Identifying Central Individuals in Organised Criminal Groups and Underground Marketplaces[⋆]

Jan William Johnsen and Katrin Franke

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

Abstract. Traditional organised criminal groups are becoming more active in the cyber domain. They form online communities and use these as marketplaces for illegal materials, products and services, which drives the Crime as a Service business model. The challenge for law enforcement of investigating and disrupting the underground marketplaces is to know which individuals to focus effort on. Because taking down a few high impact individuals can have more effect on disrupting the criminal services provided. This paper present our study on social network centrality measures' performance for identifying important individuals in two networks. We focus our analysis on two distinctly different network structures: Enron and Nulled.IO. The first resembles an organised criminal group, while the latter is a more loosely structured hacker forum. Our result show that centrality measures favour individuals with more communication rather than individuals usually considered more important: organised crime leaders and cyber criminals who sell illegal materials, products and services.

Keywords: Digital Forensics · Social Network Analysis · Centrality Measures · Organised Criminal Groups · Algorithm Reliability.

1 Introduction

Traditional Organised Criminal Groups (OCGs) have a business-like hierarchy [10]; with a few leaders controlling the organisation and activities done by men under then, which does most of the criminal activities. OCG are starting to move their operations to the *darknet* [2,6], where they form *digital undergrounds*. These undergrounds serves as meeting places for likeminded people and marketplace for illegal materials, products and services. This also allow the criminal economy to thrive, with little interference from law enforcements [6].

The transition between traditional physical crime to cyber crime changes how criminals organise. They form a loosely connected network in digital undergrounds [11,8]; where the users can be roughly divided into two distinct

[⋆] The research leading to these results has received funding from the Research Council of Norway programme IKTPLUSS, under the R&D project "Ars Forensica - Computational Forensics for Large-scale Fraud Detection, Crime Investigation & Prevention", grant agreement 248094/O70.

groups [6]: a minority and a majority. The division is based on individual's technical skill and capabilities, and the group names reflects how many individuals are found in each group. The minority group have fewer individuals, however, they have higher technical skills and capabilities. They support the majority – without the same level of skills – through the Crime as a Service (CaaS) business model [6]. The consequence is that highly skilled criminals develop tools that the majority group use as a service. This allows entry-level criminals to have greater impact and success in their cyber operations.

The challenge is identifying key actors [12] in digital undergrounds and stopping their activities. The most effective approach is to target those key actors found in the minority group [6]. We represent the communication pattern between individuals as a network, and then use Social Network Analysis (SNA) methods to investigate those social structures. An important aspect of SNA is that it provides scientific and objective measures for network structure and positions of key actors [5]. Key actors – people with importance or greater power – typically have higher centrality scores than other actors [5,12].

We substitute the lack of available datasets of OCG and digital undergrounds with the Enron corpus and Nulled.IO, respectively. The Enron corpus have been extensively studied [7,4,3], where Hardin et al. [7] studied the relationships by using six different centrality measures. While Nulled.IO is a novel dataset for an online forum for distributing cracked software, and trade of leaked and stolen credentials. SNA methods have also been used by Krebs [14] to analyse the network surrounding the airplane highjackers from September 11th, 2001.

The dataset types in these studies are highly varied: ranging from a few individuals to hundreds of them; networks that are hierarchical or are more loosely structured; and complete and incomplete networks. The *no free lunch* theorem [15] states that there is no algorithm that works best for every scenario. The novelty of our work is that we evaluate the results of centrality measures for two dataset with distinctly different characteristics. Our research tries to answer the following research questions: 1) How does centrality measures identify leading people inside networks of different organisational structures and communication patterns? and 2) How good are they to identify people of more importance (i.e. inside the smaller population)? The answers to these questions are particularly important for law enforcement, to enable them to focus their efforts on those key actors whose removal has more effect for disruping the criminal economy.

2 Materials and Methods

2.1 Datasets

Although the Enron MySQL database dump by Shetty and Adibi [13] is unavailable today, we use a MySQL v5 dump of their original release[1]. The corpus contains 252 759 e-mail messages from 75 416 e-mail addresses. Nulled.IO[2] is

[1] `http://www.ahschulz.de/enron-email-data/`

[2] `http://leakforums.net/thread-719337` (recently became unavailable)

an online forum which got their entire database leaked on May 2016. The forum contains details about 599 085 user accounts, 800 593 private messages and 3 495 596 public messages. The distinction between private and public is that private messages are between two individuals, while public messages are forum posts accessible by everyone. These datasets have very different characteristics: Enron is an organisation with strict hierarchical structure, while Nulled.IO is flat and loosely connected network.

