DATA DRIVEN SECURITY

Grupo: Yvette & John

Enero 2020

PRACTICA

Para desarrollar la actividad práctica del curso nuestro grupo ha seleccionado la temática de ***Threat Intelligence***, la pregunta que buscamos responder es:

> ***Are there some types of threats where there is a larger degree of "agreement" between different RBLs?***

Para ello hemos tomado 9 diferentes Realtime Blackhole Lists (RBL) del repositorio de GitHub del proyecto ***FireHOL***, obtuvimos más de 60.000 IPs catalogadas como amenaza, a partir de allí recolectamos más información relevante para comparamos el porcentaje de coincidencia que existe entre los ipset en función de la categoría a la cual pertenecen y a su ubicación geográfica.

A continuación se resume la información sobre los ipset seleccionados para la práctica.

Table 1: Data description table

RBL	*Threat*	*IPs*	*Country*	*Source*
Malware				
hpHost	Malware	19240	103	http://iplists.firehol.org/?ipset=hphosts_emd
eSentire	Malware	7228	36	http://iplists.firehol.org/?ipset=esentire_downs1_ru
Cybercrime	Malware	2106	93	http://iplists.firehol.org/?ipset=cybercrime

RBL	Threat	IPs	Country	Source
Spam				
NiX Spam	Spam	6805	154	http://iplists.firehol.org/?ipset=nixspam
Urandom	Spam	258	12	http://iplists.firehol.org/?ipset=urandomusto_mailer
Clean-MX	Spam	4517	80	http://iplists.firehol.org/?ipset=cleanmx_phishing
Abuse				
Sblam.com	Abuse	8124	120	http://iplists.firehol.org/?ipset=sblam
MyIP.ms	Abuse	851	44	http://iplists.firehol.org/?ipset=myip
GPF Comics	Abuse	13230	166	http://iplists.firehol.org/?ipset=gpf_comics

Como primer paso para el desarrollo de la practica hemos tomado 9 diferentes Realtime Blackhole Lists (RBL) del repositorio de GitHub del proyecto ***FireHOL***, de donde obtuvimos más de 60.000 IPs catalogadas como amenaza, a partir de allí recolectamos más información relevante para determinar el porcentaje de coincidencia que existe entre los ipset en función de la categoría a la cual pertenecen y a su ubicación geográfica.

Step 1: Data ingestion

En primer lugar seleccionamos los 9 ipsets divididos en 3 categorías:

Malware - Spam - Abuse

Posteriormente realizamos la descarga de los datos, tomamos las IPs, la información más relevante de los encabezados y generamos un data frame por cada fuente.

A continuación se puede evidenciar la composición de cada uno de estos data frames, para ello seleccione una fuente y numero de observaciones.

Datasets Practica.

Seleccione el Dataset:

Malware 1 - hpHost

Numero de observaciones:

10

IP	Abrv	Country	Longitud	Latitud	Threat	Source
1.32.250.75	SG	Singapore	103.85	1.29	Malware	hpHost
1.34.10.192	TW	Taiwan, Province of China	121.30	24.99	Malware	hpHost
1.34.91.201	TW	Taiwan, Province of China	121.47	25.01	Malware	hpHost
1.53.252.174	VN	Viet Nam	107.08	10.35	Malware	hpHost
1.81.5.185	CN	China	109.75	38.29	Malware	hpHost
1.85.5.74	CN	China	108.93	34.26	Malware	hpHost
1.85.11.115	CN	China	108.93	34.26	Malware	hpHost

Step 2: Data Processing

Una vez cargada la información limpiamos los datasets de datos con valor NA o cero, determinamos la geolocalización de cada IP y generamos los diferentes data frames necesarios para determinar gráficamente si existe algún grado de coincidencia entre las IPs registradas como amenazas con el fin de cuantificar la confiabilidad de cada fuente en función de su relación con las demás.

