Leiden University

Computer Science
Data Science Track

Identifying key players in child exploitation networks on the Dark Net

Name:	Alain Fonhof
Date:	29/08/2018
1st supervisor:	Dr. F.W. Takes
2nd supervisor:	Dr. C.J. Veenman
External supervisor:	Madeleine van der Bruggen

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Child pornography is one of the key social security problems. With the digitization of society the scale of this problem is increasing. Whereas law enforcement used to find a few dozen or hundreds of photos on the spot in a large child pornography case, it is now terabytes of data that can be distributed worldwide. The dark net, which is a hidden part of the world wide web, enables so-called child exploitation networks in which users communicate and distribute resources in a secure and anonymous manner. As such, it has become increasingly valuable to have the right tools that help law enforcement prioritize whom they focus on. We propose a social network approach to identify key players and discover the structure of these child exploitation networks. This approach enables law enforcement to use the relational data of encrypted messages and messages of different languages. In this thesis we will study two real-world data sets that were collected from two different online discussion forums on the dark net. These social networks are connected by *topic-to-person* relationships and we will research the effect of projecting it to a *person-to-person* network. This thesis presents and discusses ways of applying and interpreting established social network algorithms to both the *topic-to-person* and the *person-to-person* network. The findings in this thesis will help law enforcement gain more insights into the behaviour of users on these dark networks. Firstly, we were able to reach a 79% rank-order correlation with previous research of the Dutch National Police and discover up to 81.25% of the admins on a network. Secondly, we detected an anti-lurker and application policy on a child exploitation network. Thirdly, we found that adding weight to the *person-to-person* network during projection boosts performance. Lastly, we discuss a method that enables us to remove up to 25% of the edges in the *person-to-person* network without losing information.

Acknowledgements

I would like to thank Frank Takes, Madeleine van der Bruggen and Arthur van Bunningen for their discussions and supervising me during this research. Furthermore, I would like to thank the Dutch National Police for the internship and the opportunity to do research on this sensitive topic.

Contents

–If you can do what you do best and be happy, you're further along in life than most people.

Leonardo DiCaprio

1

Introduction

The Dutch National Police defined their goals to combat child pornography in the security agenda 2015-2018 by the Ministry of Justice and Security [1]. The Ministry has set out two common objectives for the Dutch National Police: 1) reduce child pornography and child sex tourism and 2) increase the commitment to signals of abuse, actual abuse and the relieving of victims.

Child pornography is one of the key social security problems but is also a cross-border problem due to the nature of the internet. Child pornography is classified as sexual violence against children on visual material. The distribution and deliberate watching of child pornography is also seen as sexual violence. Child pornography has serious damaging effects on the victims and the victimization never ends after the offence because images can never be fully erased from the internet. The impact can range from physical to long-term psychological effects such as depression, anxiety, PTSD, low self-esteem, difficulty establishing healthy relationships and ongoing humiliation [2]. The Rutgers Nisso Groep survey has shown that 20% of Dutch women and 4% of Dutch men - in their opinion - have experienced a form of sexual harassment (excluding offensive comments) under the age of 16 [3]. With the digitization of society the scale of this problem is increasing. The digitization has also made it easier for pedophiles to get access to child pornography. With the creation of the dark net this access has also been made more secure and anonymous. The dark net is a part of the world wide web that exists on hidden networks that use the internet but require specific software, configurations or authorization to access [4]. Dark net websites are

accessible only through networks such as Tor and I2P. While Tor focuses on providing anonymous access to the internet, I2P specializes on allowing anonymous hosting of websites. Identities and locations of dark net users stay anonymous due to the layered encryption system which makes it almost impossible to track these users. Due to the anonymity of the dark net the contents of these websites contain primarily illegal products and services. Another consequence of this anonymity is the rise of many platforms where users can openly share and talk about child pornography.

The pace at which developments such as digitization and internationalization occur and the increasing impact they have on our society and thus also on crime within our society, means that new challenges arise and that existing phenomena can change in nature. For example, where before dozens or hundreds of photos were found on the spot in a large child pornography case, it is now about terabytes of data that can be distributed worldwide. The Dutch National Police has set out goals to start more investigations focused on child abuse, production of child pornography and child sex tourism. These complex cases concern proactive and regular investigations into actual abuse and are extensive and therefore labor intensive. Due to these factors it has become increasingly valuable to have the right tools that help the police prioritize whom they focus on.

