Teoría de Grafos para la Identificación de Nodos Maliciosos en la Red

Tatiana S. Parlanti[1], Carlos A. Catania[1], and Luis G. Moyano[2]

[1] Universidad Nacional de Cuyo, Facultad de Ingenieria, Laboratorio de Sistemas Inteligentes (LABSIN), Mendoza, Argentina
[2] División de Física Estadística e Interdisciplinaria, Centro Atómico Bariloche, Bariloche, Argentina
{tatiana.parlanti,harpo}@ingenieria.uncuyo.edu.ar

Resumen Se explora el reconocimiento de las botnets en una red a partir de su representación como grafo, extrayendo características a sus nodos y poniendo a prueba algoritmos de agrupamiento. Se logra la separación del 88 % de las botnets junto al ∼0,14 % de los nodos benignos.

Keywords: Redes complejas, seguridad de redes, aprendizaje máquinas

1. Introducción

Una estrategia posible para la detección de nodos maliciosos en la red consiste en la construcción de modelos estadísticos utilizando información obtenida mediante técnicas aplicadas en las áreas de las redes complejas. Se trata de algoritmos tomados de la teoría de grafos para extraer características de las redes para luego ser utilizadas como información de entrada en modelos de aprendizaje estadístico. Estas características permiten capturar la estructura topológica del grafo y permite exponer aspectos adicionales de los nodos maliciosos [2,3].

En este trabajo se explora la viabilidad de reconocer el comportamiento de las botnets en una red de computadoras a partir de su representación como grafo, extrayendo para tal fin distintas características que dan información de sus nodos en cuanto al grado y centralidad de cada uno. Para ello, se pondrán a prueba diferentes algoritmos de *clustering* o agrupamiento, con el objetivo de diferenciar los nodos, no las conexiones, que estén infectados de los que no. A diferencia de otros trabajos anteriores que han considerado el número de paquetes transferidos entre los distintos nodos, en este trabajo se analizan el número de bytes transferidos, ya que se considera que contar con el dato del tamaño de los paquetes puede ofrecer mayor información para el reconocimiento de los canales de comando y control (C&C) de una botnet.

El trabajo se encuentra enmarcado en el proyecto de tesis doctoral de la Lic. Parlanti, realizado en la Universidad Nacional de Centro de la Pcia. de Bs.As. bajo la dirección de los Dres. Moyano y Catania, financiado por CONICET.

2. Materiales y Métodos

Se utilizó la base de datos CTU-13, el cual es un conjunto de 13 capturas con datos de tráfico de botnets que fue capturado en la Universidad CTU, República

Checa, en 2011 [4]. Particularmente se filtraron aquellas observaciones cuyo protocolo fuera del tipo TCP o UDP, ya que se consideró que los otros protocolos observados no ofrecían información, y se almacenó la dirección IP del nodo de origen y destino, así como la cantidad de bytes transmitidos entre ambos nodos.

Con esta información se construyó un grafo ponderado y dirigido por cada captura, donde los pesos están dados por la suma de bytes de las distintas conexiones que comparten los mismos nodos de origen y destino, diferenciando entre cantidad de bytes transmitidos del primero al segundo (`SrcBytes`), y su correspondiente conexión inversa (`DstBytes`), eliminando previamente aquellas conexiones que no hubieran logrado transmitir byte alguno. Luego, se extrajeron características de cada nodo, descriptas en la Tabla 1, cuyas definiciones comprendidas en la Teoría de Grafos se puede consultar en [5].

Tabla 1. Descripción de las características extraídas.

