Identificando patrones ilegales dentro de la red TOR

Jorge Antonio Morales, Juan Esteban López, y Alejandro Narváez

(Universidad Icesi, Cl. 18 #122-135, icesi.edu.co)

Resumen— Este documento tiene la intención de probar que existen patrones de nombramiento en las urls onion. Se hizo una recopilación de datos de urls y fueron procesadas en programas como *Google Colab y Sklearn*. En síntesis se logró cumplir los objetivos alcanzables en su mayoría, la cantidad de datos fue muy poca para determinar los valores debido a que la muestra no fue muy representativa.

Abstract--This document is intended to prove that there are appointment patterns in the urls onion. A collection of urls data was made and processed in programs such as sklearn and codelabs. In short, we managed to achieve the most achievable objectives, the amount of data was too little to determine the values because the sample was not very representative.

Palabras Clave: DNS onions, Dark webI

I. Introducción

La extensa y vasta red que comunica a todo el mundo en la actualidad tiene su lado profundo y desconocido en el que están inmerso variedad de páginas e información no apta para la mayoría de las personas. Muchos piensan que al navegar en una red anónima como lo es TOR, se pueden salvar de cualquier intrusión o ataque, no obstante, la realidad es otra y al parecer están más comprometidos de como lo estarían en la red superficial. Es por esto que se va a tratar de identificar patrones dentro los sitios web de la red "anónima" para poder verificar si estos influyen dentro de la ilegalidad.

Este proyecto investigativo va a estar conformado de tres partes. La primera fase consiste en realizar una lista de sitios de la web Tor que hayan sido reportados o no reportados. La segunda consiste en determinar, con base en la lista de urls, cuales de ellas se encuentran maliciosas-activas y cuales no. La última consiste en identificar y realizar un seguimiento al patrón de nombramiento de las urls según su contenido.

II. Hipótesis

Para la realización de esta investigación - experimento se parte desde la siguiente pregunta interrogante: ¿Existen patrones dentro de las urls de la darknet de Tor? y si es así ¿Cómo identificarlos?

III. Objetivos

Esta investigación tiene como propósito identificar patrones que tienen las url dentro de la red Tor, para ello se utilizará una máquina virtual que emulará el sistema operativo Kali Linux para realizar todas las pruebas pertinentes y hacer más efectivo el experimento.

ALCANZABLES

- Encontrar sitios en la darknet, abriendo ventanas html con un web scraper.
- Buscar patrones de nombramiento de urls, implementando algunos métodos de búsqueda, como:
    - Comparar los tamaños de la cadena, el número de mayúsculas y minúsculas, la suma de los valores de la url.
- Clasificar las urls con contenido legal y no legal.

NO ALCANZABLES

- Cómo realizar una extensión para el navegador que al abrir un link pueda decir la probabilidad de que dicho enlace sea malicioso
- Buscar patrones de nombramiento de urls maliciosas, implementando algunos métodos de búsqueda. (complicados)

IV. Metodología

Antes de entrar en materia es necesario conocer las diferencias que hay entre la darknet y la red superficial, dado que es de suma importancia para esta investigación diferenciar y reconocer en qué lado de la web influyen y se propagan todas aquellas urls no identificadas.

La darknet a su vez forma parte de la Deep Web y consiste en contenido público que requiere de software específico y autorización para acceder al mismo. A diferencia de la red superficial que se encarga de ejecutar programas con la

función de buscar, clasificar e indexar los contenidos web, almacenando la información en bases de datos.

Este trabajo estará configurado en tres etapas que permitirán llegar a una conclusión con respecto a los patrones maliciosos en las urls dentro de la red Tor (llevando las debidas precauciones durante el proceso), y con esto, determinar si existe alguna similitud entre las urls escaneadas de su mismo tipo.

La primera fase del proyecto consiste en conseguir una lista de urls dentro de la red Tor, después se hará una búsqueda exhaustiva en la red para buscar un máximo de 100 urls.

La segunda parte consiste en determinar de entre la lista de páginas seleccionadas cuales permanecen activas y cuáles no. Para implementar lo anterior, se utilizará la herramienta *PhantomJS* que permitirá realizar una descarga automática de las urls seleccionadas, permitiendo así determinar cuáles son los enlaces externos que tiene cada una de estas páginas, para luego almacenarlas y averiguar a donde redirigen.

