REVISTA MULTIDISCIPLINARIA
DESARROLLO AGROPECUARIO, TECNOLOGICO, EMPRESARIAL Y HUMANISTA
UNIVERSIDAD TECNICA DE COTOPAXI - LA MANA. COTOPAXI. ECUADOR
ISSN: 2773-7527
Volumen X. No. X
mes xxxx

Análisis de Memoria de Malware Ofuscado en el Conjunto de Datos CIC- MALMEM-2022

Memory Analysis of Obfuscated Malware in the CIC-MALMEM-2022 Dataset

Freddy Neptali Chamorro Palacios[1], Edison Javier Guaña Moya[2], Wilson Iván Sánchez Paredes[3]

[1]Universidad Técnica Estatal de Quevedo, Quevedo, Ecuador - ORCID: 0000-0001-6819-3265

[2]Instituto Superior Tecnológico Japón, Quito, Ecuador - ORCID: 0000-0003-4296-0299

[3]Universidad Técnica de Ambato, Ambato, Ecuador - ORCID: 0009-0009-2379-4548

Correo de correspondencia: fchamorrop@uteq.edu.ec, eguana@itsjapon.edu.ec, wilsons92@live.com

Información del artículo

Tipo de artículo: Artículo original

Recibido: 10/09/2023

Aceptado: 24/02/2024
Revista: DATEH

Resumen

El malware de ofuscación de memoria es una técnica sofisticada utilizada por los ciberdelincuentes para evitar la detección por parte de los programas antivirus y dificultar el análisis de la misma por parte de los investigadores de seguridad. El presente trabajo de investigación está basado en un conjunto de datos creado para representar escenarios en un ambiente real. Está compuesto por ransomware, troyanos y spyware, proporcionando un conjunto de datos con el fin de probar sistemas de detección de malware ofuscado. Un análisis inteligente de los datos utilizados para el presente estudio permite encontrar patrones comunes para identificar un malware ofuscado. La investigación planteada propone realizar un exhaustivo análisis para localizar tipos de malwares, relaciones y diferencias significativas que permitan extraer indicadores que puedan revelar la presencia de malware ofuscado en memoria. Existe un equilibrio de los datos entre el volcado de memoria benigno y su contraparte, el malware. Así mismo, el grupo compuesto por ransomware, troyanos y spyware en sus diferentes categorías tiene un alto equilibrio según el análisis realizado.

Palabras clave: Malware, Ofuscación, ransomware, troyano, spyware.

Abstract

Memory obfuscation malware is a sophisticated technique used by cybercriminals to avoid detection by antivirus programs and make it difficult for security researchers to analyze. This research work is based on a data set created to represent scenarios in a real environment. It is composed of ransomware, trojans and spyware, providing a data set for the purpose of testing obfuscated malware detection systems. An intelligent analysis of the data used for the present study allows finding common patterns to identify obfuscated malware. The proposed research proposes to carry out an exhaustive analysis to locate types of malwares, relationships and significant differences that allow extracting indicators that can reveal the presence of malware obfuscated in memory. There is a balance of data between the benign crash dump and its malware counterpart. Likewise, the group made up of ransomware, Trojans and spyware in their different categories has a high balance according to the analysis carried out.

Keywords: Malware, Obfuscation, ransomware, trojan, spyware.

Volumen X. No. X
mes xxxx
ISSN: 2773-7527

INTRODUCCIÓN

El Código ofuscado es sinónimo de seguridad y es una herramienta utilizada para mejorar la seguridad de su código. Generalmente, la ofuscación convierte un archivo, software o proyecto en un tipo de código más difícil de entender para los humanos (Marqués y Somekawa, 2012), (Llalli, 2022).

El malware ofuscado es una amenaza importante en el panorama de la ciberseguridad actual y utiliza sofisticadas técnicas de encubrimiento para evadir la detección tradicional y complicar el análisis de los sistemas de seguridad (Moya, 2023). Este tipo de malware se caracteriza por su capacidad de modificar su código de formas sofisticadas para ocultar su existencia y comportamiento malicioso. Utiliza técnicas de ofuscación como cifrado de comandos, manipulación de datos y cifrado simple, lo que dificulta que los programas antivirus y las herramientas de seguridad reconozcan su patrón y lo neutralicen de manera efectiva.

