Universidad Nacional De Asunción

Facultad Politécnica

Ingeniería en Informática

Propuesta de Trabajo Final de Grado

Web Application Firewall **con detección de anomalías usando características específicas de tráfico HTTP y** ***One-Class*** **SVM**

Autores:

Nico Epp
Año de ingreso: 2010 - CI: 3.184.199
Ralf Funk
Año de ingreso: 2011 - CI: 3.886.529

Tutor:

Msc. Cristian Cappo

San Lorenzo - Paraguay
Agosto - 2017

1. Motivación

Las aplicaciones web han tenido un gran auge en la última década, convirtiéndose en herramientas de uso masivo y frecuente para una gran cantidad de usuarios. Pero debido a que las mismas son accesibles a través de la red, están expuestas a una gran variedad de ataques [Gim15]. Muchas de las aplicaciones web actualmente no están construidas de acuerdo a las mejores prácticas de seguridad, posibilitando que dichas aplicaciones queden vulnerables a diferentes ataques. Esto se debe a la falta de consciencia sobre la importancia de la seguridad y en muchos casos también a una falta de tiempo, ya que se suele priorizar el desarrollo de funcionalidades por encima de la seguridad. Esta es la situación de aplicaciones existentes como también lo puede ser para aplicaciones futuras. Por lo tanto se necesitan soluciones para mitigar los riesgos presentes.

En este trabajo nosotros investigaremos sobre mecanismos externos especializados en la detección de ataques, con el fin de mitigar los riesgos creados por las vulnerabilidades presentes en las aplicaciones web.

Sistemas de detección de intrusos (IDS - *Intrusion Detection System*) son programas o dispositivos especializados para monitorear las actividades en un sistema en busca de intrusiones no autorizadas o posibles ataques [SM07]. Las respuestas frente a posibles intrusiones pueden ser variadas, desde envío de alertas hasta medidas concretas de mitigación y contención de los posibles ataques. En este último caso se puede hablar también más específicamente de Sistemas de prevención de intrusos (IPS - *Intrusion Prevention System*) [SM07].

Los IDS pueden basarse en varias fuentes de datos para sus análisis, como por ejemplo el tráfico de una red o los registros de acciones en un sistema operativo [TG15]. Como las aplicaciones web utilizan mayormente el protocolo HTTP (*Hypertext Transfer Protocol*) [FGM$^+$99] para sus comunicaciones, se necesita un IDS que pueda monitorear el tráfico HTTP, analizando los paquetes enviados y recibidos a través de las conexiones de red. En este caso se puede hablar más específicamente de cortafuegos para aplicaciones web (WAF - *Web Application Firewall*) [TG15].

Nuestra propuesta puede ser considerada un WAF que tiene la finalidad de detectar ataques contra aplicaciones web. Cabe mencionar que debido a que WAF es un término más específico que IDS, muchos de los conceptos expuestos a continuación aplican a ambos términos, pero nosotros nos enfocamos únicamente en WAF en este trabajo.

Los WAF pueden utilizar dos métodos distintos para la detección de intrusiones. Una forma puede ser la búsqueda de patrones de ataques conocidos, llamado también método basado en firmas de ataques (*signature-based detection*). Otro método empleado es la búsqueda de desviaciones o anomalías en el tráfico HTTP, ya que estas pueden indicar ataques (*anomaly-based detection*) [TG15].

Para que un WAF pueda utilizar eficazmente el método por firmas, es necesario que el mismo mantenga una lista actualizada de las firmas de todos los ataques conocidos. Esto resulta en un aumento del uso de tiempo y recursos por parte de los WAF al momento de analizar el tráfico de red, ya que la lista de ataques descubiertos crece y probablemente nunca deje de crecer [KV03].