El tercer y último paso consiste en identificar cuál es el patrón de seguimiento o comportamiento similar que tienen estas páginas maliciosas. Para esta fase se usará un programa desarrollado por nosotros que nos permitirá hacer un seguimiento de la longitud de las cadenas, la cantidad de mayusculas y minusculas de las urls, así mismo se realizarán comparaciones en base a sus letras consecutivas.

V. Análisis y resultados

Con respecto a la herramienta *PhantomJS[8]*, la cual hubiese permitido capturar grandes cantidades de datos de urls de la red *Tor*, no se pudo efectuar en su totalidad debido a que éste no permite abrir los enlaces .onion, por lo que a su vez también se descartó la idea de clasificar las páginas como activas e inactivas debido a que la mayoría estaban caídas.

A pesar de lo anterior, se logró hacer de forma manual lo que el *PhantomJS* debió de hacer de forma automática: se procedió a ingresar los datos uno a uno encontrados en la red *Tor*, también se creó un mecanismo semi-automático para clasificar las urls con ayuda de *vue.js* como librería de clasificación.

En total se logró reunir 129 urls .onion, estas urls fueron clasificadas como: buscador, email, foros hosting, erótico, drogas, murder, markets y hacking. Además de de que fueron categorizadas como ilegales o no ilegales dependiendo del contenido.

Surgieron varios problemas debido a la imposibilidad de utilizar el web-scraper que afectaron directamente el análisis y seguimiento de los patrones para urls maliciosas, ya que sesgó bastante la información y las estadísticas. Estas son unas de las funcionalidades que no se pudieron implementar debido a este inconveniente:

- *Obtener el jpg de la página:* Es necesario para visualizar la página sin necesidad de entrar en ella.
- *Cantidad de enlaces, enlaces javascript, enlaces a otros dominios, enlaces a la surface web:* Necesario para facilitar la obtención de más direcciones urls y clasificarlas.
- *Dominio externo onion al que se dirige más*
- *Dominio surface web al que se dirige más*

Por otra parte, se creó una interfaz para poder calcular los valores de las urls. Además del uso de *Google Colab* que es una herramienta de programación online en Google el cual tiene como propósito usar las librerías de *Sklearn* que funciona con Machine Learning para realizar predicciones con el fin de analizar los datos obtenidos de la clasificación de urls

Después de procesar los datos, se pudieron generar tablas usando los programas de *Pandas[4]*, *Matplotlib[5]* para importar los datos y hacer las gráficas que explican el comportamiento de las urls según las especificaciones dadas:

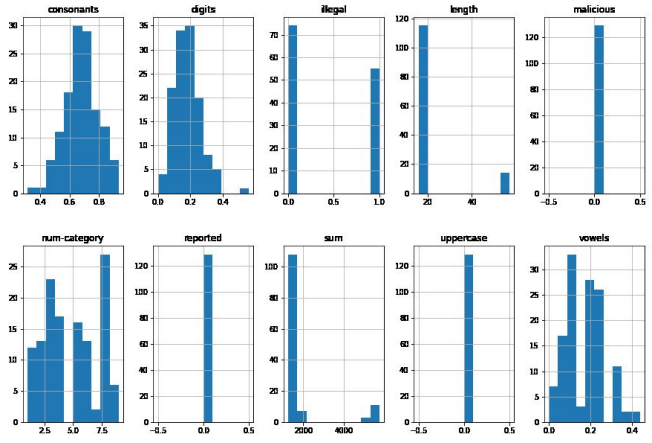

Figura 1

La figura 1 representa las gráficas de cada una de las categorías encontradas, donde se puede observar su distribución y comportamiento.