El uso de la minería de datos para detectar malware oculto se ha convertido en un área importante y en crecimiento de la investigación en ciberseguridad (Guaña et al., 2022). Utilizando técnicas avanzadas de análisis de datos, combinadas con algoritmos de aprendizaje automático y minería de datos, se puede identificar patrones ocultos y comportamientos inusuales en el tráfico y la ejecución de programas. Este enfoque inteligente puede ayudar a revelar las sutilezas del comportamiento del malware oculto y los cambios en el entorno digital, permitiendo respuestas más flexibles y efectivas para proteger los sistemas y las redes contra estas amenazas complejas y esquivas.

Villaroel (2022), en su estudio denominado "Revisión de algoritmos de detección de malware ofuscado basado en machine learning" indica que los desarrolladores de malware mejoran constantemente sus técnicas para atacar sistemas de forma eficaz. Una de esas técnicas es la ofuscación de código, que dificulta la detección de malware utilizando los mecanismos tradicionales utilizados por el software antivirus actual.

En su estudio "Detección de malware con modelo de lenguaje y su clasificación mediante SVM" desarrollado por Valencia y Galicia (2016), detalla que, debido a la presencia de malware, se vuelve más difícil de analizar. Los desarrolladores se centran en métodos de prevención y detección de Ofuscación para destruir el disco duro si un análisis detecta malware. En este artículo publicamos un análisis dinámico de seis tipos de malware. Troyanos, gusanos, virus, troyanos espías, puertas traseras y rootkits también Software blanco que utiliza el modelo de lenguaje n-gram en las llamadas a Win API de cada ejemplo. Finalmente, se utilizó máquinas de vectores de soporte de kernel polinómico para la predicción de clasificación de malware. Se logró un 70% de rendimiento y clasificación de malware, El software de White tiene una calificación del 100% para determinar si es un producto malicioso.

"Meta análisis de la aplicación de aprendizaje automático en la detección de malware" es un estudio realizado por González y Daniel (2023) describe que el crecimiento de las ciberamenazas y el desarrollo continuo de tecnologías de ataque crean la necesidad de sistemas de detección más potentes y eficaces (Moya J., 2023). En este contexto, el aprendizaje profundo ha surgido como un enfoque prometedor que explota su capacidad para extraer patrones y características complejos de grandes cantidades de material. El objetivo del trabajo es evaluar si la aplicación de técnicas de aprendizaje profundo en la detección de malware da resultados positivos para su uso en este entorno. Esto se logra utilizando estas técnicas (CNN, RNN, codificadores automáticos) para la detección de malware. Los métodos utilizados incluyeron un meta análisis detallado de estos estudios siguiendo las directrices PRISMA y teniendo en cuenta algunas características de estratificación del meta análisis, como la plataforma en la que se recibió el ataque y el tipo de ataque (Labarca et al., 2020). Analizado simplemente los resultados obtenidos durante el meta análisis fueron muy heterogéneos, lo que hizo imposible asegurar la exactitud de sus conclusiones numéricas. Sin embargo, no hay duda de que un análisis crítico de todo el proceso proporciona una interpretación positiva de la aplicación de la tecnología de aprendizaje profundo en la detección de malware.

En el trabajo "Detección de malware usando herramientas de Big Data" de Valero Campaña describe el desarrollo del proyecto rb-Malware (propiedad de ENEO Tecnología) (Valero Campaña, 2015). El propósito de este proyecto es detectar archivos maliciosos que ingresan a la red local. En particular, se describirá en detalle el desarrollo de las partes relacionadas con el análisis de malware. Varios sensores de red son responsables de capturar y almacenar archivos en el rb-S3. Luego, el programa rb-sequence-oozie recupera estos archivos y los envía a un sistema de detección de malware. Este sistema de registro ejecuta lotes y exporta los resultados a Apache Kafka. Finalmente, los datos se leen y muestran mediante la interfaz web de rb-Malware. Este proyecto está basado en el proyecto BinaryPig desarrollado por Endgame en 2013. El marco de detección de BinaryPig se basa en un clúster de Hadoop para análisis escalables. Las tareas que se ejecutan durante la ejecución por lotes ejecutan herramientas que se utilizan para detectar malware. Estas herramientas de detección incluyen YARA, VirusTotal, Kaspersky, Metascan y

Volumen X. No. X
mes xxxx
ISSN: 2773-7527

ClamAV. Finalmente se realizaron pruebas para evaluar la versión 0.4 del proyecto rb-Malware

MATERIALES Y MÉTODOS

La metodología implementada para la presente investigación de minería de datos CRISP-DM según (Pete, 2000) se describe en términos de un proceso jerárquico. Modelo, que consta de conjuntos de tareas descritas en cuatro niveles de abstracción (de general a específico): fase, tarea genérica, tarea especializada e instancia de procesos.