The challenge of analysing our datasets are the large amount of information they contain. Every piece of information would be of potential interest in a forensic investigation, however, we limit the information to that which represents individual people and the communication between then. We use this to create multiple *directed graphs* (digraphs), where individuals are modelled as *vertices* and the communication between them as directed *edges*. A digraph G is more formally defined as a set V of vertices and set E of edges, where E contains ordered pairs of elements in V. For example, (v_1, v_2) is an ordered pair if there exists an edge between vertices v_1 and v_2, called *source* and *target* respectively.

2.2 Centrality Measures

Centrality measures are graph-based analysis methods found in SNA, used to identify important and influential individuals within a network. We evaluate five popular centrality measures for digraphs: *in-degree* (C_{deg^-}), *out-degree* (C_{deg^+}), *betweenness* (C_B), *closeness* (C_C) and *eigenvector* (C_E). They are implemented in well-known forensic investigation tools, such as IBM i2 Analyst's Notebook [1].

The centrality measures differs in their interpretation of what it means to be 'important' in a network [12]. Thus, some vertices in a network will be ranked as more important than others. Figure 1 illustrate how vertices are ranked differently. The number of vertices and edges affects the centrality values. However, normalising the values will counter this effect and allow us to compare vertices from networks of different sizes. Our analysis tool *Networkx* uses a scale to normalise the result to values $[0, 1]$.

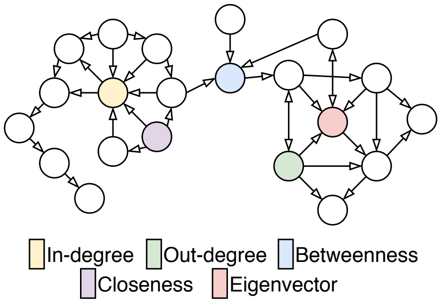

Fig. 1. Highest ranking vertices in a digraph

3 Experiment

We first constructed three weighted digraphs to represent the communication between users in Enron and Nulled.IO. Only a few database (DB) tables and fields had the necessary information to build the digraphs: *message (sender)* and *recipientinfo (rvalue)* for Enron, and *topics (starter_ id)*, *posts (author_ id)* and *message_ topics (mt_ starter_ id, mt_ to_ member_ id, mt_ to_ count, and mt_ replies)* for Nulled.IO. The digraph construction method can be generalised as: find the sender and receiver of messages. Represent them as unique vertices in a digraph (if not already exists) and connect them with an edge from sender to receiver. These operations was repeated for every public/private and e-mail message. Finally, edges' weights was initialised once it was first created, and incremented for each message with identical vertices and edge direction. The data extraction and analysis was performed on a desktop computer, with a MySQL server and Python, with packages Networkx v1.11 and PyMySQL v0.7.9.

3.1 Pre-filtering and Population Boundary

The digraph contruction included a bit more information than necessary, which have to be removed before the analysis. However, we want to find a balance between reducing the information without removing valuable or relevant information. To analyse the hierarchical structure of Enron (our presumed OCG), we have to remove vertices which does not end with '@enron.com'. Additionally, we removed a few general e-mail addresses which could not be linked to unique Enron employees. A total of 691 vertices was removed by these steps.

We have previously identified user with ID 1 as an administrator account on the Nulled.IO forum [9], used to send out private 'welcome'-messages to newly registered users, which can skew the results. Thus, we only remove edges from ID 1 to other vertices – when its weight equals one – to achieve the goal of information preservation. The private network, that originally had 295 147 vertices and 376 087 edges, was reduced to 33 647 (88.6%) and 98 253 (73.87%), respectively. The public thread communication network did not undergo any pre-processing. From its original 299 702 vertices and 2 738 710 edges, it was reduced to 299 105 (0.2%) and 2 705 578 (1.22%), respectively.

The final pre-processing step was to remove isolated vertices and self-loops. Isolated vertices had to be deleted because they have an infinite distance to every other vertex in the network. Self-loops was also removed because it is not interesting to know a vertex' relation to itself. The reduction in vertices and edges for the Nulled.IO public digraph was a consequence of this final step.