El método de detección de anomalías no requiere una lista de firmas, sino que en su fase de entrenamiento establece modelos que representan al tráfico normal. Se basa en la premisa de que la mayoría de los ataques se diferencian en alguna forma del tráfico normal. En adelante usamos el término anomalía para referirnos a los ataques, para ser consistentes con la literatura relacionada. Así, durante la fase de monitoreo, este método compara el tráfico con los modelos establecidos anteriormente con el fin de detectar desviaciones significativas, es decir, aquellos paquetes HTTP que son considerados anomalías [KV03]. La fase de entrenamiento es obligatoria una vez al inicio del uso del sistema y después solamente si hay cambios en el tráfico normal, por ejemplo después de una actualización a una de las aplicaciones web que protege el WAF en cuestión.

El método por anomalías tiene la ventaja de poder detectar anomalías novedosas desde el momento que aparezcan, mientras que los métodos por firmas dependen de la actualización de su lista de ataques conocidos [KV03].

A pesar de esto, WAF basados en anomalías no son tan comunes como aquellos basados en firmas [SP10]. Esto se debe en parte a que suele ser más complicado establecer modelos significativos para diferenciar muestras normales de anómalas y, como consecuencia, hay menos posibilidades de detectar eficazmente las anomalías. De esta manera los métodos por anomalías corren el peligro de caer en extremos. Por un lado, si se concentran en detectar todas las anomalías, pueden marcar equivocadamente muestras normales como anómalas (más errores de falsos negativos). Por otro lado, si los métodos priorizan no bloquear ninguna muestra normal, puede que muchas anomalías no sean detectadas (más errores de falsos positivos) [TG15].

Nuestra propuesta es un WAF basado en detección de anomalías en el tráfico HTTP, buscando mejorar algunas de las propuestas que ya han sido presentados por otros investigadores, como por ejemplo Kruegel [KV03], Giménez [Gim15] y Torrano-Giménez [TG15].

Para la detección de anomalías se puede utilizar varias estrategias. Una opción es emplear herramientas estadísticas, como podemos ver en los trabajos de Kruegel [KV03], Giménez [Gim15] y Torrano-Giménez [TG15]. También se puede utilizar herramientas del área de aprendizaje de máquinas (ML - *Machine Learning*) [TG15] para tratar de detectar las anomalías. Podemos ver ejemplos de estos en los trabajos de Sommer [SP10], Buczak [BG16], Parhizkar [PA15] y también en el trabajo ya mencionado de Torrano-Giménez [TG15].

Las herramientas de ML han sido empleadas con mucho éxito en varias áreas de la computación, como por ejemplo en sistemas de recomendación de productos, clasificación de imágenes, reconocimiento óptico de caracteres, entre otros [TG15].

Una de las áreas de ML son los problemas de clasificación, y la detección de anomalías se puede encarar como un problema de esta área. ML puede utilizar varias herramientas para clasificar los datos de entrada en varios grupos o clases. En este contexto se habla de aprendizaje supervisado si se especifican todas las clases posibles de antemano, usando solamente muestras para el entrenamiento de las que se conocen sus clases. Muestras nuevas serán asignadas a la clase a la que más se parezcan. Se ha-

bla de aprendizaje no supervisado cuando no se provee muestras con clases conocidas de antemano y la herramienta trata de encontrar las clases presentes en las muestras. Este segundo caso está también estrechamente relacionado con los problemas de agrupamiento (*clustering*) [TG15].

Aplicado a un WAF, se puede usar clasificación supervisada con una clase para tráfico normal y otra (o también varias otras) para tráfico anómalo. Un primer desafío con este abordaje es que se necesita volver a entrenar el clasificador cuando aparece un nuevo tipo de anomalía. Si no se vuelve a entrenarlo con los nuevos tipos de anomalías, es posible que el mismo clasifique equivocadamente una anomalía como normal en el caso de una anomalía nueva que no se ajusta suficientemente a las clases de anomalías vistas anteriormente por el clasificador. Un segundo desafío es la necesidad de obtener muestras de todos los tipos de anomalías conocidas para realizar un entrenamiento completo.

