de la probabilidad estimada para cada una de dichas palabras [5]:

$$A_s(U_k) = -log\left(\frac{1}{L}\sum_{i=1}^{L}\log(f_i^k)\right) \tag{1}$$

Este índice será positivo y tanto mayor cuanto menor sea la probabilidad de la secuencia observada. De esta forma, se podrá clasificar un URI como normal o anómalo de acuerdo al *umbral de detección*, θ, como

$$Clase(U) = \begin{cases} Normal & si\ A_s(U) < \theta \\ Anomalo & si\ A_s(U) \geq \theta \end{cases} \tag{2}$$

Por otra parte, esta aproximación plantea un problema de *entrenamiento insuficiente* relacionado con la posible aparición de palabras que no han sido observadas durante el proceso de entrenamiento y que, en consecuencia, tendrían asociada una probabilidad nula. Para solucionarlo se establece una probabilidad fija mínima para cualquier palabra observada, denominada *probabilidad de fuera de vocabulario*, p_{OOV}.

B. Generación de firmas

El modelado anteriormente descrito permite evaluar la probabilidad de normalidad de las distintas palabras que componen la URI, por lo que, dada una URI que se determina anómala (ataque), es posible identificar y seleccionar los segmentos que contribuyen en mayor proporción a dicha clasificación. De esta forma, se delimitan y extraen las palabras o secuencias de palabras que superan el denominado *umbral de generación de firma para un segmento*, ϕ, incluyendo los delimitadores correspondientes. Cada uno de estos fragmentos será candidato a formar parte de una nueva firma.

Por otra parte, el propio índice de anomalía de una URI es indicativo del grado de normalidad de la misma, por lo que, para minimizar el posible impacto de los FP, se establece un *umbral de generación de firmas*, Ψ, de tal forma que únicamente las URI cuyo índice de anomalía supere dicho umbral serán consideradas en el proceso de generación de firmas. En consecuencia, dado un URI, U, se determina que un segmento t es anómalo y se incorpora a una firma si se cumple

$$(A_s^t(U) \geq \phi)\ \wedge\ (A_s(U) \geq \Psi),\ con\ \Psi > \theta \tag{3}$$

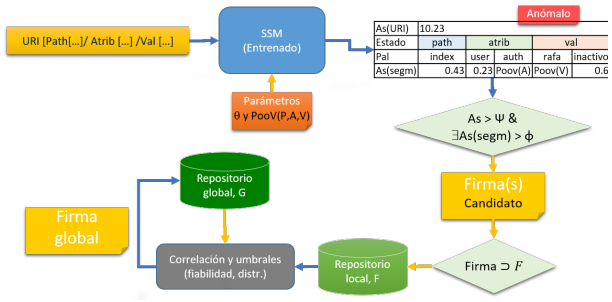

Fig. 1. Funcionamiento del sistema propuesto

siendo $A_s^t(U)$ el índice de anomalía del segmento.

La operación del sistema propuesto en este trabajo se esquematiza en la Fig. 1. Por un lado, cada uno de los AIDS desplegados y entrenados con su tráfico local evalúan las URI de entrada y, para aquellas suficientemente anómalas, extraen los segmentos candidatos a firmas, que serán agrupados convenientemente en una nueva firma integrada en un repositorio local de firmas. Como se puede observar, a partir de los modelos entrenados y ajustados en varios servidores se infieren repositorios de firmas locales que son agrupadas y analizadas para extraer un repositorio global con firmas válidas para todos los servidores. La generación de un repositorio global cooperativo de firmas será abordado en las siguientes fases del proyecto en curso, centrándose este trabajo en el sistema generador de firmas.

III. Ajuste de umbrales de la generación de firmas

Para la extracción de las firmas locales es necesario ajustar experimentalmente el sistema para seleccionar el punto óptimo de operación, que influirá en la tasas finales de detección y de falsos positivos. Consecuentemente, es necesario ajustar 3 parámetros: θ, ϕ y Ψ, ya que el valor de p_{OOV} depende del conjunto de entrenamiento. Así, el valor del umbral de generación de firma para un segmento, ϕ, debe ser inferior al de la probabilidad mínima registrada en el diccionario, esto es, $\phi < min(\{f_i\})$, para asegurar que las palabras que constituyen la firma no han sido observadas previamente. Así mismo, parece lógico pensar que las URI candidatas a generación de firmas sean un subconjunto de aquellas detectadas como anómalas, lo que exige que se cumpla $\theta < \Psi$. También resulta coherente que, para controlar el número de FP que pueden dar lugar a firmas, haya que ajustar el valor de Ψ. A continuación, proponemos un procedimiento de ajuste del umbral de generación de firmas en el que acotamos la tasa máxima de FP aceptada. Este algoritmo parte de la suposición de que la tasa de FP objetivo que tengamos en el conjunto de entrenamiento será similar a la que obtendremos durante la explotación del sistema.