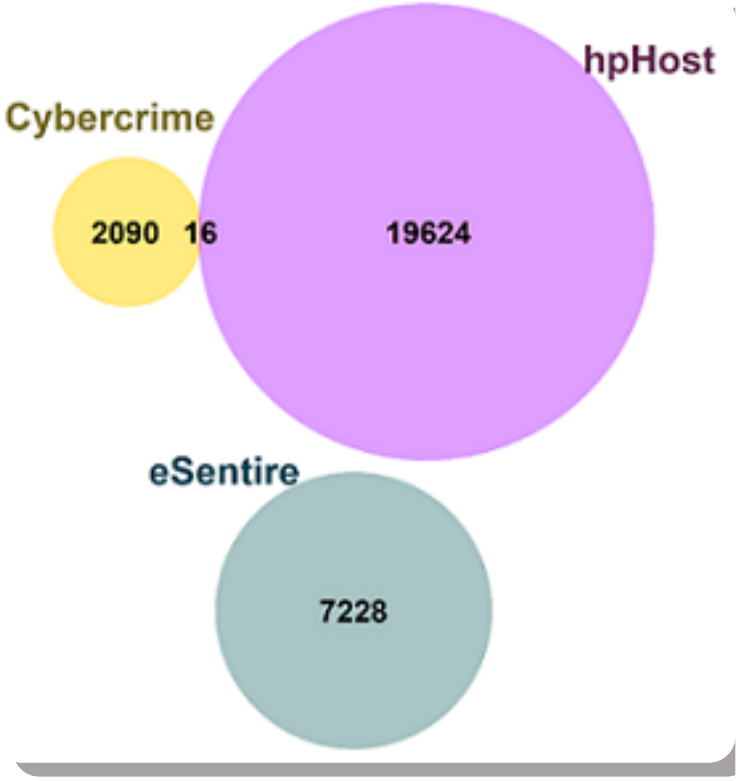
Para este análisis empleamos gráficos como Diagramas de Venn e Histogramas, además de dos aplicaciones que nos permiten interactuar con el contenido de los datasets y la localización geográfica de cada IP.

Step 3: Reporting

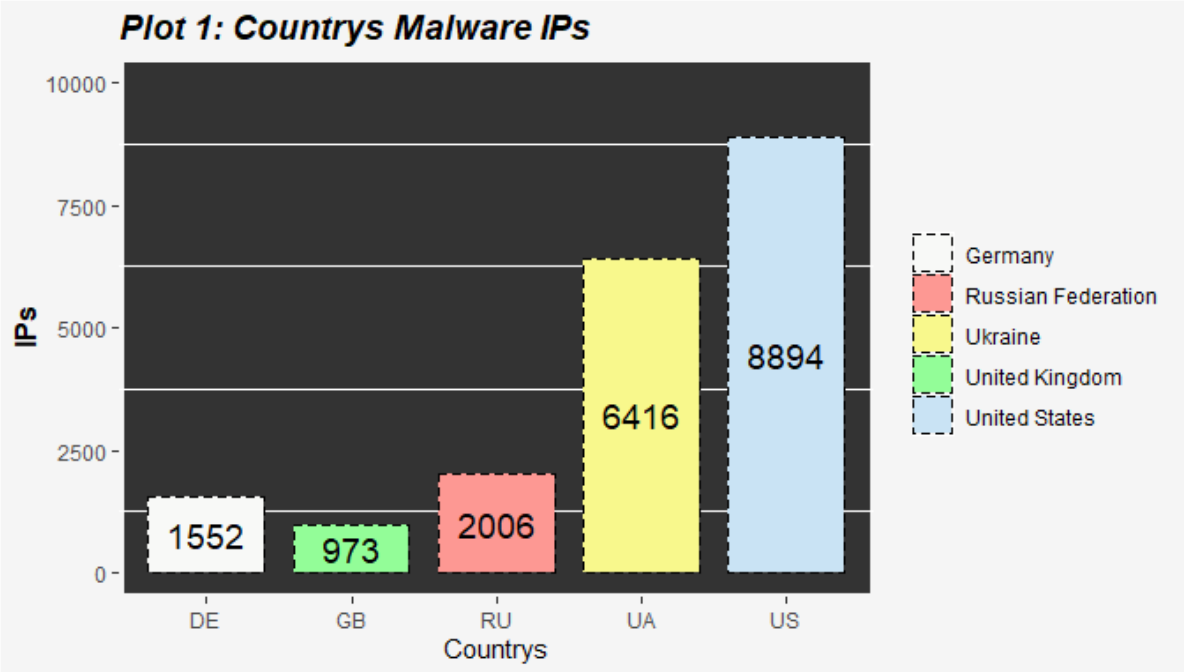
Venn Diagram category 1: MALWARE

Como se evidencia en la imagen, la relación entre las IPs publicadas como Malware por las tres fuentes seleccionadas no nos permita inferir algún grado

de confiabilidad sobre estos datos. Tras analizar los datasets creados se logró una coincidencia de menos del 1% en las IPs, lo que hace necesario recolectar más información y realizar más pruebas de correlación.