4 Results

The results are found in Tables 1 and 2, sorted in descending order according to vertices' centrality score. Higher scores appear on top and indicates more importance. The results are limited to the top five individuals due to page limitiations.

The goal of our research is to identify leaders of OCG or prominent individuals who sell popular services; people whose removal will cause more disruption to the criminal community. To evaluate the success of centrality measures, we first had to identify people's job positions or areas of responsibilities in both Enron and Nulled.IO. We combined information found on LinkedIn profiles, news articles and a previous list[3] to identify Enron employees' position in the hierarchy. For Nulled.IO users we had to manually inspect both private and public messages to estimate their role or responsibility. The total number of possible messages to inspect made it difficult to determine the exact role for each user.

4.1 Enron

sally.beck is within the top three highest ranking individuals in all centrality measures, except for eigenvector centrality. Her role in Enron was being a Chief Operating Officer (COO); responsible for the daily operation of the company and often reports directly to the Chief Executive Officer (CEO). Her result correspond to expectations of her role: a lot of sent and received messages to handle the daily operation.

Table 1. Top ten centrality results Enron

UID	C_{deg^-}	UID	C_{deg^+}	UID	C_B
louise.kitchen	0.03374	david.forster	0.07393	sally.beck	0.02152
steven.j.kean	0.02900	sally.beck	0.06559	kenneth.lay	0.01831
sally.beck	0.02884	kenneth.lay	0.04982	jeff.skilling	0.01649
john.lavorato	0.02803	tracey.kozadinos	0.04955	j.kaminski	0.01555
mark.e.taylor	0.02685	julie.clyatt	0.04907	louise.kitchen	0.01145

UID	C_C	UID	C_E
sally.beck	0.39612	richard.shapiro	0.37927
david.forster	0.38500	james.d.steffes	0.33788
kenneth.lay	0.38362	steven.j.kean	0.27800
julie.clyatt	0.38347	jeff.dasovich	0.27090
billy.lemmons	0.38293	susan.mara	0.25839

kenneth.lay and *david.forster* are two individuals with high rankings in all centrality measures, except for eigenvector centrality. They are CEO and Vice President, respectively. *kenneth.lay* and his second in command *jeff.skilling* was the heavy hitters in the Enron fraud scandal.

Although there where a few CEOs in the Enron corporation, many of the higher ranking individuals had lower hierarchical positions. Most notably this occurred in eigenvector centrality, however, this is because of how this measure works. Finally, our result also show that centrality measures usually ranks the same individuals as being more important than others.

4.2 Nulled.IO

Unique Identifier (UID) 0 in the public digraph appears to be a placeholder for deleted accounts, because the UID does not appear in the member list and the username in published messages are different. UID 4, 6, 8, 15398, 47671 and 301849, among others, provides free cracked software to the community, with most of them being cheats or bots for popular games. While UID 1337 and 1471 appears to be administrators.

[3] `http://cis.jhu.edu/~parky/Enron/employees`

Table 2. Top five public and private centrality results

Public centrality results									
UID	C_{deg^-}	UID	C_{deg^+}	UID	C_B	UID	C_C	UID	C_E
15398	0.23695	1471	0.00393	0	0.00959	1471	0.03564	1337	0.28764
0	0.16282	8	0.00321	15398	0.00855	8	0.03553	0	0.27157
1337	0.06466	193974	0.00294	1337	0.00461	118229	0.03542	15398	0.25494
4	0.05656	47671	0.00273	1471	0.00334	169996	0.03540	334	0.23961
6	0.04276	118229	0.00266	193974	0.00219	47671	0.03520	71725	0.22798

Private centrality results									
UID	C_{deg^-}	UID	C_{deg^+}	UID	C_B	UID	C_C	UID	C_E
1	0.08412	1	0.42331	1	0.41719	1	0.40665	61078	0.45740
1471	0.05028	51349	0.00773	1471	0.02369	51349	0.28442	51349	0.30353
1337	0.04289	88918	0.00695	334	0.02286	88384	0.28102	1	0.24505
8	0.03970	47671	0.00617	1337	0.02253	10019	0.28080	88918	0.21214
15398	0.03967	334	0.00600	15398	0.02129	61078	0.28043	193974	0.19651

UID 1 in the private digraph is found on top of (almost) all centrality measure. Although this account appear as a key actor, it was mostly used to send out thousands of automatic 'Welcome'-messages, 'Thank you'-letters for donations and support and other server administrative activities.

