Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 108C (2017) 2388–2392

Procedia Computer Science

International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland

Feasibility Study of Social Network Analysis on Loosely Structured Communication Networks

Jan William Johnsen and Katrin Franke

Norwegian University of Science and Technology (NTNU), Norway
jan.w.johnsen@ieee.org and kyfranke@ieee.org

Abstract

Organised criminal groups are moving more of their activities from traditionally physical crime into the cyber domain; where they form online communities that are used as marketplaces for illegal materials, products and services. The trading of illicit goods drives an underground economy by providing services that facilitate almost any type of cyber crime. The challenge for law enforcement agencies is to know which individuals to focus their efforts on, in order to effectively disrupting the services provided by cyber criminals. This paper present our study to assess graph-based centrality measures' performance for identifying important individuals within a criminal network. These measures has previously been used on small and structured general social networks. In this study, we are testing the measures on a new dataset that is larger, loosely structured and resembles a network within cyber criminal forums. Our result shows that well established measures have weaknesses when applied to this challenging dataset.

© 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the International Conference on Computational Science

Keywords: Digital forensics; Social network analysis; Centrality measures; Criminal networks

1 Introduction

Law enforcement agencies report that cyber crime activity is growing and become more aggressive and technically proficient [3, 7] – although the majority of cyber criminals in online marketplaces have relatively low technical skills and capabilities. This suggests that a minority of cyber criminals use marketplaces to sell easy access to sophisticated tools and expetise, through a business model called Crime-as-a-Service (CaaS) [3]. Which allow lesser skilled cyber criminals to have more impact and success in their cyber attacks. A focus on identifying and disrupting criminals in the smaller and more technical skilled group will have a larger impact on stopping illegal activities in underground marketplaces. Because their skills and expertise are difficult to replace by the larger group, with lower technical skills.

Social Network Analysis (SNA) methods has been proposed [9] for the application of identifying central individuals within criminal networks. More specifically, centrality measures are used to determine central individuals by analysing their position in a network [8], represented

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the International Conference on Computational Science
10.1016/j.procs.2017.05.172

by a graph as defined in Section 2. In previous research, centrality measures has been used to analyse relational structures in organisations [4, 5, 2] and terrorist groups [6]. The network size in these studies are between 30 and 150 individuals. Centrality measures have shown promising results to find central individuals in small and organised networks – although the networks has been incomplete or is just a sample from the total population.

However, real world datasets are neither small nor organised, and they often requires data preprocessing before they can be analysed. Although centrality measures has performed good on networks of smaller sizes by finding interesting individuals, this does not mean they will also perform good on larger and more loosely structured [1] networks. This paper is guided by the research question: *How can graph-based methods be applied to identify important individuals within a real-world online communication network?* Our research question seeks to determine the feasibility of centrality measures in applying it to the area of civil and criminal investigations.

2 Methodology

We extracted information to represent the communication within Nulled.IO as graphs: users and the messages between them, represented as vertices and edges respectively. It has not been pre-filtered and is used in its original form (detailed in Section 3) except for separating public and private messages; which results in two graphs with public communication between 26.11.2012 - 06.05.2016 and private communication between 14.01.2015 - 06.05.2016.

The reason for this division is twofold: (i) communication patterns is likely to be different between them, and (ii) civil investigators only have access to public communication in their investigation, whereas criminal investigators will have access to both.

The four centrality measures under evaluation are: *degree*, *betweenness*, *closeness* and *eigenvector*. They differ in the interpretation of *important*, thus different individuals will be ranked as more important in the same network; illustrated in Figures 1 - 4.

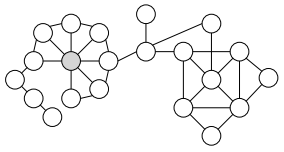

Figure 1: Largest degree centrality

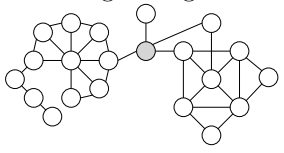

Figure 2: Largest betweenness centrality

A (undirected) graph $G = (V, E)$, where V is the set of vertices and E is the set of edges, is represented in terms of the binary adjacency matrix A. Degree centrality is the most basic measure as it only counts directly adjacent vertices. For a vertex $v \in V$, it is defined by $C_D(v) = \sum_{u=1}^{n} A_{v,u}$, where $n = |V|$. The centrality measures discussed in this paper do not consider the diagonal elements in A [8], where $v = u$, because the relationship to oneself is not important.