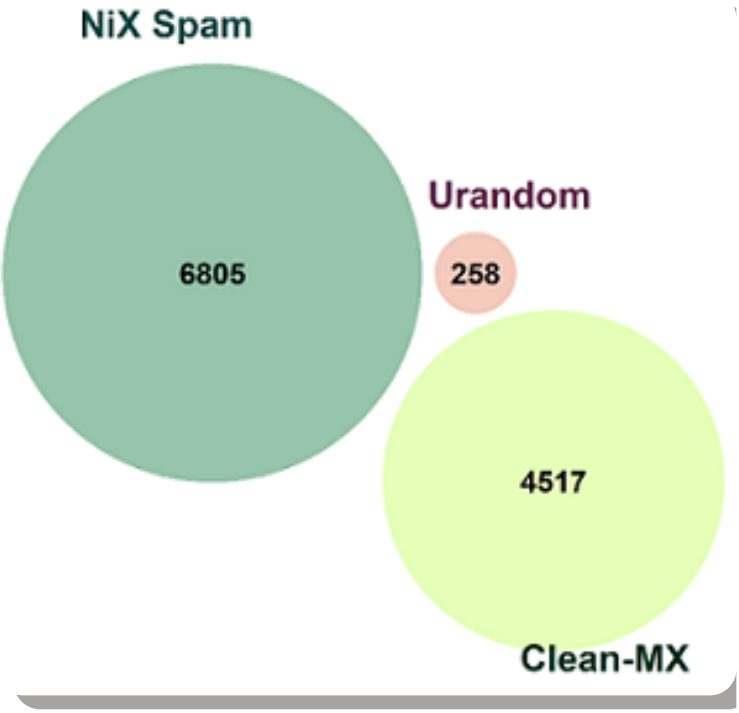


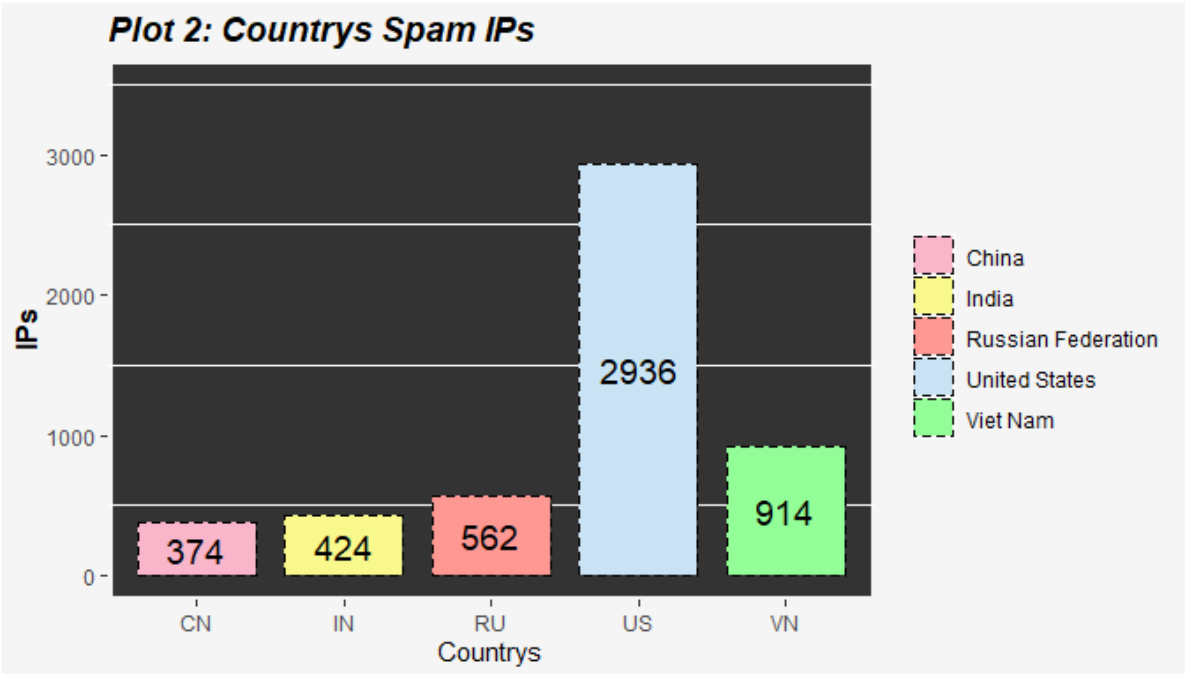
Dado que el Diagrama de venn no nos brindó información útil para determinar la confiabilidad de los datos publicados en los ipsets mantenidos por hpHost, eSential y Cybercrime, hemos tomado los datos referentes a los 5 países con el mayor número de IPs registradas como malware según estas 3 bases de datos. En la gráfica que se presenta a continuación podemos evidenciar que poco más del 30% de estas IPs catalogadas como malware provienen de Estados Unidos, además, alrededor del 20% de las IPs restantes, provienen de Ucrania.