Estos dos desafíos se trata de superar con la estrategia conocida bajo el nombre de clasificación de una sola clase (OCC - *One-Class Classification*) [KM09]. Se busca definir una sola clase, la clase positiva, y clasificar las muestras de acuerdo a si pertenecen o no a dicha clase. La fase de entrenamiento utiliza solamente muestras de la clase positiva, de forma que una muestra que no se ajuste a la clase positiva sea clasificada como no perteneciente a la misma (clase negativa). Esto provee robustez frente a la aparición de muestras negativas nuevas. Esta estrategia ha sido utilizada con éxito en varias áreas, como detección de spam, reconocimiento de rostros, detección de fallas en maquinarias, entre otros [KM09]. Aplicado a un WAF, la clase positiva estará conformada solamente por el tráfico normal y todos los tipos de anomalías no pertenecerán a dicha clase. Además, una gran ventaja con este abordaje es que no se necesita muestras anómalas para el entrenamiento.

Para nuestra propuesta implementaremos un WAF que emplea OCC con herramientas de ML para detectar tráfico HTTP anómalo. Con este abordaje solamente se necesita entrenar una vez el clasificador con tráfico normal y, mientras no cambie la aplicación, no debería ser necesario volver a entrenarlo, aún con la aparición de nuevas anomalías.

Los algoritmos o herramientas utilizados en ML son muy diversos, como por ejemplo árboles de decisiones, redes neuronales, algoritmos genéticos, entre otros [TG15]. Una de estas herramientas, que ha sido utilizada con mucho éxito en las tareas de clasificación, es la máquina de vectores de soporte (SVM - *Support Vector Machine*). Una versión modificada del SVM ha sido propuesta como una de varias alternativas para afrontar tareas de OCC [SPST$^+$01]. Varios investigadores ya han empleado exitosamente este clasificador *One-Class* SVM en problemas de distintas áreas, como por ejemplo clasificación de textos, clasificación de rostros, detección de spam, detección de fallas en máquinas, entre otros [KM09].

En este trabajo utilizaremos un *One-Class* SVM para detectar tráfico HTTP anómalo, combinando propuestas de varios investigadores para obtener una alta eficacia del clasificador.

Para que un WAF basado en detección de anomalías pueda diferenciar el tráfico HTTP normal del anómalo, es necesario que existan características del tráfico que posibiliten esa diferenciación. Ejemplos de esos rasgos pueden ser la longitud de la petición, la aparición de ciertos caracteres con significado especial, entre otros [KV03] [NTGA+11]. Además se debe expresar esos rasgos en un formato procesable para las herramientas de detección. La mayoría de las herramientas de ML no pueden trabajar con los datos crudos y necesitan un paso de preprocesamiento de datos. Asumiendo la existencia de esas características distintivas, el éxito del WAF depende de encontrar dichos rasgos y de representarlos en una forma entendible para el mecanismo de detección [TG15]. En esta parte el conocimiento experto sobre el tráfico HTTP ayuda a seleccionar las características más útiles para ese fin. Vemos un ejemplo de esta selección en los trabajos de Kruegel [KV03] [KVR05], donde el autor utiliza el conocimiento sobre la estructura de paquetes HTTP para obtener rasgos más específicos, que denomina modelos de anomalías, y así mejorar la detección.

En el área de ML, las características de los datos de entrada se representan con números, llamados también *features*. Por ejemplo para el caso de un WAF, el primer número puede indicar la cantidad de caracteres de la petición HTTP, el segundo la cantidad de dígitos presente y el tercero puede representar la entropía calculada a partir de toda la petición. De esta forma, la eficacia de detección de anomalías de nuestro clasificador *One-Class* SVM depende en gran parte de nuestros procesos de *feature extraction*, es decir, de los procesos de preprocesamiento de datos que extraen las características distintivas del tráfico HTTP y las representan con números [TG15].

En este trabajo utilizaremos conocimiento experto sobre tráfico HTTP para extraer características útiles para la detección de anomalías, basandonos en los aportes de Kruegel y otros autores con trabajos relacionados. Ya que trabajamos con un clasificador *One-Class* SVM, en el resto del trabajo utilizamos el término *features* para referirnos a las características mencionadas, para ser consistentes con la terminología del área de ML.