Betweenness centrality looks at how often a vertex sits in the *geodesic* (shortest path) between two other vertices. A vertex is considered more important because it can act like a *broker* – i.e. arrange or negotiate plans and deals – and have more influence on the network by choosing to withhold or distort information [8]. Figure 2 highlights the vertex in the network with the highest betweenness centrality score, because it sit in between two large subgraphs and one vertex. Betweenness centrality for a vertex v is defined by $C_B(v) = \sum \frac{\partial_{u,v,w}}{\partial_{u,w}}$, where $\partial_{u,w}$ is the total number of shortest paths between vertex u and w, and $\partial_{u,v,w}$ is the number of those paths that pass through v, and $u \neq v \neq w$.

Closeness centrality looks at the distance between one vertex and all the other vertices. A vertex is considered more important if it has a short distance to other vertices. In other words: the sum of distances to other vertices is low. Figure 3 highlights the vertex in the network with the best closeness centrality score, because it has the shortest distance to all the other vertices. Closeness centrality for a vertex v is defined in $C_C(v) = [\sum_{u=1}^{n} d(v,u)]^{-1}$, where $d(v,u)$ is the distance (length of the shortest path) connecting v to u.

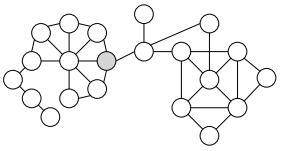

Figure 3: Largest closeness centrality

Eigenvector centrality expand on the idea of degree centrality, as it considers the edges to adjacent vertices. A vertex's score is not dependent on how many vertices it is connected to, but on many its adjacent vertices are connected to. This means that a vertex is important only if its neighbours are important – if they also have a higher degree centrality in the network. Figure 4 highlights the vertex in the network with the highest eigenvector centrality score. Eigenvector centrality for a vertex v is defined in $C_E(v) = \frac{1}{\lambda} \sum_{u=1}^{n} A_{u,v} C_E(v_u)$, where $\lambda \neq 0$ is some constant. The eigenvector value of vertex v is weighted by the sum of degree centralities of adjacent vertices.

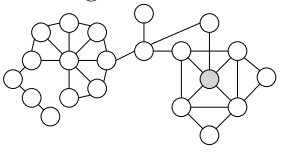

Figure 4: Largest eigenvector centrality

3 Case Study Design

The database dump[1] used in our analysis is from an online forum (accessible from the clearnet) for distributing cracked software and trading stolen credentials. It is a 9.45 GB file, which was leaked 12.05.2016, with details about 599 085 user accounts, including 800 593 private and 3 495 596 public messages. It was used as a substitute for less available darknet forums datasets, because forum users in both dark- and clearnet rely on electronic messaging to communicate, plan and organise. Although similarities between dark- and clearnet forums has not been shown in previous research, it is not unlikely to expect they are formed by similar social forces.

Table 1: Database tables and fields of interest

Table	**Fields**
topics	tid, posts, starter_id, starter_name, forum_id
posts	pid, author_id, author_name, topic_id, new_topic
message_topics	mt_id, mt_starter_id, mt_to_count, mt_to_member_id, mt_replies

Table 1 show a full list of database (DB) tables and fields used to extract the needed information for constructing the graphs. The resulting two graphs was then exported in a Graph Exchange XML Forumat (GEXF), to ease later analyses. Two DB tables was combined to construct the public communication graph. DB table *topics* contains information on the author of forum threads, so field *starter_id* is treated as source vertex. For each forum thread (topic), field *topic_id* was used to retrieve all messages posted on that topic ID from DB table *posts*. Field *author_id* was treated as the target vertex, and for each message found the edge weight between two vertices was incremented.

DB table *message_topics* hold the metadata for private communication, where field *mt_starter_id* is used as source vertex, *mt_to_member_id* as target vertex, and *mt_to_count* +

[1] http://leakforums.net/thread-719337

mt_replies as edge weight. The edge weight is the sum of messages sent to the recipient and the number of replies. The data extraction and analysis was performed on an Ubuntu 15.10 desktop computer, with Python scripts that we wrote for this purpose. The software used in this case study was MySQL and Python, with packages Networkx and MySQLdb.

4 Results

This section contains the result from four centrality measures on two (undirected) graphs. Which are divided into public and private, as seen in Table 2 and 3 respectively. Tables are sorted in descending order by their centrality value, because higher values indicate more central positions in the respective measures. They are limited to the first five results due to page limitations, however, it is enough to demonstrate that users are ranked differently according to centrality measures' interpretation of important.

