Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 13, No. 4, Mes Abril, 2020
ISSN: 2306-2495 | RNPS: 2343
Pág. 39-52
http://publicaciones.uci.cu

Tipo de artículo: Artículo original
Temática: Educación a Distancia y Tecnologías para la Educación
Recibido: 15/12/19 | Aceptado: 18/03/2020 | Publicado: 02/04/2020

Análisis y caracterización de conjuntos de datos para detección de intrusiones

Analysis and characterization of data sets for intrusion detection

Olia Castellanos Leyva [1*], Milton García Borroto [2]

[1] Empresa Tecnologías de la Información para la Defensa. Calle 296 entre Ave. 207 y 203. Boyeros, La Habana, Cuba. oleyva@xetid.cu
[2] Departamento de Informática. Facultad de Informática. Universidad Tecnológica de La Habana "José Antonio Echeverría". Calle 114 # 11901 / Ciclovía y Rotonda. Marianao, La Habana, Cuba

[*] Autor para correspondencia: oleyva@xetid.cu

Resumen

La variedad de incidentes de seguridad que aparecen en las redes cada día hace que los investigadores deban renovar constantemente sus propuestas de solución. La detección de intrusiones es un problema de seguridad que evidencia esta problemática. Dentro de esta área la detección de anomalías modela el comportamiento normal de la red e identifica las anomalías como desviaciones de dicho modelo. Aunque existen numerosas investigaciones sobre este tema, aún queda mucho por hacer para elevar los niveles de detección y disminuir el nivel de falsos positivos. Uno de los mayores retos para lograr un sistema de detección de anomalías eficaz y eficiente reside en la elección de un conjunto de datos exhaustivo y realista que permita su evaluación. El conjunto de datos que se escoja debe reflejar los escenarios actuales del tráfico de red, proveer información estructurada del mismo y poseer variedad de intrusiones. El conjunto de datos más utilizado por la comunidad científica ha sido KDDCUP99 y sus distintas versiones. Sin embargo, existen numerosos argumentos por los que han ido surgiendo otros conjuntos de datos para sustituirlo. Con la intención de lograr el conjunto de datos más representativo se han desarrollado propuestas como ISCX2012, UNSW-NB15 y CICIDS2017, entre muchas otras. El presente trabajo pretende proporcionar un análisis y caracterización de los conjuntos de datos más utilizados, de manera que pueda emplearse como punto de partida para elegir aquel que mejor se ajuste a las necesidades de cada investigación y que refleje el comportamiento actual del tráfico de las redes.

Palabras clave: intrusiones; detección de anomalías; conjunto de datos

Abstract

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 13, No. 4, Mes Abril, 2020
ISSN: 2306-2495 | RNPS: 2343
Pág. 39-52
http://publicaciones.uci.cu

The variety of security incidents that appear on the networks every day causes researchers to constantly renew their proposed solutions. Hence, intrusion detection is a security problem that evidences this problem. Within this area, anomaly detection models normal network behavior and identifies anomalies as deviations from that model. Although there are numerous investigations on this subject, there is still much to do to raise the levels of detection and decrease the level of false positives. One of the biggest challenges in achieving an efficient anomaly detection system is the choice of a comprehensive and realistic data set that allows its evaluation. The dataset chosen should reflect the current scenarios of network traffic, provide structured information and have a variety of intrusions. The dataset most used by the scientific community has been KDDCUP99 and its different versions. However, there are numerous arguments for which other datasets have emerged to replace it. With the intention of achieving the most representative dataset, proposals such as ISCX2012, UNSW-NB15 and CICIDS2017, among many others, have been developed. The present work intends to provide an analysis and characterization of the most used data sets, so that it can be used as a starting point to choose the one that best suits the needs of each investigation and that reflects the current behavior of network traffic.

Keywords: *intrusions; anomaly detection; dataset*

Introducción

Desde la consolidación de Internet como medio de interconexión global, los incidentes de seguridad en la red vienen incrementándose de manera alarmante. Este hecho, unido a la progresiva dependencia de las organizaciones hacia sus sistemas de información, ha provocado la necesidad de implantar mecanismos de control ante eventos que pongan en peligro su seguridad. Los sistemas de detección de intrusiones (IDS) forman parte de estos mecanismos y se emplean con el objetivo de identificar eventos que puedan considerarse amenazas. Estos sistemas pueden emplearse para proteger una red de computadoras o un host en particular y de acuerdo al enfoque que utilicen pueden detectar comportamientos anormales o tipos de ataques específicos.