En el nivel superior, el proceso de extracción de datos se organiza en varias fases; cada fase consta de varias tareas genéricas de segundo nivel. Este segundo nivel se llama genérico, porque pretende ser lo suficientemente general como para cubrir todas las situaciones posibles de extracción de datos. Las tareas genéricas pretenden ser lo más completas y estables posible. medios completos cubriendo tanto todo el proceso de minería de datos como todas las posibles aplicaciones de minería de datos. Estable significa que el modelo debería ser válido para desarrollos aún imprevistos como nuevas técnicas de modelado (Pete, 2000).

En la figura 1, se visualiza los procesos identificables para la metodología CRIS-MD.

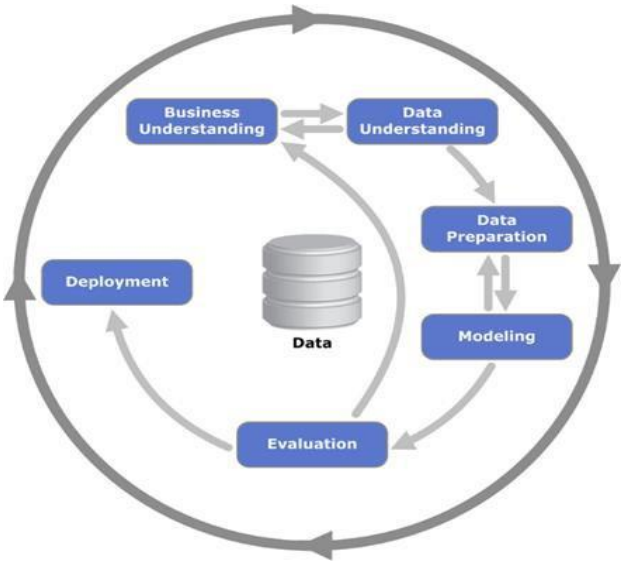

Figura 1. Metodología CRIS-MD

Comprensión de los datos.

El conjunto de los datos pertenece al Instituto Canadiense de Ciberseguridad de la universidad de New Brunswik (Portilla, 2022), quien ha desempeñado un papel fundamental de la innovación en ciberseguridad en Canadá. Este conjunto de datos usa el modo de depuración para el proceso de volcado de memoria. Esto funciona para representar un ejemplo más preciso de lo que un usuario promedio tendría en ejecución en el momento de un ataque de malware. La página da la opción de descargar, entregando un único archivo en formato csv con los datos establecidos.

El conjunto de datos utilizado para la revisión en el presente trabajo contiene 58596 registros y 57 columnas.

Las columnas utilizadas en el desarrollo de la investigación son category y class, cuaya información establece los malwares ofuscados en la data.

Preparación de los datos

Los datos carecen de valores atípicos y/o perdidos, por lo que no es necesario una limpieza de estos. Sin embargo, las categorías están muy detalladas. Estos datos se obtuvieron mediante la instrucción **"select Category as Recategory from ['MalwareOfuscado'] GROUP BY Category;"** en SQL Server, según se muestra en la figura 2.

Figura 2. Categorías

Se procede a recategorizar el conjunto de datos de la columna en familias principales de Ransomware, Spyware y troyano, y a su vez en subcategoría, como se muestra en la siguiente tabla.

Malware	Recategory
Ransomware	Maze
Ransomware	Pysa
Ransomware	Conti
Ransomware	Shade
Ransomware	Ako
Spyware	TIBS
Spyware	CWS
Spyware	Transponder
Spyware	180solutions

Volumen X. No. X
mes xxxx
ISSN: 2773-7527

Spyware	Gator
Trojan	Reconyc
Trojan	Zeus
Trojan	Emotet
Trojan	Scar
Trojan	Refroso

Tabla 1. Tipos de Malware

Durante la fase de preparación de los datos, se generaron dos nuevas variables denominadas "malware" y "recategory". Estas fueron creadas a partir de un análisis y evaluación de las variables preexistentes "category" y "class". Las variables resultantes, "malware" y "recategory", serán objeto de análisis y evaluación detallados en las fases subsiguientes del proceso, lo que permitirá un examen exhaustivo de su relevancia y contribución en el estudio o investigación desarrollada.

