1 Objetivos

- Aplicar los conocimientos vistos en clase sobre las técnicas de análisis estático de malware
- Implementar el algoritmo K-means para la clasificación de familias de malware

2 Preámbulo

FAMILIAS DE MALWARE

Para que un malware se considere parte de una familia, no es necesario que el código sea 100% idéntico. Las diferencias más comunes entre el malware de una familia están relacionadas a la configuración, las direcciones para el Command and Control, y las características con que evolucionan, de ello se pueden desprender nuevas subfamilias.

Por ejemplo, en el 2019 los investigadores de Mandiant identificaron 186 familias únicas de malware, entre decenas de miles de ejemplares de malware. Esto es importante porque permite implementar controles similares en base a riesgos ya conocidos.

3 Desarrollo

El laboratorio consiste en la creación de un dataset de características a partir de ejemplos de malware real proporcionados. Existen familias entre los ejemplos y se debe etiquetar a que grupo o familia pertenecen los ejemplares proporcionados.

El laboratorio será desarrollado de forma individual. Se debe entregar un enlace a un repositorio de Github con el código fuente y el modelo de clusteres, así como la explicación de las métricas de rendimiento (técnica del codo y coeficiente de silhouette).

NOTA: se proporcionan ejemplares reales de malware, para efectos de aplicar los conocimientos académicos de análisis estático 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.

Se proporciona una carpeta con el nombre MALWR.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.

Parte 1

Creación del dataset

Se debe realizar un análisis estático utilizando la herramienta pefile sobre los archivos de malware proporcionados. Con la información que se obtenga del análisis se construirá el dataset inicial. Recuerde lo aprendido sobre el PE header, las secciones, las llamadas que realiza, etc.

NO se debe utilizar Virustotal ni cualquier otro motor de antivirus para obtener información sobre los archivos.

Exploración y pre procesamiento de datos

Analice la data proporcionada y determine qué técnicas de ML utilizará para la implementación de los clusteres solicitados.

Parte 2

Implementación del modelo

Utilice el algoritmo de K means para crear los clusteres a partir de los datos preprocesados del dataset. Utilice las siguientes métricas para verificar la calidad de los clusteres obtenidos:

- Método del “codo”
- Coeficiente de silhouette

Etiquete cada observación según el clúster indicada por el algoritmo.

Conclusión

¿Para qué número de clústeres se obtiene el coeficiente de sillhouete más alto? ¿Coincide con el método del codo ¿Cuántas familias existen entre los ejemplares de malware proporcionados?