Muchos son los sistemas de detección de intrusiones que se han desarrollado utilizando para ello las más diversas técnicas y, sin embargo, los investigadores concuerdan en que aún queda mucho por hacer para elevar los niveles de detección y disminuir la cantidad de falsos positivos de estas herramientas. La variedad de incidentes de seguridad que aparecen cada día y la complejización de los entornos de redes de computadoras hace que los investigadores deban renovar constantemente sus propuestas de solución.

Entre los elementos de mayor importancia para desarrollar un IDS se encuentra la elección del conjunto de datos a utilizar. Este debe ser lo más exhaustivo y realista posible de modo que permita al IDS entrenar el reconocimiento de

ataques o comportamientos anormales en la red y luego evaluar su desempeño. El conjunto de datos que se escoja debe reflejar los escenarios actuales del tráfico de red, proveer información estructurada del mismo y poseer variedad de intrusiones.

Desde la publicación del primer conjunto de datos formalizado para detección de intrusiones, DARPA98, se han publicado diversos trabajos que recogen la actividad de redes de computadoras en disímiles ámbitos, con configuraciones variables y con la presencia de diferentes eventos y ataques. El conjunto de datos KDDCUP99 y sus distintas versiones han sido los más utilizados por la comunidad científica (Özgür y Erdem, 2016). No obstante, (McHugh, 2000), (Tavallaee y otros, 2009), (Creech y Hu, 2013), (Siddique y otros, 2019) exponen numerosos argumentos por los que ha sido necesario que surgieran otros conjuntos de datos para sustituirlo. Con la intención de lograr el conjunto de datos más representativo se han desarrollado propuestas que en los últimos años han sido más numerosas. El presente trabajo caracteriza los conjuntos de datos más utilizados según las investigaciones consultadas y aquellos que han surgido en los últimos años. Por otro lado, proporciona una caracterización de los mismos de manera que pueda emplearse como punto de partida para elegir aquel que mejor se ajuste a las necesidades de cada investigación.

Materiales y métodos

Para desarrollar la presente investigación se analizaron los conjuntos de datos más utilizados por los investigadores en la evaluación del desempeño de los modelos propuestos para detectar anomalías en el tráfico de red. Se tuvieron en cuenta las investigaciones de este tipo que se han publicado en los últimos años y los criterios que comúnmente se utilizan para comparar o evaluar los conjuntos de datos. Se seleccionaron, además, los conjuntos de datos que recogen la información del tráfico de una red y no de un host en particular.

Trabajos relacionados

Aunque se han realizado importantes estudios sobre la generación de conjuntos de datos para detección de intrusiones, no son muchos los trabajos sobre la evaluación y valoración de los mismos de manera general. (Gharib y otros, 2016) En la mayoría de los casos, cuando se publica la recolección de datos para detección de intrusiones, se ofrecen comparaciones en las que se resaltan características muy particulares del nuevo conjunto datos y/o se compara con pocos conjuntos de datos de todos los existentes. Además, es importante tener en cuenta que en los últimos años

se han desarrollado nuevas recolecciones de datos en aras de lograr una mejor representación del comportamiento de las redes modernas y los ataques que en ellas se ejecutan y que, por lo tanto, deben tenerse en cuenta.

En (Nehinbe, 2011) se realiza una evaluación crítica de las principales deficiencias que presentan los conjuntos de datos para detección y prevención de intrusiones. Se detallan además las dificultades que comúnmente se enfrentan en la creación de un conjunto de datos de este tipo y sus principales fuentes de obtención. Su contribución fundamental se basa en la sugerencia de posibles métodos que se pueden adoptar para mitigar los problemas descritos. En el análisis se incluyen los conjuntos de datos DARPA, DEFCON, CAIDA, KDD, NSL-KDD, LBNL, SLINGbot y UMass. Estos son catalogados como conjuntos de datos de repositorios públicos y se caracterizan en cuanto a su disponibilidad a la comunidad científica, la existencia de varias versiones de los mismos y los elementos generales de las situaciones en las que se obtuvieron los datos.