Modelamiento

En esta fase se implementa un análisis de los datos con la ayuda de las herramientas SQL Server y Power BI, que son consideradas como tecnologías pertenecientes a Microsoft.

SQL Server es un sistema de gestión de bases de datos relacionales (RDBMS) desarrollado por Microsoft. Es una poderosa plataforma para almacenar, recuperar, administrar y proteger grandes conjuntos de datos. SQL Server utiliza un lenguaje de consulta llamado Transact-SQL (T- SQL) para interactuar con la base de datos y realizar operaciones como consultar, insertar, actualizar y eliminar datos. El sistema ofrece características como almacenamiento avanzado, herramientas de administración, seguridad de datos, análisis, integración con otras tecnologías de Microsoft y soporte para entornos locales y en la nube (Azure). SQL Server se usa ampliamente en aplicaciones empresariales para administrar y mantener datos críticos de manera eficiente y segura.
Este potente gestor de bases de datos se utilizó en la etapa de preparación de datos para crear las nuevas columnas que permitirán ajustar el análisis exhaustivo.

Power BI es una plataforma de análisis de datos desarrollada por Microsoft que permite visualizar y compartir información a través de paneles interactivos y reportes dinámicos.

Este conjunto de herramientas ofrece capacidades para conectar, transformar y modelar datos desde diversas fuentes, incluyendo bases de datos, archivos, servicios en la nube y aplicaciones. Utilizando una interfaz intuitiva, los usuarios pueden crear visualizaciones atractivas, tableros interactivos y reportes personalizados mediante arrastrar y soltar elementos visuales

RESULTADOS Y DISCUSIÓN

En esta fase se evalúa los resultados obtenidos mediante el análisis realizado con la ayuda de la herramienta de power BI.

La figura 3 muestra la relación porcentual de los datos antes de ser preparados, donde se visualiza aquellos que son considerados benignos representan 93.75 % de los datos.

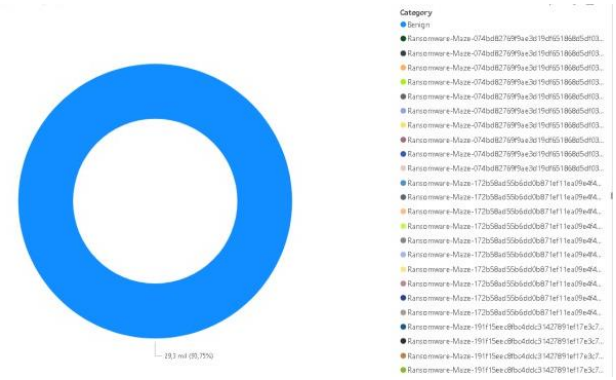

Figura 3. Relación Categorías/Benigno

La figura 4, muestra la porción de los datos entre los malwares ofuscados que se identificaron y su proporción equilibrada en el ambiente realizado.

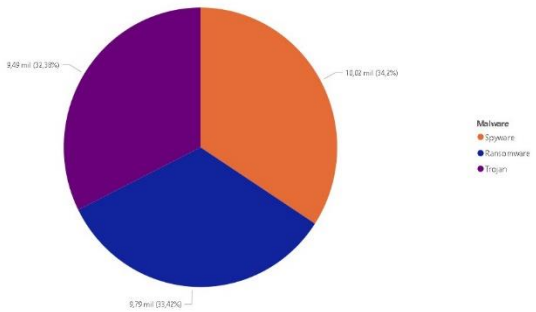

Figura 4. Relación Recategorización Nivel 1 Malware

ISSN: 2773-7527
Volumen X. No. X
mes xxxx

La figura 5 muestra un gráfico de barras de la recategorización en el nivel 2 y el malware al que pertenece.