	Característica	**Descripción**
ID	Grado Entrante	Número de aristas que entran a un vértice
OD	Grado Saliente	Número de aristas que salen de un vértice
IDW	Grado Entrante Ponderado	Suma de los pesos de las aristas que entran a un vértice
ODW	Grado Saliente Ponderado	Suma de los pesos de las aristas que salen de un vértice
BC	Centralidad de Intermediación	Número de caminos más cortos que pasan por un vértice y minimizan la suma de los pesos de las aristas
LCC	Coeficiente de Agrupamiento Local	Cuantifica qué tan interconectado está un vértice con sus vecinos, a partir de las conexiones de sus vecinos entre sí. Se consideró el grafo como no dirigido, y no se tuvieron en cuenta los pesos de las aristas
AC	Centralidad Alfa	Variante de la centralidad de autovector, donde el vértice está sujeto a distinta importancia dependiendo de factores externos. Se tomaron los valores $\alpha = 0{,}01,\ e = 1$

Finalmente, cada uno de los nodos fue etiquetado a partir de la información provista por el equipo de StratosphereIPS [1], bajo el supuesto de que todo nodo que no tuviera la identificación de "botnet" sería considerado "normal".

Una vez calculadas cada una de las características antes mencionadas, se procedió a utilizarlas como conjunto de entrenamiento de los distintos modelos de agrupamiento, a excepción de las características de la novena captura, la cual fue conservada para testear en una etapa posterior a la abarcada en este trabajo.

Entre los algoritmos de *clustering* conocidos [7], se compararon los siguientes: **k−means** usando tanto el algoritmo Hartigan-Wong como el Lloyd-Forgy, tomando por k los cuadrados de los números naturales entre 2 y 15, ambos inclusive; así como **CLARA** usando la distancia euclideana y la Manhattan, con los mismos valores de k que para $k-$means, tomando 300 muestras sobre las que se aplica el algoritmo PAM.

Para llevar a cabo los experimentos aquí planteados, así como la extracción de características, se utilizó un procesador AMD Ryzen 7 1700 Eight-Core de 64 GB de memoria, con sistema operativo Ubuntu 20.04.4 LTS. Se trabajó con las versiones `3.8.10` de `Python` y `4.2.1` de `R`. El análisis de los distintos grafos para la extracción de características se llevó a cabo usando el paquete `igraph` (versión

1.3.1) de R, así como la biblioteca homónima (versión 0.9.9) de Python. Así se calcularon LCC y AC en el primero, mientras que el resto en el segundo.

3. Resultados y Discusión

3.1. Sobre las Características

En primer lugar se realizó una comparación sobre la información que brindan los paquetes y bytes transmitidos entre nodos. Si bien pareciera existir cierta correlación entre ambas, como se observa en la Figura 1, al hacer foco en una muestra particular de los datos, en este caso aquellas observaciones con 16 paquetes o menos y a lo sumo 1000 bytes, se aprecia que para una misma cantidad de paquetes la cantidad de bytes puede ser diferente (gráfica derecha), lo que representa un dato de interés. Es por ello que, a diferencia de trabajos como [2,3], en lugar de tomar la cantidad de paquetes como peso, se decidió utilizar la cantidad de bytes.

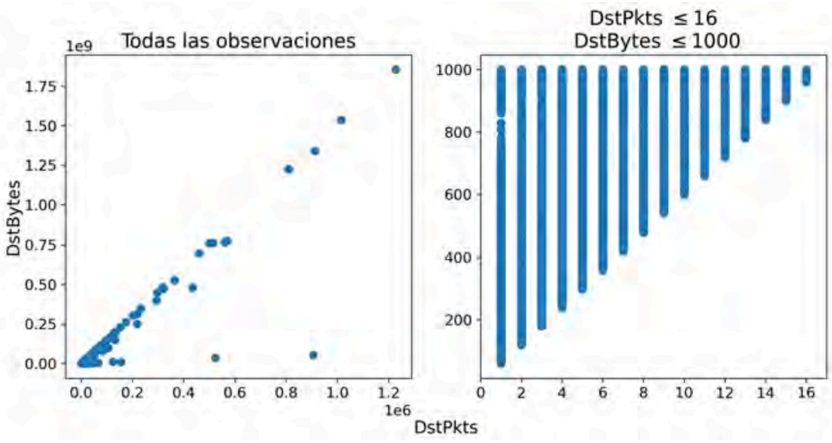

Figura 1. La gráfica de la izquierda compara la relación entre paquetes y bytes transmitidos desde nodos destino para cada una de las observaciones; la de la derecha se concentran en los casos en que se transmitieron a lo sumo 16 paquetes y 1000 bytes como máximo. Las conexiones desde nodo origen presentan un comportamiento similar.