En el trabajo presentado por (Shiravi y otros, 2012) se comparan los conjuntos de datos CAIDA, Internet Traffic Archive, LBNL, DARPA-99, KDD-99, DEFCON y ISCX2012. La caracterización presentada de cada uno se basa en la presencia de una configuración de red y datos del tráfico lo más realista posible, el etiquetado de los datos, la captura de todas las interacciones de red, la recolección de todas las trazas y la diversidad de escenarios de ataques. El principal aporte de este trabajo es la publicación del conjunto de datos ISCX2012, cuya valoración según los criterios planteados es bastante buena.

Por su parte (Gharib y otros, 2016) proveen una revisión de conjuntos de datos generados entre 1998 y 2016. Propone una guía para evaluarlos basándose en la configuración de red utilizada para generar el tráfico; la completa recolección de los paquetes de red; el etiquetado de los datos; la recolección de todas las interacciones de red; la captura del tráfico completo, sin que se remuevan partes del mismo; los protocolos disponibles en los datos recolectados; la presencia de diversidad de ataques; la anonimización de datos; la heterogeneidad de las fuentes del tráfico recolectado; la presencia de características calculadas y extraídas de los datos, así como su fundamentación; además de la documentación con una explicación detallada de los entornos, configuraciones de red, sistemas operativos, escenarios de ataques, entre otros elementos utilizados para la generación del conjunto de datos. Se analizan los conjuntos de datos: DARPA, KDD'99, DEFCON, CAIDA, LBNL, CDX, KYOTO, TWENTE, UMASS, ISCX2012 y ADFA2013. Por otro lado, como parte de la metodología propuesta se incluye una fórmula para evaluar los conjuntos de datos según los criterios presentados. Aunque en el trabajo solo se evalúan los conjuntos KDD'99 y

KYOTO. Como extensión de esta investigación se presenta en (Sharafaldin y otros, 2018) un nuevo conjunto de datos, CICIDS2017, que es incluido en la misma comparación presentada por su predecesor.

(Siddique y otros, 2019) formaliza una crítica exhaustiva sobre la familia de conjuntos de datos KDDCUP99, DARPA y NSL-KDD; y basándose en las características propuestas por (Shiravi y otros, 2012) realiza una comparación con otros conjuntos de datos que consideran mejores opciones. Estos son: NGIDS-DS, ISCX-UNB, TUIDS, UNSW-NB15 y MAWILab. Se compara, además, su desempeño ante algunas técnicas/algoritmos de aprendizaje automatizado. A los criterios utilizados se suma el tipo de tráfico recogido en los conjuntos de datos: normal, malicioso o mixto; y la presencia de datos anonimizados.

El presente trabajo ofrece una caracterización de 10 conjuntos de datos utilizados en el diseño, desarrollo y evaluación de IDSs, de los cuales 4 fueron generados en los últimos 5 años. El mismo pretende realizar una caracterización general que muestre aquellos rasgos que los diferencian y que permita a los investigadores valorar si son apropiados o no para el objetivo de sus estudios. Por otro lado, procura incorporar rasgos que resalten cuánto se ajusta cada conjunto de datos a las particularidades de las redes modernas. Para ello se incluyen criterios utilizados en la bibliografía consultada, aunque algunos se fusionan para formar características más generales, y se adicionan otros que se consideran importantes para la evaluación de un IDS. La mayoría de los conjuntos de datos que se analizan fueron escogidos dada su notable utilización por la comunidad científica en el desarrollo de IDS. Otros fueron seleccionados por ser propuestas más recientes y no estar incluidas en los trabajos consultados o encontrarse en análisis dispersos como parte de otras investigaciones. Se consideró importante reunirlos todos en un mismo estudio, de modo que pueda servir de guía y consulta.