In order to support these challenges new automated methods are being developed. These methods can process large amounts of data and help in either automating certain tasks or give the detectives better insights into the underlying data. These insights can result in a more effective and thought out control strategy by the police. The underlying data that we will be focusing on in this thesis is originating from online discussion forums on the dark net. Pedophiles use these forums to communicate and distribute resources or information. Members can post a statement or question under the general heading of the forum, to which other members can respond. The community-aspect and sense of belonging is much greater. Offenders get to know each other as if it were real life and develop long lasting relationships. These networks completely normalize child exploitation material, as their main goal is the promotion and distribution of child pornography [5]. Such platforms are often moderated and organized in a professional manner. The number of users can range from a few thousand to well over one million. Unfortunately the police does not have the resources to investigate every visitor. Therefore it is important to target key players that are vital to the existence of these forums. In this thesis we will explore this data and attempt to automatically identify these key players using methods from the field of social network analysis.

Social Network Analysis (SNA) is strongly related to the broader field of network science [6]. SNA focuses on analyzing and gaining insight from social data. Social networks are classified as the sum of all professional, friendship and family ties [6]. It consists of the connections in a society and determines the spread of knowledge, behavior and resources. In 1991 one of the first papers was published that explored the opportunities for the application of network analytic techniques to the problems of criminal intelligence analysis, paying particular attention to the identification of vulnerabilities in different types of criminal organizations — from terrorist groups to narcotics supply networks [7]. Two decades later a lot has changed, whereas state-of-the-art in 1991 consisted of simple visualization charts and mainly manual work nowadays most tasks are computerized and capable of handling millions of users. New algorithms are developed to compute metrics about users, the flow of information and to visualize a network. Recently the possibilities and limitations of a data-driven approach to study criminal networks were explored [8]. This work discussed several use cases that used SNA. One of these cases was "operation Blackbird", which was a Dutch investigation against a criminal group involved in organized cannabis cultivation. After the investigation, even though three key suspects were arrested, the process of cannabis cultivation continued. With SNA they concluded that this was due to the fact that these suspects were connected to other well connected criminals who could replace them. It was also concluded that further control strategies take into account the active and important participation of women and direct relatives in the organization of criminal activities. In another case a police unit wanted to target synthetic drug producers that presented their services to a large criminal network. Through SNA and the value chain a calculated intelligence strategy was developed to identify four chemists that were interchangeable. Three of them were efficiently targeted. This resulted in considerable delays of the synthetic drugs production and disruption of the network. One of the biggest limitations however is that SNA is considered to be too slow for law enforcement. This is because data has to be gathered over a period of time before analysis can be done and because it requires in-depth qualitative analysis. These use cases however show opportunity of using SNA to get insights into the bigger picture of a criminal network and develop a suiting control strategy.

In order to construct our social network we will use the data from the dark net forums. A forum consists of topics and allows multiple people to respond on a certain topic. This activity results in a *topic-to-person* network. Links exist only between a topic and person. Which implies that people are linked through a topic that they comment on, because they are interested in the same topic. A network with two different entities is in network-terminology called a two-mode or bipartite network. Unfortunately, a significant collection of notions and tools to analyze network structure are based on one-mode networks. This is due to the fact that in a two-

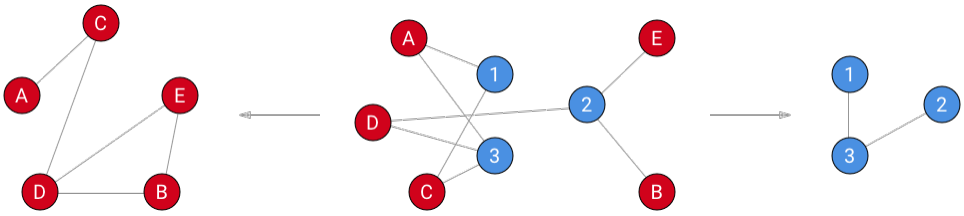

Figure 1.0.1: Projection from a bipartite graph (center) to its ⊤-projection (right) and it ⊥-projection (left).