2. Justificación

Resumiendo lo expuesto anteriormente, en esta investigación combinamos tres áreas de estudio para proponer una solución a la problemática descrita. Dichas áreas se pueden observar también en la Figura 1.

- IDS con en detección de anomalías: nuestra propuesta busca detectar posibles ataques al reconocerlos como tráfico anómalo. Utilizaremos el método de detección de anomalías debido a las ventajas mencionadas en la sección anterior.
- Caracteristicas del tráfico HTTP: utilizar conocimiento experto sobre la estructura del tráfico HTTP puede ayudar para diferenciar las peticiones normales de las anomalías o ataques, como se puede ver en los trabajos de Kruegel [KV03] [KVR05].

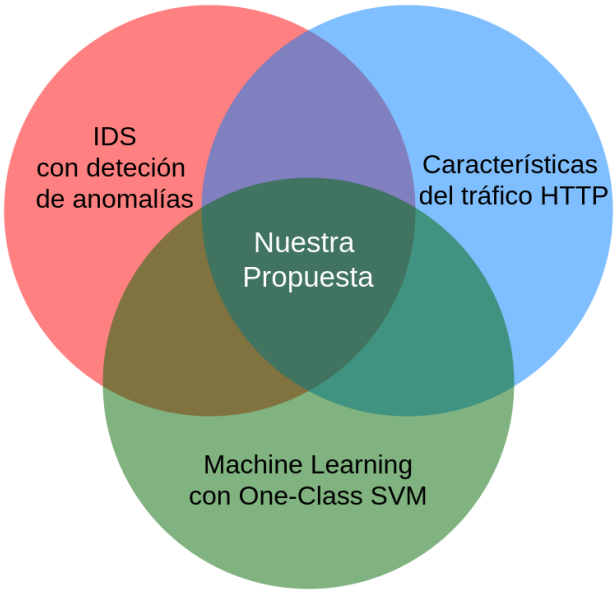

Figura 1: Diagrama de las áreas de estudio de la investigación

- ML con *One-Class* SVM: utilizamos este clasificador, debido a que otros investigadores obtuvieron buenos resultados al aplicarlo a distintos problemas de OCC [KM09] y la detección de tráfico anómalo se puede enfocar como un problema de OCC.

En los trabajos de Kruegel se combina las dos primeras áreas descritas anteriormente, IDS y características del tráfico HTTP. Sus ideas ya fueron aplicadas en varios trabajos en años subsecuentes, pero no hemos encontrado una investigación que las combine con un clasificador *One-Class* SVM.

En nuestra opinión un WAF que combina las ideas de Kruegel con este clasificador puede ser de gran utilidad para la protección de aplicaciones web, y eso buscamos confirmar en el marco de este trabajo.

3. Objetivos

3.1. Objetivo general

Detectar tráfico HTTP anómalo entre aplicaciones web y sus usuarios con el fin de mitigar los riesgos de ataques contra dichas aplicaciones, utilizando un WAF basado en *One-Class* SVM.

3.2. Objetivos específicos

1. Diseñar conjuntos de características (*features*) específicas para tráfico HTTP basado en aportes de otros investigadores de la literatura.

    - Se diseñará nuevos conjuntos de *features* para tráfico HTTP, partiendo de los trabajos de Kruegel y combinando sus ideas con trabajos de la literatura especializada.

2. Evaluar la eficacia de un WAF basado en *One-Class* SVM para detectar tráfico HTTP anómalo.

    - Se construirá un WAF, utilizando los nuevos conjuntos de *features* diseñados y un clasificador *One-Class* SVM. Después se evaluará mediante distintas pruebas la eficacia de detección de tráfico anómalo de esa implementación sencilla.

3. Analizar la viabilidad de utilizar el WAF propuesto para detección de ataques en tiempo real.

    - Se analizará mediante una implementación sencilla en qué medida los conjuntos de *features* propuestos y el clasificador seleccionado afectan al tiempo de respuesta de las distintas aplicaciones web que están siendo protegidas por el WAF.