Criterios utilizados

- Año: Dado que nuevos escenarios de ataques surgen diariamente se consideró importante incluir el año de creación del conjunto de datos en aras de lograr actualidad en cuanto a los patrones de comportamiento de las redes de computadoras.
- Disponibilidad: La disponibilidad del conjunto de datos es otro elemento que también se consideró importante ya que, en aras de comparar diferentes métodos de detección de intrusiones, los datos deben estar al alcance de los investigadores. En la caracterización establecida este criterio toma valores de: público si el conjunto de

datos está totalmente disponible o por solicitud si la única vía de acceder al mismo es a través del envío de correo a la(s) personas/entidades pertinentes.

- Documentación: Para lograr una correcta y completa interpretación de un conjunto de datos es importante contar con la descripción de elementos como la estructura de red utilizada en la recolección del mismo, escenarios de ataques presentes, entre otras. De ahí que la disponibilidad de documentación que ayude a la comprensión del contenido de los datos es importante. (Gharib y otros, 2016)
- Formato: Los conjuntos de datos de detección de intrusiones en la red aparecen en variados formatos, así que para lograr una visión objetiva se dividieron en tres formatos: tráfico de red basado en paquetes que incluye el tráfico con la carga útil, tráfico basado en flujos que incluye sólo la información general de las conexiones de red y en algunos casos se incluyen combinaciones de los formatos anteriores u otros como registros (logs de un host).
- Anonimato: De acuerdo a las particularidades de cada conjunto de datos como el entorno de red donde fue obtenido, pueden retirarse o mantenerse anónimas algunas características o valores como direcciones IP o la carga útil. Este criterio establece la presencia o no de este fenómeno en cada conjunto de datos analizado. (Gharib y otros, 2016)
- Tiempo: Otro criterio a tener en cuenta es el tiempo durante el cual se efectuó la recogida de los datos, ya que es un elemento a tener en cuenta si se desean observar efectos periódicos como comportamientos en horarios específicos del día, fines de semana, entre otros. (Maciá y otros, 2018)
- Realismo del tráfico: El realismo del tráfico describe los orígenes de los datos recolectados como real o emulado. Real significa que el tráfico de red real se capturó dentro de un entorno de red productivo. Emulado significa que el tráfico de red real se capturó dentro de un banco de pruebas o entorno de red emulado. (Gharib y otros, 2016)
- Tráfico completo: Esta propiedad indica si el conjunto de datos contiene el tráfico de red completo de un entorno de red con varios hosts, enrutadores, etc. Si el conjunto de datos contiene solo tráfico de red de un solo host o solo algunos protocolos del tráfico de red, el valor se establece en no. (Gharib y otros, 2016)
- Subconjuntos: Para los investigadores es importante poder comparar el desempeño de varios IDS, pero aún comparándolos con los mismos datos puede resultar difícil si no se realiza el entrenamiento y evaluación con los mismos subconjuntos. Por lo tanto, esta propiedad refleja la presencia o no de subconjuntos predefinidos para realizar el entrenamiento y evaluación.

- Etiquetado: Según el método de detección de intrusiones que se emplee (supervisado o no supervisado) el etiquetado de los datos pasa a ocupar un rol determinante, ya que es necesario para entrenar métodos supervisados y para evaluar el desempeño de los no supervisados. Esta característica toma valor positivo si en el conjunto de datos existen al menos dos clases para clasificar los datos como normal y ataque. (Shiravi y otros, 2012)

Resultados y discusión

El aumento de los escenarios de ataques, la evolución de las estructuras de red y la creación de software cada vez más complejos reafirman la necesidad de crear conjuntos de datos actualizados, de modo que permitan desafiar los IDS desarrollados con una evaluación más cercana a los escenarios reales. Durante mucho tiempo se han venido utilizando, como base para las investigaciones, conjuntos de datos que ya acumulan 20 años de antigüedad, por lo que se han desarrollado nuevas recolecciones de datos. Sin embargo, para lograr la correcta evaluación de los IDS desarrollados y sus comparaciones objetivas es necesario que los investigadores hagan uso de propuestas más actuales, como las expuestas en el presente trabajo.