mode network it would be the same as comparing apples to oranges and an one-mode network has only one entity. Therefore we want to project our network to a *person-to-person* network. Projection is done by linking two people in the one-mode network if they both have a link to the same topic in the two-mode network. See Figure 1.0.1 for an example. It is important to realize there is much information in the two-mode network that is lost after projection. For example, we lose how many topics two people respond to or how big a certain topic is that connects people because in a one-mode network these two people are simply linked. It is also important to note the fact that different two-mode graphs can lead to the same projection [9]. The projection can also lead to properties that are a result of the projection process rather than the underlying social structure. This is caused by large neighbourhood overlap in one of the entity sets in the two-mode network. Therefore a topic that is linked to many people, which we call a 'big linker', reduces the importance of a link in the projection. For example, if two people are linked because they commented on a big generic 'Introduce yourself' topic this link is of less significance than if they commented on a specific 'How to encrypt your connection' topic. For these reasons we want to research the impact of projection. We will explore different kinds of projection algorithms and look into certain pre/post-processing filters that we can apply. One example of a filter on the two-mode network is to remove a number of big linkers.

We will analyze the *person-to-person* network by dividing the network structure in three different levels:

1. Macro level; network topology
2. Meso level; groups within the network
3. Micro level; individual actors

It is important to first understand the network topology before we can clarify the individual positioning and groups. These groups can be based on topic categories or user roles. For instance multiple topics belong to the 'technical' category. User roles originate from the structured organization of these forums. Users share their knowledge in certain domains or get promoted

to a moderator. The labeling of these groups is done in previous research by the Dutch National Police which we will specify in Section 4.2. An approach to distinguish between groups is to look at certain metrics that might indicate a variance in the distribution of two groups.

We are interested in the following research questions. These will provide us with more insights into the network structure of child pornography on the dark net.

1. Which method can identify actors that play a significant role in the network?
2. Which type of projection is best suited to replicate the underlying social structure?
3. What is the impact of filters on the projected one-mode network?
4. Can we identify certain groups of users with only the relational data?

Thesis Overview

In Chapter 2 we formulate the notions and algorithms that are used in this thesis. Related research is discussed in Chapter 3. In Chapter 4 the used data sets and certain characteristics are presented. The proposed methodology is explained in Chapter 5. Next, the experiments are evaluated in Chapter 6. Finally, conclusions are drawn and future work is suggested in Chapter 7.

–That there's some good in this world, Mr. Frodo; and it's worth fighting for.

Samwise Gamgee

2

Preliminaries

In this chapter we will discuss the terminology used in this thesis. We use basic notions for the topology of a two-mode and one-mode network.

2.1 Two-mode networks

The type of network that we will work with is known as a two-mode or bipartite network. A two-mode network is made up of two different sets of vertices and ties exist only between nodes belonging to different sets. A distinction is often made between the two vertex sets based on which set is considered more responsible for tie creation (primary or top vertex set) than the other (secondary or bottom vertex set). In this thesis we will use the notation defined by Latapy et al. [10]. We denote a bipartite graph as $G = (\top, \bot, E)$ where $\top$ is the set of top vertices, $\bot$ is the set of bottom vertices and $E \subseteq \top \times \bot$ is the set of edges. We will be working with a set of forum topics and people. Since a person comments on a topic we consider the topics to be our top vertex set. In addition we consider our edges to be undirected because topics can not make an edge. When two people comment on the same topic this creates a connection between these people through this topic.

To analyze the topology of the two-mode network we will use the following measurements: number of top and bottom nodes, number of edges, average degree of all, top and bottom nodes, density, average clustering coefficient

and the average min- and max-clustering coefficient, all defined below.

First we will denote the number of top and bottom nodes with $n_\top = |\top|$ and $n_\bot = |\bot|$ respectively and the total number of nodes with $n = n_\top + n_\bot$. Second we denote the number of edges with $m = |E|$. We can then define the top and bottom average degree as $k_\top = (m/n_\top)$ and $k_\bot = (m/n_\bot)$ respectively. This leads to a total average degree in the graph $G' = (\top \cup \bot, E)$ as $\langle k \rangle = (2m/(n_\top + n_\bot)$. The bipartite density is defined by $\delta(G) = (m/n_\top n_\bot)$ which is the fraction of existing links related to possible ones. Clustering coefficient does not make sense in a bipartite graph since it relies on the enumeration of the triangles in the graph. However there are no triangles in a bipartite graph. Therefore another notion is discussed by Latapy et al. [10] that defines the clustering coefficient of a single node by the average clustering coefficient with other nodes excluding nodes that share no neighbors (see Equation 2.1.1). The clustering coefficient of a pair $cc_\bullet(u, v)$ can be computed in three ways. Equation 2.1.2 is the generalization of the basic notion [11]. If u and v share no neighbor then $cc_\bullet(u, v) = 0$ and if they have the same neighborhood $cc_\bullet(u, v) = 1$. The drawback of this notion however is that if one of the two nodes has a higher degree than the other then $cc_\bullet(u, v)$ will be small. Even if one of the neighbors is completely contained in the other. In order to capture this we define the min- and max-clustering. Min-clustering is equal to 1 when one neighborhood is included in the other (see Equation 2.1.3). Max-clustering is equal to 1 when both neighborhoods are the same and decreases swiftly if the degree of one of the nodes increases (see Equation 2.1.4). To get the average clustering coefficient of a graph we can use $cc_N(G) = \frac{1}{n} \sum_{u \in V} cc_\bullet(u)$.