Respecto al tiempo que demandó la extracción de características de cada grafo, en la Tabla 2 se observa el promedio de éste así como la desviación estándar en segundos. Se destaca que BC es la que requiere mayor tiempo, tomando unas ~ 21 horas promedio para su cálculo. Así también, es la que tiene mayor variación, junto con AC, en relación a la cantidad de vértices de cada grafo.

3.2. Agrupamiento

Como se explicó en la Sección 2, una vez calculadas las características de interés sobre un total de 2 874 213 nodos, de los cuales 25 están infectados, se entrenaron distintos modelos de agrupamiento, con el objetivo de diferenciar entre nodos malignos y benignos. Dada la desproporción entre ambos, es de esperar que durante el agrupamiento se encuentre un cluster principal, que contenga a la

Tabla 2. Tiempo promedio en capturas (seg.) de la extracción de características.

Característica	Tiempo Promedio	Desviación Estándar
BC	77355,149458	84083,199324
AC	507,160385	908,155491
LCC	0,087248	0,068996
ODW	0,036261	0,026509
IDW	0,036241	0,026511
OD	0,003525	0,002684
ID	0,003377	0,00256

Tabla 3. Agrupamiento usando $k-$means y CLARA. Se especifican: (i) porcentaje de nodos no infectados que quedaron fuera del cluster principal; (ii) porcentaje de nodos infectados que quedaron fuera del cluster principal; (iii) tiempo empleado (seg.).

	k-means						CLARA					
k	Alg. Hartigan-Wong			Alg. Lloyd-Forgy			Dist. Euclidea			Dist. Manhattan		
	(i)	(ii)	(iii)	(i)	(ii)	(iii)	(i)	(ii)	(iii)	(i)	(ii)	(iii)
4	0,0005	0	28,057	0,0005	0	30,651	0,0206	36	49,593	0,0203	36	29,625
9	0,003	8	48,293	0,0034	8	64,265	**0,1243**	**84**	**107,559**	**0,1376**	**88**	**68,215**
16	0,0289	36	66,089	0,0379	68	103,720	0,2964	88	185,435	0,2599	88	121,003
25	0,0875	80	95,352	0,0872	80	208,62	0,3179	88	277,426	0,3439	88	179,842
36	**0,1861**	**84**	**171,429**	**0,1876**	**84**	**480,524**	0,4437	88	381,026	0,4164	88	257,241
49	0,1991	84	326,304	0,3693	88	811,729	0,6354	88	498,378	0,5725	88	345,199
64	0,4067	88	425,199	0,4903	88	1846,021	0,9636	88	646,739	0,9193	88	448,270
81	0,5692	88	515,123	0,7593	88	2724,904	1,134	88	823,091	1,1839	92	579,934
100	0,9523	88	572,051	0,8478	92	4589,811	1,1173	92	1023,322	1,1234	92	716,886
121	0,8453	92	569,522	1,3271	92	5973,175	1,2073	92	1331,300	1,2207	92	898,435
144	1,2516	100	828,449	1,3512	92	11048,307	2,3659	100	1735,531	1,4612	92	1204,218
169	2,1877	100	791,413	1,9773	100	13722,263	1,3678	100	2159,850	1,2173	92	1552,681
196	79,4276	100	594,627	2,3576	100	17771,940	1,9427	92	2755,081	9,7819	100	2048,137
225	79,2468	100	700,582	2,3729	100	22760,47	2,632	92	3475,695	11,2066	100	2652,382

mayoría de los nodos benignos, por tener características similares. En el Tabla 3 se observan los resultados obtenidos usando $k-$means y CLARA, para diferentes valores de k, algoritmos y distancia, especificando las siguientes métricas utilizadas para su evaluación:

- **i)** Porcentaje de nodos no infectados que quedaron fuera del cluster principal.
- **ii)** Porcentaje de nodos infectados que quedaron fuera del cluster principal.
- **iii)** Tiempo empleado (segundos).