Conjuntos de datos

DARPA 1998/99 se han convertido en los conjuntos de datos más utilizados para evaluar el comportamiento de los sistemas de detección de intrusiones. Creados en el Massachusetts Institute of Technology (MIT) Lincoln Lab, recogen 7 y 5 semanas de datos de tcpdump sin procesar, respectivamente. Incluyen cuatro tipos de ataques: probe, remote to local, denial of service (DoS), y user to root. Todas las versiones de este conjunto de datos se pueden encontrar en https://www.ll.mit.edu/r-d/datasets (Lippmann y otros, 2000)

KDDCUP 99 se basa en el conjunto de datos DARPA y se encuentra entre los conjuntos de datos más extendidos para la detección de intrusos. El conjunto de datos contiene atributos básicos sobre conexiones TCP y atributos de alto nivel, como el número de inicios de sesión fallidos, pero no direcciones IP. KDD CUP 99 abarca más de 20 tipos diferentes de ataques (por ejemplo, DoS o desbordamiento de búfer) y viene con un subconjunto de prueba explícito. El conjunto de datos incluye 5 millones de puntos de datos y se puede descargar libremente en http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html . (Tavallaee y otros, 2009)

NSL-KDD mejora el KDD CUP 99. Una crítica importante contra el conjunto de datos KDD CUP 99 es la gran cantidad de registros duplicados y redundantes. Por lo tanto, los autores de NSL-KDD eliminaron duplicados del conjunto de datos KDD CUP 99 y crearon subconjuntos más sofisticados. El conjunto de datos resultante contiene alrededor de 150,000 puntos de datos y se divide en subconjuntos predefinidos de entrenamiento y prueba para métodos de detección de intrusos. El mismo puede ser descargado desde https://www.unb.ca/cic/datasets/nsl.html (Tavallaee y otros, 2009)

LBNL se crea con el interés de analizar las características del tráfico de red dentro de las redes empresariales y no con la intención de proporcionar un conjunto de datos para detección de intrusiones. Sin embargo, pudiera utilizarse como tráfico de fondo ya que contiene un comportamiento de usuario casi exclusivamente normal. Contiene más de 100 horas de tráfico de red en formato basado en paquetes, que abarcan desde la reexaminación de temas previamente estudiados para el tráfico de área amplia (por ejemplo, tráfico web), hasta la investigación de nuevos tipos de tráfico no evaluados en la literatura hasta ese momento (por ejemplo, protocolos de Windows). Los archivos que contienen los registros de este conjunto de datos pueden encontrarse en http://www.icir.org/enterprise-tracing (Pang y otros, 2005)

TUIDS está compuesto por un subconjunto de datos de 50 características extraídas del tráfico de red a nivel de paquete y otro subconjunto de datos de 24 de características extraídas del flujo de red. Aunque luego fue publicado también un tercer subconjunto de datos enfocado principalmente en ataques DDoS. Sus datos fueron generados dentro de un entorno emulado con alrededor de 250 clientes. Cada subconjunto abarca un período de siete días y los tres subconjuntos contienen alrededor de 250,000 flujos. (Gogoi y otros, 2012) (Bhuyan y otros, 2015) Para hacer uso de los datos los interesados deben poner se en contacto con los autores.

ISCX2012 es un conjunto de datos generado bajo un enfoque sistemático que se basa en el concepto de perfiles con descripciones detalladas de los escenarios que se desean emular. Según los autores los perfiles α definen escenarios de ataque, mientras que los perfiles β caracterizan el comportamiento normal del usuario, como escribir correos electrónicos o navegar por la web. Este enfoque permite una generación continua de nuevos conjuntos de datos. Durante 7 días se generaron más de 2 millones de registros que contienen varios tipos de ataques como fuerza bruta SSH, DoS o DDoS, que pueden ser obtenidos en https://www.unb.ca/cic/datasets/ids.html . (Shiravi y otros, 2012)

UNSW-NB15 es creado con el uso de la herramienta IXIA Perfect Storm en un pequeño entorno emulado durante 31 horas. La generación de tráfico sintético con comportamiento malicioso incluye nueve tipos de ataques como backdoors, DoS, exploits, worms y fuzzers. Está compuesto por 49 características extraídas con herramientas como Argus, Bro-IDS, entre otras. En el sitio oficial, https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets , los investigadores pueden encontrar 4 archivos de tipo csv que contienen los registros de ataques y comportamiento normal, así como otros archivos con los datos de origen de cada una de las herramientas utilizadas. (Moustafa y otros, 2015)