4. Descripción de la propuesta

Nuestra propuesta en el marco de este trabajo consiste en un WAF que puede ser colocado frente a varias aplicaciones web con el fin de monitorear todo el tráfico HTTP entre dichas aplicaciones y sus usuarios. En la Figura 2 se puede observar la arquitectura general de nuestra propuesta.

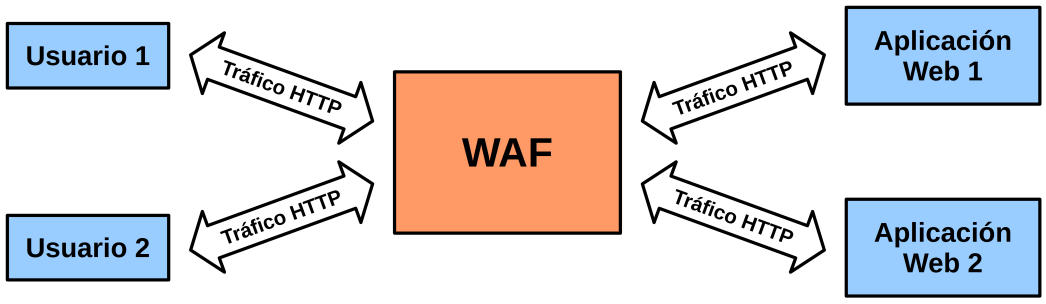

Figura 2: Diagrama de la arquitectura general de la propuesta

La implementación tendrá la opción de habilitar o deshabilitar la detección en tiempo real. Con la detección deshabilitada, el WAF se limitará a registrar el tráfico para que posteriormente se pueda realizar el entrenamiento del clasificador con muestras normales de tráfico. Una vez entrenado, se podrá habilitar el modo de detección. En este modo el WAF analizará todas las peticiones entrantes y registrará todas aquellas que sean clasificadas como anómalas. Se podrá configurar acciones adicionales que el WAF deberá realizar como respuesta a una anomalía detectada, siendo una opción el bloqueo de la petición en cuestión para evitar que posibles ataques lleguen hasta las aplicaciones.

5. Alcance y limitaciones

Según nuestro relevamiento inicial, el trabajo de Kruegel [KV03] es la investigación pionera relacionada con IDS que se enfoca en las particularidades del tráfico HTTP, especialmente el ánalisis de los parámetros y valores dentro de la petición. Con más de 600 citas hasta la fecha según Google Scholar, este trabajo ha sido utilizado como base para muchas investigaciones posteriores.

De esta manera, la investigación del estado del arte que haremos en el marco de nuestro trabajo se limitará a las publicaciones que citan a Kruegel en sus referencias.

Para la evaluación cuantitativa de nuestro WAF propuesto necesitamos conjuntos de datos. Según [TG15] los conjuntos utilizados en las investigaciones sobre WAF deberían cumplir las siguientes características:

- Deberían ser públicamente accesibles para que varios investigadores los utilizen y puedan comparar resultados.
- Deberían contener tráfico HTTP.
- Deberían contener muestras etiquetadas según sean normales o anómalas, o inclusive especificar el tipo de ataque al que pertenecen.
- Deberían tener ataques novedosos.

Los trabajos relacionados sobre WAF utilizan distintos conjuntos de datos para sus pruebas, pero la mayoria de esos conjuntos no cumple una o varias de las características mencionadas. Encontramos unicamente dos conjuntos que son adecuados para nuestras pruebas y se trata de CSIC 2010 [TGPVlM10] y CSIC TORPEDA 2012 [TGPl12]. Así podemos realizar comparaciones con los resultados de otros trabajos que utilizan estos conjuntos.

La implementación de un WAF que haremos en el marco de este trabajo busca ser funcional y sencilla. Se trata de una prueba de concepto, y no buscamos obtener una aplicación terminada que incluya todas las funcionalidades necesarias para utilizarla directamente en ambientes de producción.

6. Plan de actividades

A continuación se detallan las actividades que nosotros nos proponemos para el desarrollo de este trabajo de investigación. Además, en la Tabla 1 se puede observar las duraciones estimadas para dichas actividades.

