1 Objetivos

- Aplicar los conocimientos vistos en clase sobre las técnicas de análisis estático y dinámico de malware
- Implementar un modelo de ML que utilice la secuencia de llamadas a las APIs, en base a las propuestas de artículos sobre el análisis dinámico, para la clasificación de distintas categorías de malware.

2 Preámbulo

El análisis dinámico ofrece información sobre el comportamiento de un malware y cómo interactúa con el sistema que infecta. Al registrar, observar y analizar este comportamiento es posible evadir las técnicas de ofuscamiento que hacen difícil un análisis estático, pues el malware ejecuta las funciones cuyo código intenta ocultar.

Entre la información relevante que ofrece el análisis dinámico se encuentra la secuencia de llamadas a las APIs. A diferencia de un análisis estático donde podemos obtener el conjunto de APIs que un malware utiliza, la secuencia de llamadas muestra el orden en el tiempo en el que estas APIs son ejecutadas, información que se puede utilizar para derivar nuevas características en un modelo de aprendizaje de máquina, como los n-gramas.

3 Desarrollo

Parte 1 - Análisis de un malware

La primera parte consiste en el análisis de dos ejecutables de Windows proporcionados. Se proporciona una carpeta con el nombre MALWR2.zip en CANVAS, la cual posee la contraseña *infected*

Para los usuarios de Windows se debe utilizar una VM con Linux para trabajar. Se debe descargar el archivo y descomprimirlo en la ubicación deseada. Luego se debe descomprimirlo y NO se debe manipular manualmente ningún archivo, de hacerlo se corre el riesgo de ejecutarlo e infectarse.

NOTA: se proporcionan ejemplares reales de malware, para efectos de aplicar los conocimientos académicos de análisis estático y dinámico de malware, y es responsabilidad del alumno(a) cualquier uso adicional que no sea el indicado en este laboratorio. Luego de finalizar el laboratorio se deben eliminar todos los ejemplares.

Análisis estático

1. Utilice la herramienta pefile para examinar el PE header y obtenga las DLL y las APIs que los ejecutables llaman. ¿Qué diferencias observa entre los ejemplos? ¿Existe algún indicio sospechoso en la cantidad de DLLs y las APIs llamadas?

2. Obtenga la información de las secciones del PE Header. ¿Qué significa que algunas secciones tengan como parte de su nombre "upx"? Realice el procedimiento de desempaquetado para obtener las llamadas completas de las APIs.
3. Según el paper "Towards Understanding Malware Behaviour by the Extraction of API Calls", ¿en que categoría sospechosas pueden clasificarse estos ejemplos en base a algunas de las llamadas a las APIs que realizan? Muestre una tabla con las APIs sospechosas y la categoría de malware que el paper propone.
4. Para el archivo "sample_vg655_25th.exe" obtenga el HASH en base al algoritmo SHA256.
5. Para el archivo "sample_vg655_25th.exe", ¿cuál es el propósito de la DLL ADVAPI32.dll?
6. Para el archivo "sample_vg655_25th.exe", ¿cuál es el propósito de la API CryptReleaseContext?
7. Con la información recopilada hasta el momento, indique para el archivo ""sample_vg655_25th.exe" si es sospechoso o no, y cual podría ser su propósito.

Análisis dinámico

8. Utilice la plataforma de análisis dinámico https://www.hybrid-analysis.com y cargue el archivo "sample_vg655_25th.exe". ¿Se corresponde el HASH de la plataforma con el generado? ¿Cuál es el nombre del malware encontrado? ¿En que consiste este malware?
9. Muestre las capturas de pantalla sobre los mensajes que este malware presenta a usuario. ¿Se corresponden las sospechas con el análisis realizado en el punto 7?

Parte 2 - Clasificación de malware en base a secuencias de llamadas a las APIs

Obtenga del repositorio https://github.com/khas-ccip/api_sequences_malware_datasets el archivo VirusSample.csv

A partir de este dataset se deberán implementar dos modelos de clasificación de malware. Se sugiere la lectura del artículo "New Datasets for Dinamyc Malware Classification" donde se explica cómo se obtuvo la información del dataset, y el artículo "Contextual Identification of Windows Malware through Semantic Interpretation of API Call Sequence" secciones 1 y 2 para obtener ideas de como los analistas de malware han implementado los modelos de clasificación de secuencias de APIs, así como los algoritmos más utilizados.

Los modelos deben contemplar todas las fases de machine learning: exploración de datos, pre – procesamiento, ingeniería de características, implementación y validación (70% entrenamiento y 30 pruebas), K folds para k = 10, y cálculo y explicación de las métricas de Accuracy, Precision y Recall para cada categoría de malware.

El artículo "New Datasets for Dinamyc Malware Classification" sirve como un benchmark para comparar modelos de clasificación, ¿se lograron obtener mejores métricas que las obtenidas en el artículo para la clasificación de malware?