UGR16 es un conjunto de datos basado en el flujo que centra su atención en capturar efectos periódicos en un entorno ISP. Abarca un período de cuatro meses y contiene 16.900 millones de flujos unidireccionales. Los flujos se etiquetan como normales, de fondo o de ataque y las direcciones IP se anonimizan. No obstante, la mayor parte del tráfico está etiquetado como fondo que podría ser normal o un ataque. UGR16 contiene un conjunto de ataques diseñados específicamente para entrenar y probar algoritmos IDS, entre los que se encuentran DoS, UDP Scan, SSH Scan, SPAM, etc. El tiempo dedicado a la recolección de los datos lo hace adecuado para probar algoritmos que consideran la evolución del tráfico estacionario durante el día / noche y los días de semana / fines de semana. Según los autores, la principal ventaja de este conjunto de datos sobre otros es que el tráfico de fondo es muy representativo del tráfico de Internet, ya que se captura de los sensores en una red ISP donde se encuentran muchos perfiles diferentes de clientes. Para obtener el conjunto de datos se acceder a https://nesg.ugr.es/nesg-ugr16 (Maciá y otros, 2018)

NGIDS-DS2017 contiene tráfico de red en formato basado en paquetes, así como archivos de registro basados en host. Los autores proponen una métrica para evaluar el realismo de los conjuntos de datos para evaluar IDS, que sirvió como base para la creación de NGIDS-DS2017. En un entorno emulado se generó comportamiento normal de usuario, así como ataques de siete familias de ataque diferentes (DoS, worms o reconnaissance). El conjunto de datos etiquetado contiene aproximadamente 1 millón de paquetes y está disponible públicamente en https://research.unsw.edu.au/people/professor-jiankun-hu . (Haider y otros, 2017)

CICIDS 2017 se creó en un entorno emulado durante un período de 5 días, contiene tráfico de red en formato de paquetes y los resultados del análisis del tráfico en flujo bidireccional, obtenidos con la herramienta CICFlowMeter y descritos por 85 características. Incluye una amplia gama de tipos de ataque como SSH, brute force, heartbleed, botnet, DoS, DDoS, web y ataques de infiltración. Para generar el tráfico de fondo se implementaron perfiles de

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 13, No. 4, Mes Abril, 2020
ISSN: 2306-2495 | RNPS: 2343
Pág. 39-52
http://publicaciones.uci.cu

comportamiento abstracto de 25 usuarios basados en protocolos como HTTP, HTTPS, FTP, SSH y protocolos de correo electrónico. Este conjunto de datos se puede obtener en https://www.unb.ca/cic/datasets/ids-2017.html (Sharafaldin y otros, 2018)

En la tabla 1 se muestra la caracterización de los conjuntos de datos analizados en relación a los criterios definidos.

Tabla 1. Caracterización de conjuntos de datos para detección de intrusiones

Conjunto de datos	Año	Disponibilidad	Documentación	Formato	Anonimato	Tiempo	Realismo del tráfico	Tráfico completo	Subconjuntos	Etiquetado
DARPA	1998/ 1999	público	si	paquetes, otros	no	7/5 semanas	emulado	no	si	si
KDDCUP 99	1998	público	si	otros	no	no conocido	emulado	no	si	si
NSL-KDD	1998	público	si	otros	no	no conocido	emulado	no	si	si
LBNL	2004/ 2005	público	no	paquetes	si	5 horas	real	si	no	no
TUIDS	2011	por solicitud	no	paquetes, flujos	no	21 días	emulado	si	si	si
ISCX2012	2011	público	si	paquetes, flujos	no	7 días	emulado	no	no	si
UNSW-NB15	2015	público	si	paquetes, otros	no	31 horas	emulado	si	si	si
UGR16	2016	público	alguna	flujos	si	4 meses	real	no	si	si
NGIDS-DS2017	2016	público	no	paquetes, otros	no	5 días	emulado	si	no	si
CICIDS 2017	2017	público	si	paquetes, flujos	si	5 días	emulado	si	no	si

Los conjuntos de datos caracterizados, por lo general, son liberadas de forma pública a la comunidad científica con un cúmulo de datos que usualmente incluyen los paquetes de red capturados en las interacciones de un entorno que casi siempre es emulado. Un análisis general de ellos muestra que más de la mitad de los conjuntos estudiados fueron

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 13, No. 4, Mes Abril, 2020
ISSN: 2306-2495 | RNPS: 2343
Pág. 39-52
http://publicaciones.uci.cu

generados en los últimos 10 años, a los que se suman otros que por tener propósitos muy específicos no fueron incluidos en el presente estudio.