Venn Diagram category 2: SPAM

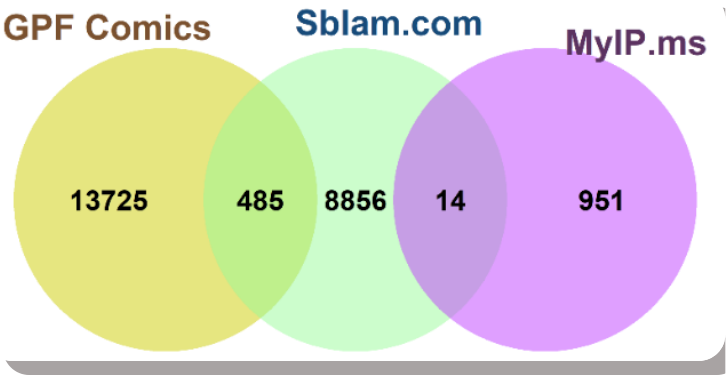
De forma similar al paso anterior, el Diagrama de Venn generado a partir de las tres fuentes seleccionadas como Spam, nos demuestra que no hay ninguna coincidencia entre las más de 11.000 IPs analizadas en esta categoría, por lo tanto, ya que este resultado no nos permite inferir la confiabilidad de los datos ni su relación. A continuación generamos una gráfica sobre los 5 países con mayor cantidad de IPs catalogadas como Spam según las bases de datos:





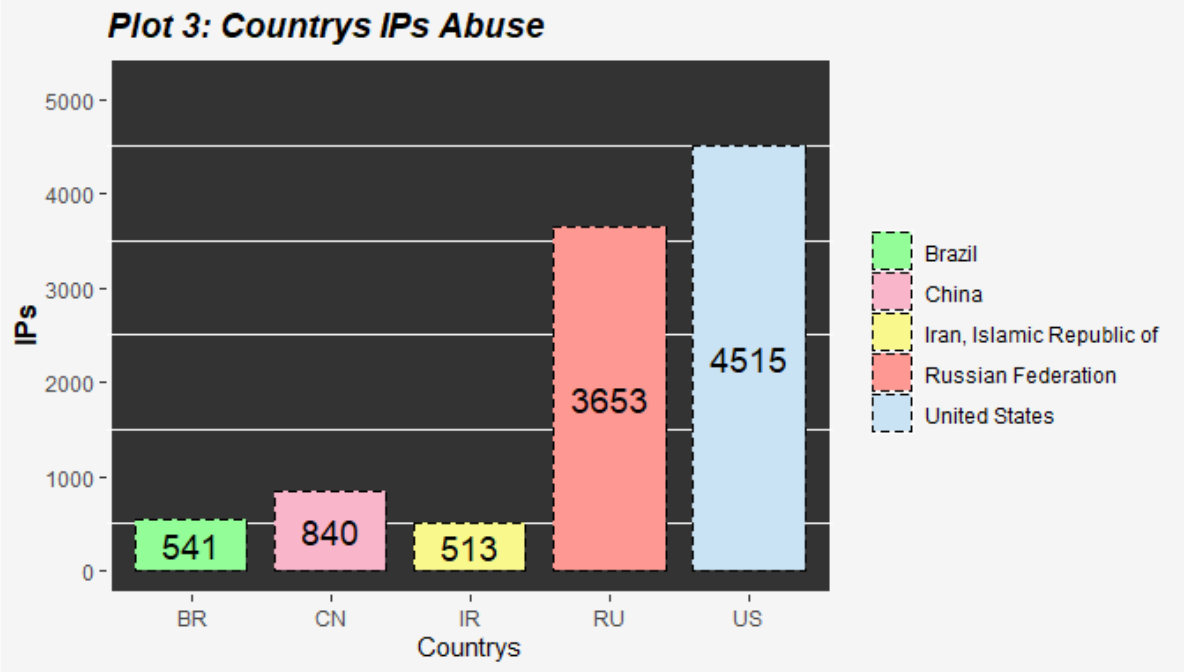
En esta segunda grafica podemos evidenciar que la tendencia de los países con mayor numero de IPs catalogadas como amenza, para este caso Spam, la lidera Estados Unidos. Los datos procesados nos indican que alrededor del 30% de las amenazas detectadas han provenido de esta región, por lo que podriamos inferir que se debe prestar especial atención al trafico que provenga de Norteamérica.

