1 Objetivos

- Aplicar los conocimientos sobre threat hunting, data science y dominio experto para la detección de dominios DGA.

2 Preámbulo

La búsqueda y caza de amenazas en el tráfico de red es una tarea desafiante: del tráfico total solamente un segmento muy pequeño es realmente malicioso, por ejemplo, algunas conexiones de una computadora infectada tratando de conectarse con un servidor remoto en espera de instrucciones (C2), a través de dominios DGA, entre miles de conexiones benignas.

La idea del threat hunting es aplicar el conocimiento sobre tráfico de red para depurar la cantidad de tráfico a analizar, y sobre ese tráfico restante que es sospechoso, aplicar data science y dominio experto para encontrar las conexiones realmente maliciosas.

3 Desarrollo

El laboratorio consiste en analizar un archivo con registros de red, con más del 99% de registros como benignos, y encontrar y detectar conexiones hacia dominios DGA.

Se deberá desarrollar de forma individual o en parejas, y se debe entregar un jupyter notebook con el desarrollo del laboratorio. La fecha de entrega es el sábado 30 de abril.

Parte 1 – Filtrado y preprocesamiento

Para este ejercicio se utilizará el archivo large_eve.json que se encuentra en CANVAS. Este archivo contiene miles de registros capturados a través del IDS Suricata. Así mismo necesitará instalar la herramienta Flare https://github.com/austin-taylor/flare, y la librería whois.

1. Cargue la información del archivo large_even.json en una lista, muestre la cantidad de registros total (deben ser 746, 909).
2. Debido a que estamos buscando dominios DGA, del total de registros, solamente estamos interesados en los registros DNS. Cargue únicamente aquellos registros en cuya llave se encuentra “DNS”.

- Muestre la nueva cantidad de registros filtrados.
- Muestre la información de 2 registros cualesquiera.

- Debido a que la data consiste en json anidados, utilice la característica json_normalize para normalizar la información y asignarla en un dataframe.
- Como estamos buscando dominios DGA, debemos filtrar los registros DNS para aquellos registros tipo A (son aquellos que mantienen una dirección IP asociada a un dominio), se debió bajar la cantidad a 2849 registros.
- Filtre los dominios únicos.
- Del dataframe de dominios únicos de tipo A, obtenga el TLD (top level domain) utilizando Flare como una columna nueva.

Parte 2 – Data Science

- Utilice su clasificador entrenado en el laboratorio 1, y clasifique los dominios del dataframe de la parte 1. Filtre aquellos claisificados como DGA y muéstrelos.
- Utilice el clasificador de DGA que incluye Flare, y clasifique los dominios (agregue a su dataframe la columna de clase). Filtre aquellos considerados como DGA y muéstrelos. ¿Son los mismos dominios que los de su modelo?

Parte 3 – Dominio experto

- Ahora ya tenemos un listado de dominios reducido y considerado como sospechoso, por lo que debemos aplicar dominio experto para encontrar los verdaderos registros malicioso. Utilizando la lista de Cisco Umbrella incluido en Flare, indique si los dominios encontrados por el clasificador DGA de Flare se encuentran en este top 1000000.
- Si son considerados por el clasificador como DGA, y no están en el top de Cisco Umbrella es muy probable que sean maliciosos. Para confirmar esto podemos buscar la fecha de creación del dominio. Cree una función qué en base al dominio, devuelva la fecha de creación de este (utilice la librería whois para esto).
- Muestre la fecha de creación para cada uno de los dominios clasificados como DGA. ¿Cuáles son los dominios que podemos confirmar como sospechosos?