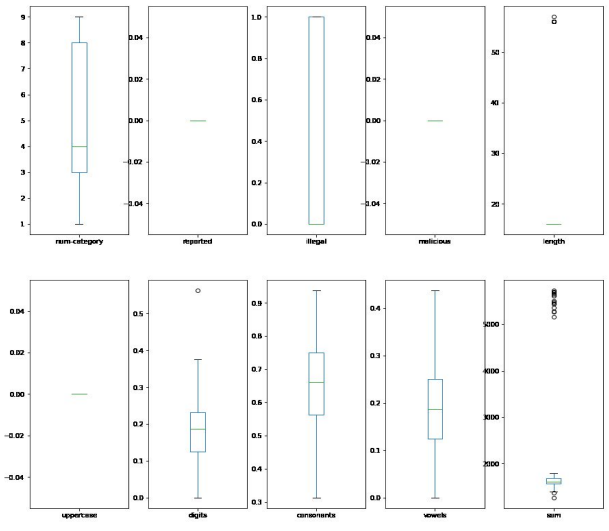

Figura 2

Teniendo en cuenta la figura 2, los diagramas de cajas y bigotes los datos atípicos en las gráficas de longitud y suma resultaron ser de utilidad debido a que en otros análisis se observó que la mayoría pertenecían a una de las categorías que se buscaba (ilegal).