Tanto para $k-$means como para CLARA, en la mayoría de los casos se logró un cluster principal que, a juzgar por (i), contiene la mayor cantidad de nodos benignos. Sin embargo, dicho cluster no es completamente homogéneo, ya que por (ii) se observa que éste también contiene botnets. Entonces, se busca maximizar el cluster principal, minimizando la cantidad de nodos infectados que pueda contener. Teniendo en cuenta esto, además del tiempo empleado, se concluye que utilizar el algoritmo CLARA con la distancia Manhattan es la mejor opción,

ya que con $k = 9$ se logra un cluster principal que sólamente excluye el 0,1376 % de los nodos no infectados y el 88 % de los infectados, es decir un total de 3956 nodos no infectados quedan excluidos, pero incluye únicamente a 3 nodos infectados, y sólo demanda $\sim 1{,}14$ minutos. Por otro lado, CLARA usando la distancia euclideana y el mismo valor de k, deja menos nodos no infectados por fuera del cluster principal (3574), pero incluye un nodo infectado más y toma ~ 2 minutos. En contraposición, para obtener resultados similares implementando $k-$means, ya sea con el algoritmo de Hartigan-Wong o el de Lloyd-Forgy, son necesarios 36 clusters. Sin embargo, en ambos casos el cluster principal contiene 4 nodos infectados (el 84 % restante queda fuera) y toman ~ 3 y 8 minutos, respectivamente. Finalmente, vale aclarar que si bien hay casos donde el 100 % de los nodos infectados quedan por fuera del cluster principal, no fueron tenidos en cuenta ya que de igual manera aumenta el número de nodos benignos que no pertenecen a dicho cluster, siendo que además toman más tiempo.

4. Conclusiones y Trabajo Futuro

De los resultados preliminares mediante técnicas de cluster se desprende que la aplicación de teoría de grafos para la extracción de características en redes con millones de nodos facilita la discriminación de comportamiento. La utilización del número de bytes transferido entre los nodos demostró ser adecuada. En general todos los algoritmos de clustering analizados fueron capaces de agrupar a la mayoría de los nodos benignos excluyendo a los nodos con comportamiento malicioso. Sin embargo, aquellas características que se focalizan en medir la centralidad de un nodo demandan un tiempo considerable, lo que dificulta su aplicación en escenarios de tiempo real. Es por esto último que como trabajo futuro se propone analizar otras técnicas para la extracción de características de los grafos como las redes convolucionales para grafos [5,6].

Referencias

1. Stratosphere research laboratory. https://www.stratosphereips.org/, última vez visitado en Agosto 2022
2. Abou Daya, A., Salahuddin, M.A., Limam, N., Boutaba, R.: Botchase: Graph-based bot detection using machine learning. IEEE Transactions on Network and Service Management 17(1), 15–29 (2020)
3. Chowdhury, S., Khanzadeh, M., Akula, R., Zhang, F., Zhang, S., Medal, H., Marufuzzaman, M., Bian, L.: Botnet detection using graph-based feature clustering. Journal of Big Data 4(1), 1–23 (2017)
4. Garcia, S., Grill, M., Stiborek, J., Zunino, A.: An empirical comparison of botnet detection methods. computers & security 45, 100–123 (2014)
5. Hamilton, W.L.: Graph representation learning. Synthesis Lectures on Artifical Intelligence and Machine Learning 14(3), 1–159 (2020)
6. Welling, M., Kipf, T.N.: Semi-supervised classification with graph convolutional networks. In: J. International Conference on Learning Representations (ICLR 2017) (2016)
7. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Transactions on neural networks 16(3), 645–678 (2005)