Dado que los escenarios de ataques presentes en las recolecciones de datos caracterizadas son muy diversos se decidió especificarlos de manera independiente, resultado que se muestra en la tabla 2.

Tabla 2. Escenarios de ataques presentes en los conjuntos de datos

Conjunto de datos	Diversidad de ataques
DARPA	Probe (Ej: portsweep, ipsweep, mscan), DoS (Ej: neptune, smurf, teardrop), R2L (Ej: imap, xsnoop, named), U2R (Ej: sqlattack, perl, xterm), Data (Ej: ppmacro, ntfsdos, secret)
KDDCUP 99	
NSL-KDD	
LBNL	Scan
TUIDS	DDoS (Ej: syn-flood, rst-flood, smurf, fraggle udp-flood, ping-flood, fraggle, icmp-flood, land), Probe (Ej: fin-scan, udp-scan, null-scan, syn-scan, xmasstree-scan), DoS (Ej: teardrop, winnuke, syndrop, newtear, bonk, saihyousen)
ISCX2012	Scan, SSH Brute Force, DDoS, HTTP DoS
UNSW-NB15	Fuzzers, Scan, Spam, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms
UGR16	Port scans, DoS, Botnet, SSH Brute Force, Spam
NGIDS-DS2017	Exploits, DoS, Worms, Generic, Reconnaissance, Shellcode, Backdoors
CICIDS 2017	Brute Force, Heartbleed, Botnet, DoS, DDoS, Web Attack (Ej: Sql Injection, Cross-Site Scripting, HTTP Brute Force), Backdoor, Ipsweep, Port Scan

Al observar los ataques presentes en los conjuntos de datos estudiados se aprecia que las familias de ataques DoS y Scan se incluyen con bastante frecuencia de manera general y, sin embargo, los reportes oficiales los muestran entre los ataques menos recurrentes. (McAfee. Threats Statistics) Además, CICIDS 2017 y UNSW-NB15 presentan la mayor diversidad de comportamientos maliciosos, entre los destacan algunos ejemplos que clasifican como malware.

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 13, No. 4, Mes Abril, 2020
ISSN: 2306-2495 | RNPS: 2343
Pág. 39-52
http://publicaciones.uci.cu

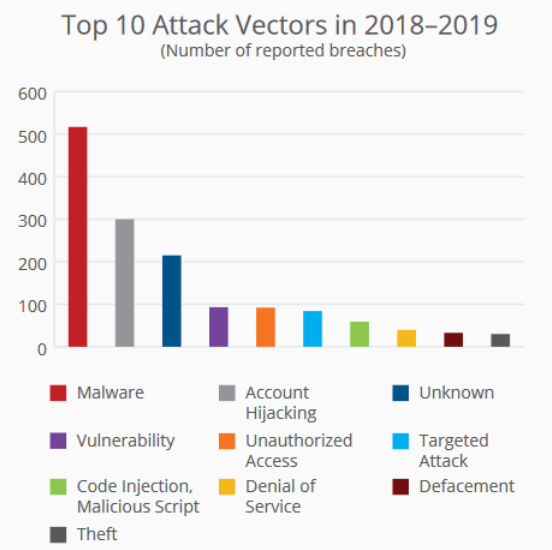

Figura 1. Lista de los 10 vectores de ataques más frecuentes en 2018-2019 según (McAfee. Threats Statistics)

Conclusiones

La cantidad de conjuntos de datos para detección de intrusiones puestos a disposición de los investigadores ha experimentado un aumento paulatino, como parte del esfuerzo por sustituir propuestas muy difundidas en este campo pero que ya se encuentran desactualizadas con respecto a la realidad actual de las redes en cuanto a configuraciones, comportamiento y amenazas a las que se enfrentan.