$$cc_\bullet(u) = \frac{\sum_{v \in N(N(u))} cc_\bullet(u, v)}{|N(N(u))|} \tag{2.1.1}$$

$$cc_\bullet(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|} \tag{2.1.2}$$

$$cc_{\underline{\bullet}}(u, v) = \frac{|N(u) \cap N(v)|}{\min\left(|N(u)|, |N(v)|\right)} \tag{2.1.3}$$

$$cc_{\overline{\bullet}}(u, v) = \frac{|N(u) \cap N(v)|}{\max\left(|N(u)|, |N(v)|\right)} \tag{2.1.4}$$

2.2 Projection and filtering

A transformation from a two-mode to an one-mode network is called a projection. We notate the $\top$-projection of graph G as $G_\top = (\top, E_\top)$. Two nodes of $\top$ are linked if they share at least one neighbour in the two-mode network $E_\bot = \{(u, v), \exists x \in \top : (u, x) \in E \text{ and } (v, x) \in E\}$. The $\bot$-projection is defined analogously. In this thesis we are interested in the *person-to-person*

network and are therefore only interested in the $\perp$-projection. There are different ways to do a projection. The difference between these methods is in the weight that is assigned to each edge.

2.2.1 Projection methods

We will study the following projection methods:

1. Unweighted projection
2. Weighted projection
3. Newman collaboration model

In our first method we simply assign a weight of 1 to each edge. One of the issues that we mentioned in Chapter 1 with regards to information loss of a projection is the fact that the number of topics that two people commented on is lost. With the second method we will solve this issue by using a weighted projection where the weight of each edge is the number of common topics that they commented on: $w_{u,v} = |N(u) \cap N(v)|$. Another issue arises with big linkers in the two-mode graph that connect many people in the one-mode graph. We want to capture this effect by reducing the importance of a big linker. This results in generic topics with many comments to be less important than specific topics with a few comments. In our third method we will assign weight using Newman's collaboration model to simulate this [12]. The definition of this model can be found in Equation 2.2.1 where u and v belong to the $\perp$ node set and x to the $\top$ node set. The value of $k(x)$ is the degree of x in the two-mode network and δ_u^x is 1 if node u is linked to node x in the two-mode network or 0 otherwise. These types of projection could help us cope with the information loss. It is important to note that we inverse the weights ($w_{u,v} = \frac{1}{w_{u,v}}$) for algorithms that use distance such as average shortest path, closeness centrality and betweenness centrality (defined in Section 2.3 and 2.4).

$$\begin{aligned} w_{u,v} &= \sum_x \frac{\delta_u^x \delta_v^x}{k(x) - 1} \\ \delta_u^x &= \begin{cases} 1 & \text{if } (u,x) \in E \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{2.2.1}$$

2.2.2 Filtering

In order to combat the effect that projection has on the resulting network we can apply a filter on the nodes or edges in our two-mode network. In this thesis we want to study the effect of such a filter on the underlying

structure in the one-mode network. Since big linkers connect many people in the one-mode network we are going to study the effect of removing these. To compute the biggest linker B we simply find the node with the highest degree in the top node set (see Equation 2.2.2). To remove i biggest linkers we repeat this process i times.