A. Ajuste automático del valor de Ψ

El objetivo del mecanismo de ajuste que se propone en este trabajo es explorar un espacio de búsqueda de valores para Ψ a fin de que la tasa de FP conseguida con las firmas

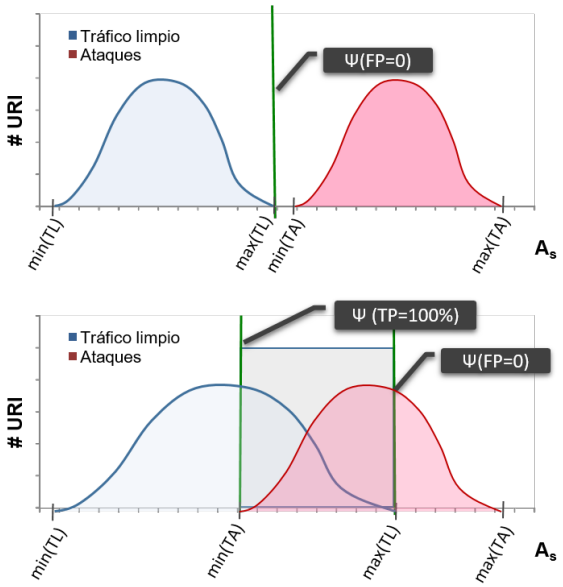

Fig. 2. Casos para el histograma de A_s.

no sobrepase un umbral determinado por el operador del servicio. En primer lugar, podemos determinar cotas para el valor de Ψ, umbral de generación de firma, a la vista de los índices de anomalía registrados durante la fase de entrenamiento.

Dado un dataset de entrenamiento con tráfico limpio (TL) y otro con tráfico de ataques (TA), es de esperar que el histograma de los índices de anomalía responda a una de las dos situaciones mostradas en la Fig. 2. En el primer caso (parte superior), que correspondería a la situación ideal, el tráfico limpio y el de ataque presentan una gran diferencia en sus diccionarios, resultando que $max(A_s(TL)) < min(A_s(TA))$, lo que implica que si elegimos $\Psi > max(A_s(TL))$ no tendremos ningún FP en el entrenamiento y detectaremos todos los ataques. Desafortunadamente, el segundo caso es el más habitual e implica que $max(A_s(TL)) > min(A_s(TA))$, por lo que valores de Ψ en el rango $[min(A_s(TA)), max(A_s(TL))]$ generarán una tasa de falsos positivos en el entrenamiento.

Así pues, el ajuste de Ψ se realizará durante el entrenamiento, evaluando iterativamente la tasa de FP encontrada en el TL cuando se utilizan las firmas generadas[1] para valores crecientes de Ψ. Esto se puede hacer con un algoritmo que parte de un valor inicial $\Psi = min(A_s(TA))$, que generará la tasa de FP máxima posible, que se computará a partir de TL. Si dicha tasa es menor que la tasa de FP objetivo, el algoritmo se detendrá, en otro caso, se incrementará el valor de Ψ y se volverá a evaluar en una nueva iteración. El resultado final será el valor de Ψ que cumple que la tasa de FP que introducen las nuevas firmas es menor que el valor objetivo.

IV. Resultados experimentales preliminares

A continuación, se presentan los resultados experimentales obtenidos relativos a la capacidad de detección y

[1]A tal efecto se ha desarrollado una sencilla herramienta SIDS denominada *InspectorLog*, que permite aplicar las firmas generadas a las URI.

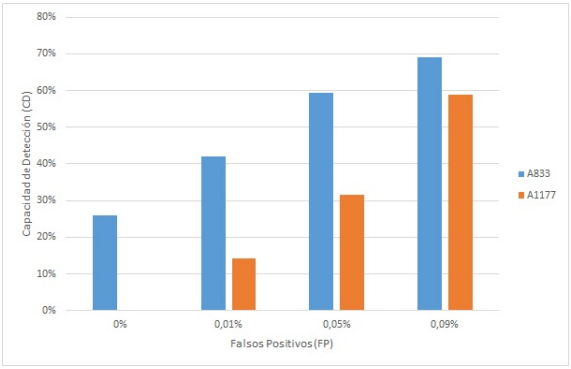

Fig. 3. Capacidad de detección de las firmas en diversos puntos de operación del AIDS.

el ajuste de umbrales. El valor de ϕ se ha ajustado a $0,9 \cdot min(\{f_i\})$, cumpliendo así la restricción de que un segmento anómalo no puede haber sido visto en el tráfico limpio. Para la experimentación se ha utilizado:

- Trafico limpio (TL): proveniente de 1 semana de tráfico real del servicio ProxyWeb de una empresa, que denominaremos H, que cuenta con 289 505 peticiones GET. Se han realizado 4 particiones para entrenamiento, test y validación.
- Tráfico de ataques (TA): se han utilizado dos dataset con 833 y 1 177 URI de ataques, respectivamente, generadas a partir de las vulnerabilidades encontradas en la base de datos CVE (*Common Vulnerabilities and Exposures*) aplicables a servidores HTTP del año 2018 [5].