Venn Diagram category 3: ABUSE



Los ipsets seleccionados en la categoría Abuse muestran un comportamiento similar a los anteriores, la convergencia de las IPs no supera el 2% del total de los datos, en consecuencia, no es posible determinar la confiabilidad de las fuentes seleccionadas a partir de los datos recolectados. Se hace necesario recabar más información y realizar más análisis para dar solucion a la pregunta propuesta en esta actividad.

Los 3 ipsets seleccionados para la amenaza categorizada como Abuse, denotan la dominancia de las IPs maliciosas provenientes de Estados Unidos, además, de sobresalir países como Rusia, China y Brazil, entre otros países Europeos y Asiáticos.



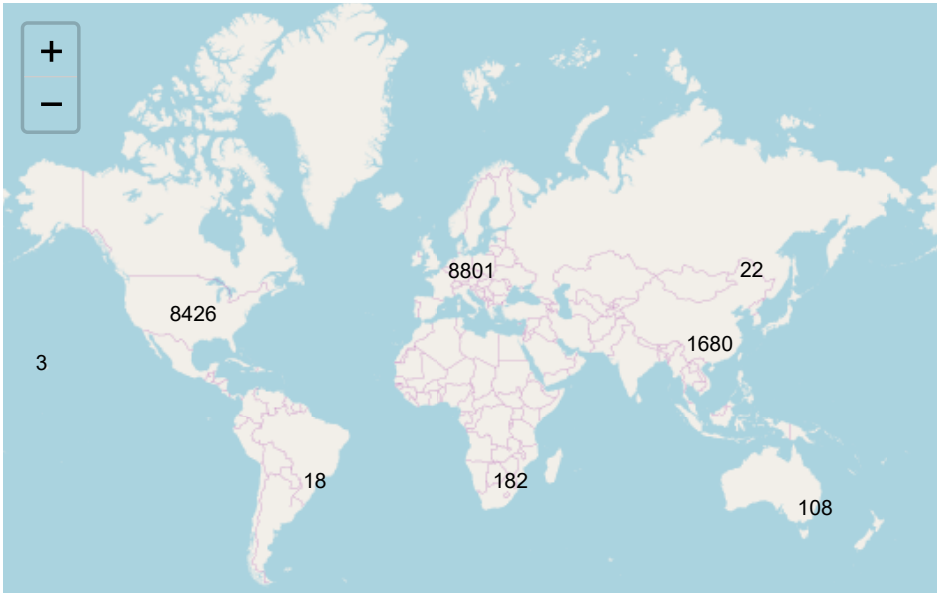
Geolocation

Por último, para contextualizar los datos analizados podemos determiar la ubicación geográfica de cada una de las IPs que han sido catalogadas como amenaza en las RBLs empleadas para esta actividad. Para ello, seleccione la fuente que desea consultar en el panel izquierdo de la aplicación que se encuentra a continuación e interactúe con el mapa para determinar la IP y su georeferenciación.

Datasets Practica.

Seleccione el Dataset:

Malware 1 - hpHost



Como finalización de este breve ejercicio se ha generado el presente reporte, los resultados arrojados tras el procesamiento y análisis grafico de los datos no nos permite inferir una solución a la pregunta planteada inicialmente, sin embargo, si nos permite establecer la relación que tienen el origen de las IPs y su comportamiento malicioso, por lo cual, se hace necesario recolectar más información al respecto para complementar y profundizar el análisis propuesto por el grupo para esta actividad practica .

PRACTICA / DATA DRIVEN SECURITY

Grupo: Yvette Ramos Oliveau & John Roldan

Enero 2020