UIDs 8, 334, 1471, 47671, 51349 and 88918 in the private digraph cracks various online accounts, such as Netflix, Spotify and game-related accounts. They usually go after the 'low hanging fruit' that have bad passwords or otherwise easy to get. Most of the users have low technical skills, however, they are willing to learn to be better and to earn more money from their scriptkid activities. They want to go into software development for economic gains or learn more advanced hacker skills and tools to increase their profit.

5 Discussion and Conclusion

Law enforcement agencies can disrupt the CaaS business model or OCG when they know which key actors to effectively focus their efforts on. However, implementations of centrality measures in forensic investigation tools are given without any explanation or advice for how to interpret the results; which inadvertently can lead to accusation of lesser criminals of being among the leaders of criminal organisations. Although the centrality measures do not perfectly identify individuals highest in the organisation hierarchy, our result show that potential secondary targets can be found via them. Secondary targets are individuals that any leader rely on to effectively run their organisation.

Contemporary centrality measures studies here most often identifed individuals with a natural higher frequency of communication, such as administrators and moderators. However, going after forum administrators is only a minor setback, as history has shown a dozen new underground marketplaces took Silk Road's place after it was shut down. Thus, the problem with current centrality measures is that they are affected by the network connectivity rather than actual criminal activities.

Our result demonstrates their weakness, as centrality measures cannot be used with any other definition for their interpretation of 'importance'. There is a lack of good interpretations to current centrality measures that fits for forensic investigations. Interpretations which are able to effectively address the growing problem of cyber crime and the changes it brings. We will continue working on identifying areas where already existing methods are sufficient, in addition to developing our own proposed solutions to address this problem.

References

1. IBM knowledge center - centrality and centrality measures, https://www.ibm.com/support/knowledgecenter/en/SS3J58_9.0.8/com.ibm.i2.anb.doc/sna_centrality.html
2. Choo, K.K.R.: Organised crime groups in cyberspace: a typology **11**(3), 270–295. https://doi.org/10.1007/s12117-008-9038-9, http://link.springer.com/10.1007/s12117-008-9038-9
3. Diesner, J.: Communication networks from the enron email corpus: "it's always about the people. enron is no different." https://pdfs.semanticscholar.org/875b/59b06c76e3b52a8570103ba6d8d70b0cf33e.pdf
4. Diesner, J., Carley, K.M.: Exploration of communication networks from the enron email corpus. In: SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA
5. Décary-Hétu, D., Dupont, B.: The social network of hackers **13**(3), 160–175. https://doi.org/10.1080/17440572.2012.702523, http://www.tandfonline.com/doi/abs/10.1080/17440572.2012.702523
6. Europol: The internet organised crime threat assessment (iOCTA) 2014, https://www.europol.europa.eu/sites/default/files/documents/europol_iocta_web.pdf
7. Hardin, J.S., Sarkis, G., Urc, P.C.: Network analysis with the enron email corpus **23**(2)
8. Holt, T.J., Strumsky, D., Smirnova, O., Kilger, M.: Examining the social networks of malware writers and hackers **6**(1), 891
9. Johnsen, J.W., Franke, K.: Feasibility study of social network analysis on loosely structured communication networks **108**, 2388–2392. https://doi.org/10.1016/j.procs.2017.05.172, http://linkinghub.elsevier.com/retrieve/pii/S1877050917307561
10. Le, V.: Organised crime typologies: Structure, activities and conditions **1**, 121–131
11. Macdonald, M., Frank, R., Mei, J., Monk, B.: Identifying digital threats in a hacker web forum. pp. 926–933. ACM Press. https://doi.org/10.1145/2808797.2808878, http://dl.acm.org/citation.cfm?doid=2808797.2808878
12. Prell, C.: Social Network Analysis: History, Theory and Methodology. SAGE Publications, https://books.google.no/books?id=p4iTo566nAMC
13. Shetty, J., Adibi, J.: The enron email dataset database schema and brief statistical report **4**, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.296.9477&rep=rep1&type=pdf
14. Valdis Krebs: Uncloaking terrorist networks | krebs | first monday, http://journals.uic.edu/ojs/index.php/fm/article/view/941
15. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization **1**(1), 67–82, http://ieeexplore.ieee.org/abstract/document/585893/