Los escenarios de ataques presentes en las recolecciones de datos no siempre corresponden con los tipos de ataques más frecuentes reportados por fuentes oficiales. En los últimos años los malware han incrementado su aparición y otros tipos de amenazas han perdido el foco de atención de los atacantes. De ahí que deban estar presentes ejemplos de virus, gusanos, troyanos, rootkits, ransomware, entre otros, en los conjuntos de datos para evaluar el desempeño de los IDS. Aunque es de suma importancia también que se mantenga una diversidad balanceada de otros tipos de ataques, con el objetivo de lograr representatividad de forma proporcional a su uso.

Aunque existen muchos y variados conjuntos de datos, se recomienda el uso de los conjuntos CICIDS 2017 y UNSW-NB15, debido a que estos poseen una amplia variedad de ataques, contienen representaciones actuales de los comportamientos usuales de las redes, incluyen la información del tráfico completo del entorno de red y cuentan con documentación lo suficientemente completa como para lograr la correcta interpretación de su contenido.

Referencias

- BHUYAN, Monowar H.; BHATTACHARYYA, Dhruba K.; KALITA, Jugal K. Towards Generating Real-life Datasets for Network Intrusion Detection. IJ Network Security, 2015, vol. 17, no 6, p. 683-701.
- CREECH, Gideon; HU, Jiankun. Generation of a new IDS test dataset: Time to retire the KDD collection. En 2013 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2013. p. 4487-4492.
- GHARIB, Amirhossein, et al. An evaluation framework for intrusion detection dataset. En 2016 International Conference on Information Science and Security (ICISS). IEEE, 2016. p. 1-6.
- GOGOI, Prasanta, et al. Packet and flow based network intrusion dataset. En International Conference on Contemporary Computing. Springer, Berlin, Heidelberg, 2012. p. 322-334.
- HAIDER, Waqas, et al. Generating realistic intrusion detection system dataset based on fuzzy qualitative modeling. Journal of Network and Computer Applications, 2017, vol. 87, p. 185-192.
- LIPPMANN, Richard, et al. The 1999 DARPA off-line intrusion detection evaluation. Computer networks, 2000, vol. 34, no 4, p. 579-595.
- MACIÁ-FERNÁNDEZ, Gabriel, et al. UGR '16: A new dataset for the evaluation of cyclostationarity-based network IDSs. Computers & Security, 2018, vol. 73, p. 411-424.
- MCHUGH, John. Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. ACM Transactions on Information and System Security (TISSEC), 2000, vol. 3, no 4, p. 262-294.
- McAfee. Threats Statistics. [En línea] McAfee Labs Threats Report, 2019. [Consultado el: 10 de diciembre de 2019] 35-40 p. Disponible en: https://www.mcafee.com/enterprise/en-us/assets/reports/rp-quarterly-threats-aug-2019.pdf
- MOUSTAFA, Nour; SLAY, Jill. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). En 2015 military communications and information systems conference (MilCIS). IEEE, 2015. p. 1-6.
- NEHINBE, Joshua Ojo. A critical evaluation of datasets for investigating IDSs and IPSs researches. En 2011 IEEE 10th International Conference on Cybernetic Intelligent Systems (CIS). IEEE, 2011. p. 92-97.

- ÖZGÜR, Atilla; ERDEM, Hamit. A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015. PeerJ Preprints, 2016, vol. 4, p. e1954v1

- PANG, Ruoming, et al. A first look at modern enterprise traffic. En Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement. USENIX Association, 2005. p. 2-2.

- SHARAFALDIN, Iman, et al. Towards a reliable intrusion detection benchmark dataset. Software Networking, 2018, vol. 2018, no 1, p. 177-200.

- SHARAFALDIN, Iman; LASHKARI, Arash Habibi; GHORBANI, Ali A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. En ICISSP. 2018. p. 108-116.

- SHIRAVI, Ali, et al. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. computers & security, 2012, vol. 31, no 3, p. 357-374.

- SIDDIQUE, Kamran, et al. KDD Cup 99 Data Sets: A Perspective on the Role of Data Sets in Network Intrusion Detection Research. Computer, 2019, vol. 52, no 2, p. 41-51.

- TAVALLAEE, Mahbod, et al. A detailed analysis of the KDD CUP 99 data set. En 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. IEEE, 2009. p. 1-6.