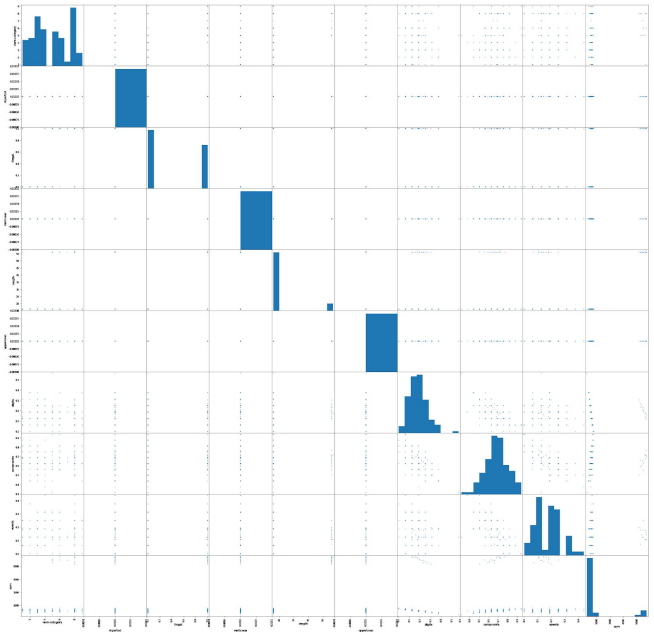

Figura 3

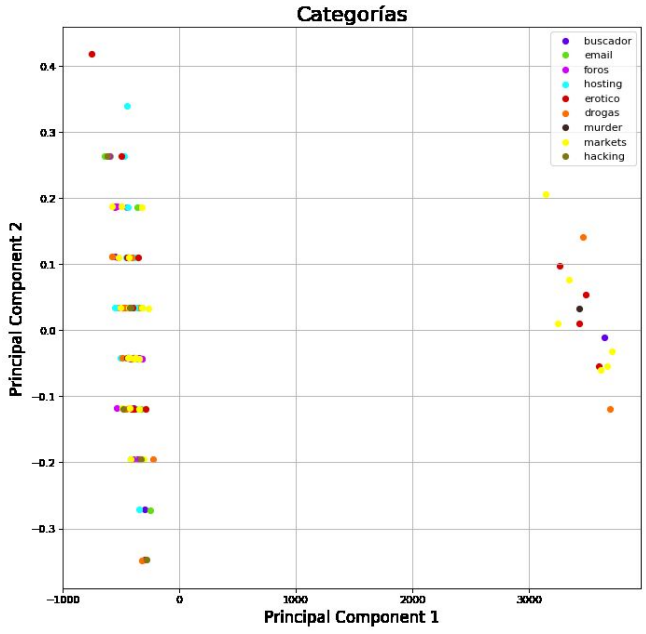

Figura 4

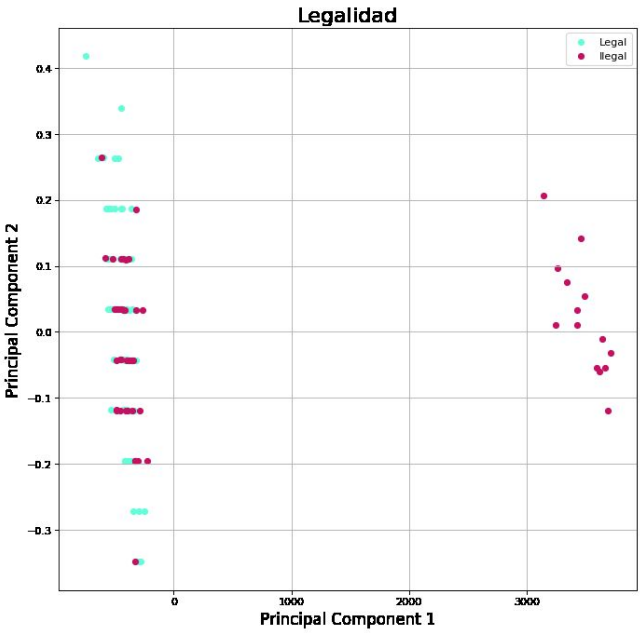

Figura 5

Con respecto a la figura 3, hubo un comportamiento entre suma de caracteres y la ilegalidad, se observó que es mucho más probable de que si una url tiene una suma de caracteres alta sea ilegal.

Se seleccionaron los datos a analizar y se transformaron en dos componentes (dígitos, consonantes, vocales y suma de caracteres para formar Principal Component 1 y Principal Component 2) y así poder mostrarlos en una gráfica de dispersión (figuras 4 y 5)

Para las figuras 4 y 5 o gráficas de dispersión se hicieron los cálculos con las características de suma, proporción de dígitos, proporción de vocales, y proporción de consonantes.

Se descartaron las que no aportaron mucho: no habían urls con letras en mayúscula, la longitud de la cadena tenía casi el mismo comportamiento que su suma, y la concatenación no sirvió debido a que no se pudo calcular con ella por ser un número demasiado grande.

Estas dos gráficas son las más representativas puesto que determinan las categorías y su posición de ilegalidad o legalidad, siendo unas categorías más ilegales que otras.

Resultados de Sklearn:

Algoritmos obtenidos de la librería de Sklearn:

- Logistic Regression (LR)
- K-Nearest Neighbors (KNN).
- Classification and Regression Trees (CART).
- Gaussian Naive Bayes (NB).
- Support Vector Machines (SVM).

Estos son los resultados con los datos que se usaron dentro del proyecto teniendo en cuenta los algoritmos:

- LR: 0.214545 (0.153429)
- KNN: 0.096364 (0.060603)
- CART: 0.066364 (0.072163)
- NB: 0.116364 (0.097047)
- SVM: 0.232727 (0.146033)

El primer valor presentado por sklearn es la media de la precisión del algoritmo de entrenamiento, mientras que el segundo es la desviación estándar.

VI. Conclusiones

Según los resultados con la herramienta de *Sklearn* se pudo obtener como conclusión que las características que se utilizaron no influyeron tanto en el resultado, además de que los datos recolectados fueron demasiado pocos por lo que la muestra no es representativa. A demás de esto, no se contaba con que las urls en onion parecen tener tamaños establecidos.

VII. Referencias:

[1] Iskander Sanchez-Rola, Davide Balzarotti, Igor Santos, "The onions Have eyes: A comprehensive structure and privacy Analysis of the Tor Hidden Services" Ph.D. Univ. Deusto, 2017.

[2] What is the Darknet?, tecnopedia, [En línea]. Disponible de: https://www.techopedia.com/definition/2395/darknet [Accedido: 7-mayo-2019]

[3] Características de la web superficial y la web profunda, Djatopagina, [En línea]. Disponible en: http://djato24.blogspot.com/2016/03/caracteristicas-de-la-web-superficial-y.html [Accedido 7-mayo-2019].

[4]pandas. https://pandas.pydata.org/

[5]matplotlib. https://matplotlib.org/

[6]scikit-learn: https://scikit-learn.org/stable/

[7]vue.js: https://vuejs.org/

[8] PhantomJS. http://phantomjs.org/.

[9] Tor Project: Anonymity Online. https://www.torproject.org

[10] google codelab :https://codelabs.developers.google.com/