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
- Determinar la similitud entre las familias de malware encontradas

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 con la configuración, las direcciones para el C&C, y las características con que evolucionan, de ello se pueden desprender nuevas sub-familias.

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, y una respuesta a incidentes más rápida en base a las tácticas y técnicas conocidas de una familia de malware.

3 Desarrollo

El laboratorio consiste en la **creación** de un dataset de características a partir de ejemplos de malware proporcionados. Existen familias entre los ejemplos y se deben determinar que familias existen entre ellos.

Se debe entregar un enlace a un repositorio de GitHub con un notebook donde se detalle: la creación del dataset, implementación del algoritmo K-means, cálculo de número óptimo de clústeres, evaluación de clústeres, y la similitud del malware que pertenezca a cada grupo. Se debe explicar cómo se obtuvo el número óptimo de familias posibles.

NOTA: se proporcionan ejemplos 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, pues existe 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.

Exploración y pre procesamiento de datos

Analice el dataset y determine qué técnicas de pre-procesamiento utilizará para eliminar características que sean poco relevantes, duplicadas, etc.

Parte 2

Implementación del modelo

Utilice el algoritmo de K-means para crear los clústeres a partir del dataset. Utilice el método del codo, genere la gráfica del error contra K (número de clústeres) para determinar de forma empírica el número óptimo de clústeres que hay (explique su razonamiento).

Luego calcule el coeficiente de Silhouette, realice la gráfica del coeficiente contra K y utilizela para determinar de mejor manera el número de clústeres o familias de malware que hay.

Etiquete cada observación según el clúster indicada por el algoritmo. Utilice el índice de Jaccard para encontrar la similitud del malware que pertenece a cada familia, y genere un grafo para cada familia.

Conclusiones

1. ¿Para qué número de clústeres se obtiene el coeficiente de Silhouette más alto?
2. ¿Coincide el coeficiente de Silhouette con el método del codo?
3. ¿Cuántas familias existen entre los ejemplares de malware proporcionados?
4. ¿Coincide el índice de Jaccard con las familias encontradas?