Table 2: Top ten public centrality results

ID	Degree	ID	Closeness
15398	0.31449	15398	0.51280
1337	0.06518	1337	0.44481
5481	0.03564	334	0.42281
16618	0.03036	3507	0.42001
410101	0.02872	2902	0.41946
ID	**Betweenness**	**ID**	**Eigenvector**
15398	0.50134	15398	0.47951
1337	0.07594	1337	0.21054
5	0.02790	334	0.14043
5481	0.02403	5481	0.11948
411677	0.02365	4782	0.10452

Table 3: Top ten private centrality results

ID	Degree	ID	Closeness
1	0.09466	1	0.37928
15398	0.03441	334	0.35757
1337	0.03275	1471	0.35631
1471	0.03194	1337	0.35437
51349	0.03074	51349	0.35118
ID	**Betweenness**	**ID**	**Eigenvector**
1	0.17174	193974	0.48531
15398	0.05871	61078	0.47249
1337	0.04811	51349	0.29031
1471	0.04593	315929	0.24046
334	0.03985	336307	0.16937

The values has been normalised so networks of different sizes can be compared with each other. Networkx can normalise the results for us. All values in Table 2 and 3 have been normalised in the range $[0, 1]$, according to equations found in [8].

We started the analysis on users that occopied similar ranks between each centrality measures and type of communication, to understand why they get their ranks. It was performed by manually inspecting the message contents, and it revealed that many of these individuals had roles such as administrator and moderators in Nulled.IO. In addition to having responsibilities and being active on the forum, they also contributed with cracked software (mostly cheats for games) and distributing user credentials. Users in eigenvector centrality, in Table 3, differed from users in the other centrality measures as the two highest ranking users (ID 193974 and 61078) was selling services of converting or trading between currencies.

Users with ID 1 and 15398 is ranking highest for degree centrality in both tables, up to 2.75 and 4.82 times larger than the second highest values respectively. But they get their values because they are connected to more neighbours than other users. This indicate that they are very active in the hacker forum by communicating with many different users.

In Table 3, user with ID 1 is 2.92 times larger than the second highest value in betweenness centrality, which indicate that this user is sitting in between a lot more users. However, results from closeness centrality indicate that the network is more connected. As it shows that users have about equally short path to all other users – as it only decreases by 0.053 after 100 users. There was only user with ID 15398 in Table 2 that had significant values in all of the centrality measures. Because of our approach to construct the public graph, this would indicate that the threads created by user ID 15398 is very popular, with 70 906 edges connecting to other users.

5 Conclusion

Organised criminal groups use anonymisation techniques to operate and move their illegal activities online. Where they form communities that are used as marketplaces for illegal materials, products and services. This drives the underground economy by providing services that facilitate almost any type of cyber crime. The challenge for law enforcement investigators is to know which individuals to focus their efforts and resources on, in order to disrupt the services provided by cyber criminals.

In this paper we assessed the performance of four graph-based centrality measures, for their ability to identify important individuals which provide valuable services to other cyber criminals. Centrality measures have previously been used to study small and structured data sets (for example Enron). However, we tested them on a newly leaked dataset that is larger, more loosely structured and a network with similarities to cyber criminal forums. Our result shows that well established graph-based measures have weaknesses when applied to this new dataset. For example, some individuals are ranked high and appears to be important to the forum. However, they actually had less important contribution to the community, and their removal would have a low impact on illegal operations from the criminal forum.

Investigators already have centrality measures available in tools they use, such as IBM i2 Analyst's Notebook. However, they need to understand that it is not a silver bullet that automatically identifies important users. To avoid accusing someone for being the leader in a cyber criminal network, further analysis is needed to confirm they are really important for investigator's goals. Focusing on wrong individuals can be illustrated with a real-world example: In 2013, Silk Road was taken down after arresting some of their administrators. After law enforcement interference, a dozen new marketplaces spawned and took Silk Road's place.

Another important aspect to improve the results is to pre-filter the dataset before analysis. This can be done by removing dependent vertices (i.e $C_D(v) = 1$), which can improve betweenness centrality results. The Nulled.IO forum should also have been analysed as a directed graph. Then additional centrality measures such as *in-* and *out-degree* would be able to identify users with high popularity and expansiveness respectively.

References

[1] K.R. Choo. Organised crime groups in cyberspace: a typology. *Trends in Organized Crime*, 11(3):270–295, 2008.

[2] J. Diesner and K. M. Carley. Exploration of Communication Networks from the Enron Email Corpus. *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*, pages 3–14, 2005.

[3] Europol. The internet organised crime threat assessment (iocta). `https://goo.gl/t83NQ4`, September 2014. Online; accessed December 9, 2016.

[4] J. Hardin, G. Sarkis, and P.C. Urc. Network analysis with the enron email corpus. 10 2014.

[5] Reece Howard. Using social network analysis measures. `https://goo.gl/OUkfrG`, July 2015. Online; accessed December 9, 2016.

[6] Valdis Krebs. Uncloaking terrorist networks. *First Monday*, 7(4), 2002.

[7] National Crime Agency. Cyber crime assessment 2016. `https://goo.gl/tKcxDN`, July 2016. Online; accessed December 9, 2016.

[8] C. Prell. *Social Network Analysis: History, Theory and Methodology.* Sage Publications Ltd., 2011.

[9] Malcolm K. Sparrow. The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks*, 13(3):251 – 274, 1991.