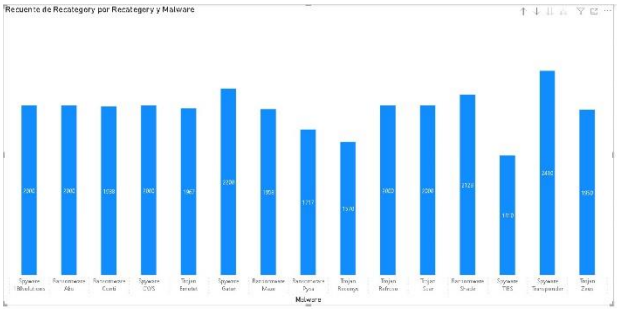

Figura 5. Relación Recategorización Nivel 2 Malware

La figura 6 muestra una equilibrada distribución de los datos con relación a las familias recategorizadas.

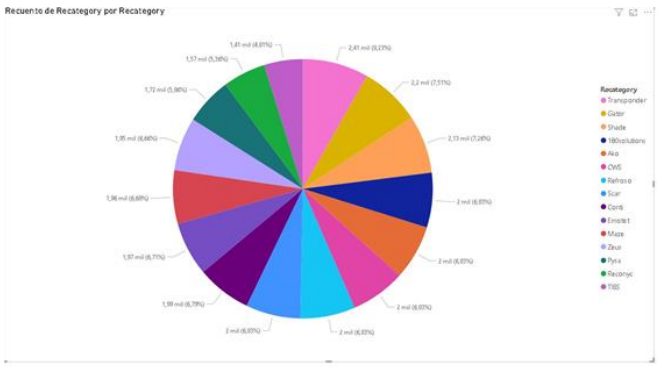

Figura 6. Relación Recategorización Malware

CONCLUSIONES

La investigación realizada analizó el conjunto de datos denominados CIC-MalMem-2022, de los malwares ofuscados en los volcados de memoria, para identificar la distribución de los diferentes softwares maliciosos.

El escenario implementado para la generación de los datos, estableció una distribución equilibrada entre los registros clasificados como benignos y malwares. Es decir, un 50% para cada clase.

Dentro de los registros maliciosos la distribución permite posicionar al malware Spyware ligeramente superior con un 34 %, lo que lo hace uno de los más numerosos en este tipo de ataques.

En la recategorización, a un nivel más inferior el malware transponder de la familia Spyware es el que más instancias abarcó, con un porcentaje del 8.23% en relación al total, mientras que la subcategoría denominada Reconyc de la familia de los troyanos fue el que menos registros de volcado de memoria generó con un 5.36%, lo que corresponde a 1570 registros.

LITERATURA CITADA

González Herrera, D. (2023). Metaanálisis de la aplicación de aprendizaje automático en la detección de malware. UNED.

Guaña-Moya, J., Sánchez-Zumba, A., Chérrez-Vintimilla, P., Chulde-Obando, L., Jaramillo-Flores, P., & Pillajo-Rea, C. (2022). Ataques informáticos más comunes en el mundo digitalizado. Revista Ibérica de Sistemas e Tecnologias de Informação, (E54), 87-100.

Labarca, G., Uribe, J. P., Majid, A., Folch, E., & Fernandez-Bussy, S. (2020). Como interpretar una revisión sistemática con comparaciones múltiples o network metaanálisis. Revista médica de Chile, 148(1), 109-117.

Llalli Chanini, S. A. N. D. R. A. (2022). Estudio de técnicas de ofuscación de código fuente para proteger información (Doctoral dissertation).

Marqués y D. Somekawa. (2012). Ofuscamiento de código para protección de programas Java contra ingeniería inversa. Repositorio institucional de UTFPR.

Moya, J. G. (2023). La importancia de la seguridad informática en la educación digital: retos y soluciones. RECIMUNDO: Revista Científica de la Investigación y el Conocimiento, 7(1), 609-616.

Moya, J. G. (2023). Revolución de la ciberseguridad en la cuarta revolución industrial. Revista Ingeniería e Innovación del Futuro (RIIF), 2(2), 6-20.

Pete Chapman, J. C. R. K. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Portilla Jaimes, J. A. (2022). Desarrollo de un modelo clasificador de malware con algoritmos de aprendizaje automático.

Valencia, A., & Galicia Haro, S. (2016). Detección de malware con modelo de lenguaje y su clasificación mediante SVM.

Valero Campaña, M. (2015). Detección de malware usando herramientas de big data.