1. Exploración bibliográfica sobre las áreas de IDS, WAF y OCC, profundizando especialmente en trabajos que utilizan el clasificador *One-Class* SVM y aquellos que basan sus *features* en los trabajos de Kruegel.

2. Diseño de nuevos conjuntos de *features* específicos para tráfico HTTP, combinando las ideas de Kruegel con aportes propios y también de otros investigadores.

3. Construcción de un WAF con los *features* diseñados y con un clasificador *One-Class* SVM.

4. Pruebas de eficacia de detección del WAF construido, buscando conjuntos de *features* y configuraciones del clasificador que mejoren la detección.

5. Pruebas de rendimiento del WAF construido, analizando en qué medida el mismo afecta al tiempo de respuesta de los sistemas que debe proteger.

6. Análisis de resultados y formulación de conclusiones.

Año	2017																	
Mes	Agosto					Setiembre				Octubre				Noviembre				
Semana	1	2	3	4	5	1	2	3	4	1	2	3	4	1	2	3	4	5
Actividad 1	x	x	x	x	x	x	x											
Actividad 2					x	x	x											
Actividad 3								x	x	x	x							
Actividad 4												x	x	x				
Actividad 5												x	x	x				
Actividad 6															x	x	x	x

Tabla 1: Duración estimada de actividades

Referencias

[BG16] Anna L Buczak and Erhan Guven. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 2016.

[FGM+99] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. Rfc 2616: Hypertext transfer protocol – http/1.1. Technical report, RFC Editor, United States, 1999.

[Gim15] José Giménez. Http-ws-ad: Detector de anomalías orientada a aplicaciones web y web services. Universidad Nacional de Asunción, 2015.

[KM09] Shehroz S Khan and Michael G Madden. A survey of recent trends in one class classification. In *Irish Conference on Artificial Intelligence and Cognitive Science*. Springer Berlin Heidelberg, 2009.

[KV03] Christopher Kruegel and Giovanni Vigna. Anomaly detection of web-based attacks. In *Proceedings of the 10th ACM conference on Computer and communications security*. ACM, 2003.

[KVR05] Christopher Kruegel, Giovanni Vigna, and William Robertson. A multi-model approach to the detection of web-based attacks. *Computer Networks*, 2005.

[NTGA+11] Hai Thanh Nguyen, Carmen Torrano-Gimenez, Gonzalo Alvarez, Slobodan Petrović, and Katrin Franke. Application of the generic feature selection measure in detection of web attacks. In *Computational Intelligence in Security for Information Systems*. Springer, Berlin, Heidelberg, 2011.

[PA15] Elham Parhizkar and Mahdi Abadi. Oc-wad: A one-class classifier ensemble approach for anomaly detection in web traffic. In *2015 23rd Iranian Conference on Electrical Engineering (ICEE)*. IEEE, 2015.

[SM07] Karen A. Scarfone and Peter M. Mell. Sp 800-94. guide to intrusion detection and prevention systems (idps). Technical report, National Institute of Standards & Technology, Gaithersburg, MD, United States, 2007.

[SP10] Robin Sommer and Vern Paxson. Outside the closed world: On using machine learning for network intrusion detection. In *Security and Privacy (SP), 2010 IEEE Symposium on*. IEEE, 2010.

[SPST+01] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 2001.

[TG15] Carmen Torrano-Giménez. *Study of stochastic and machine learning techniques for anomaly-based web attack detection*. PhD thesis, Universidad Carlos III de Madrid, 2015.

[TGPl12] Carmen Torrano-Giménez, Alejandro Pérez, and Gonzalo Álvarez. Csic torpeda 2012 http data sets. http://www.tic.itefi.csic.es/torpeda, 2012. Accessed: July-2017.

[TGPVlM10] Carmen Torrano-Giménez, Alejandro Pérez Villegas, and Gonzalo Álvarez Marañón. Csic 2010 http data sets. http://www.isi.csic.es/dataset/, 2010. Accessed: July-2017.