El primer experimento realizado utiliza el algoritmo de ajuste de Ψ propuesto anteriormente para obtener firmas con distintos umbrales de FP tolerados en el AIDS: 0%, 0,01%, 0,05% y 0,09%. Para ello se entrena el sistema con una de las cuatro particiones y se evalúa con el resto, promediando los resultados según un esquema *leave-one-out*. Los resultados finales obtenidos para las firmas generadas con los distintos dataset de ataques se muestran en la Fig. 3.

En esta figura se puede observar que, a mayor FP objetivo mayor capacidad de detección de las firmas generadas. Con respecto a los FP detectados, siempre fueron inferiores al FP objetivo del algoritmo, tomando los valores de 0%, 0.001%, 0.007%, 0.023% para los FP objetivos 0%, 0.01%, 0.05% y 0.09% respectivamente. Estos resultados avalan la hipótesis de que la tasa de FP generados por el AIDS será siempre superior a la de las firmas obtenidas.

El siguiente experimento realizado consistió en explorar los límites del sistema cuando se establece la tasa de FP a 0, para observar la capacidad máxima de detección obtenida. En la Tabla I se pueden observar los resultados para el dataset de 833 ataques. Vemos que entrenando con el tráfico limpio H1 (primera partición) tan sólo somos capaces de detectar un 33,73% de los ataques, que generarían 33 firmas. Las distintas particiones de TL empleadas (H1-H4) dan lugar a diferentes valores. Para cada experimento se muestra el valor óptimo de Ψ determinado por el algoritmo, el máximo valor que podría tomar (columna rango), la capacidad de detección de ataques, los falsos positivos encontrados y el número de firmas generadas.

Tabla I
RESULTADOS DE GENERACIÓN DE FIRMAS CON DIFERENTES PARTICIONES.

Exp	Ψ	rango	CD(%)	FP(%)	N. Firmas
H1.833	16.31	17.36	33,73	0	33
H2.833	16.27	17.31	33,73	0	66
H3.833	16.83	17.33	2	0	20
H4.833	16.29	17.33	33,7	0	33

V. CONCLUSIONES

La generación automatizada permite mejorar la capacidad de detección de los SIDS. En este artículo se ha evaluado el rendimiento de un sistema generador de firmas en el contexto de ataques en la URI así como un método para el ajuste de umbrales y reducción de FP. También se han presentado algunos resultados preliminares dentro de los límites de espacio asociados al tipo de trabajo (en curso). Los resultados muestran la capacidad de detección de ataques novedosos que no eran detectados mediante las firmas disponibles sin incrementar la tasa de FP del SIDS. Actualmente estamos trabajando con datasets de mayor tamaño que permiten seguir desarrollando y mejorando el sistema, así como en el uso cruzado de las firmas para estudiar la capacidad de generalización.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el proyecto 2020/00000172 dentro del programa de Proyectos singulares de actuaciones singulares de transferencia en los CEI en las áreas RIS3 de la Junta de Andalucía.

REFERENCIAS

[1] N. Moustafa, J. Hu, J. Slay, "A holistic review of Network Anomaly Detection Systems: A comprehensive survey", Journal of Network and Computer Applications,(128)33?55, 2019.
[2] S. Kaur, M. Singh, "Automatic attack signature generation systems: A review", IEEE Secur. Priv., (11)54–61, 2013.
[3] P. Garcia-Teodoro, J.E. Diaz-Verdejo, J. Tapiador, R. Salazar-Hernandez, "Automatic generation of HTTP intrusion signatures by selective identification of anomalies", Computers and Security, (55)159–174, 2015.
[4] J. M. Estévez-Tapiador, P. García-Teodoro, J. E. Díaz-Verdejo, "Detection of web-based attacks through Markovian protocol parsing", Proc. IEEE Symp. on Computers and Communications, 2005.
[5] R. Estepa, J.E. Díaz-Verdejo, A. Estepa, G. Madinabeitia, "How Much Training Data Is Enough? A Case Study for HTTP Anomaly-Based Intrusion Detection", IEEE Access, 8:44410–44425, 2020.