$$B = \arg\max_{v \in \top} k(v) \tag{2.2.2}$$

2.3 One-mode networks

After applying projection we have an one-mode social network. The network is considered a social network since the vertices consist of people that share and interest. Social networks have been studied extensively [6]. The key features of a social network is that they consist of one giant component, the degree distribution often follows a power law, they have high clustering and modularity and a small average path between two random connected nodes. A giant component is often called the largest connected component and there is a path between each pair of nodes in the connected component. If the degree distribution fits a power law function then the network is scale-free. One of the implications of the scale-free property is that the network robustness counts on actors with a relative high degree. If these actors are attacked the network may fall apart into subnetworks [13]. A power law function is defined as:

$$p(k) \sim k^{-\gamma} \tag{2.3.1}$$

The exponent γ describes how rapid the number of nodes fades with increasing degree. A higher γ indicates a sharper slope and accordingly less nodes with a relative higher degree. A social network is also often referred to as a small-world network [14]. In this thesis we will use basic graph notations [10]. We notate such a one-mode graph as $G' = (V, E)$ where V is the set of vertices and E is the set of undirected edges. The notation for the neighbourhood of a node v is $N(v) = \{u \in V : (u, v) \in E\}$. Each node in $N(v)$ is a neighbour of v and $k(v) = |N(v)|$ is the number of nodes in $N(v)$.

In order to analyze the network topology we will look at the following measurements: number of nodes and edges, average degree, density, average path length, diameter, degree assortativity coefficient and average clustering coefficient, all defined below.

We denote the number of nodes with $n = |V|$ and edges with $m = |E|$. The average degree is computed by averaging the degree over each node in the set (see Equation 2.3.2). The shortest path between nodes u and v is the path with the fewest number of edges. This is generally called the distance between node u and v and is denoted as $d(u, v)$ or commonly d. The average path length is denoted by $\langle d \rangle$ and is the average distance between each pair

in the network (see Equation 2.3.3). The diameter is denoted by d_{max} and is the maximum distance between two pairs in the network. The degree assortativity coefficient describes in a precise way how vertices of different types are preferentially connected amongst themselves based on their degree. The degree assortativity coefficient r is defined in Equation 2.3.4 where $a_i = \sum_j e_{ij}$ and $b_j = \sum_i e_{ij}$, and e_{ij} is the fraction of edges from a vertex with degree i to a vertex with degree j. When r > 0 the network is said to be assortative and when r < 0 it is disassortative. The average clustering coefficient c is defined in Equation 2.3.5, where a "connected triple means a single vertex with edges running to an unordered pair of others [15].

$$\langle k \rangle = \frac{1}{n} \sum_{v \in V} k(v) \tag{2.3.2}$$

$$\langle d \rangle = \frac{1}{n(n-1)} \sum_{\substack{u,v \in V \\ u \neq v}} d(u,v) \tag{2.3.3}$$

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_j}{1 - \sum_i a_i b_j} \tag{2.3.4}$$

$$c = 3 \times \frac{\text{number of trianlges in the network}}{\text{number of connected triples of vertices}} \tag{2.3.5}$$

2.4 One-mode metrics

Once we have an one-mode *person-to-person* network we can analyze these people on an individual level. For this analysis we will use the following measurements: average neighbor degree, clustering coefficient, degree centrality, closeness centrality, betweenness centrality, eigenvector and PageRank, all defined below.

Average neighbor degree is the average degree of all neighbors. We can use the eigenvector to determine the power of an actor. According to Bonacich [16] the more well-connected the actors in your neighborhood are, the more central you are. The less well-connected the actors in your neighborhood, the more powerful you are [16], since these neighbors are dependent on the well-connected actor to connect them to the network. The clustering coefficient computes the fraction of cases in which node u and v are connected with w, u and v are also connected and form a triangle. We define clustering coefficient in Equation 2.4.1, where $\left|E_{N(v)}\right| = E \cap (N(v) \times N(v))$ is the set of links between neighbours of v. Degree centrality for a node v is the fraction of nodes it is connected to, normalized by dividing by the maximum possible degree (see Equation 2.4.2). Closeness centrality computes the distance of v to each other node in the network (see Equation 2.4.3). Thus the more central a node is, the closer it is to all other nodes. We normalize $C_C(v)$ by

$n_v - 1$ where n_v is the size of the connected component that contains node v. Betweenness centrality $C_B(v)$ computes the number of shortest paths that run through v. It is defined in Equation 2.4.4 where σ_{st} is the number of shortest paths from s to t and $\sigma_{st}(v)$ is the number of shortest paths from s to t that pass through vertex v. We normalize this by dividing by the number of pairs of vertices not including v, which is $\frac{(n-1)(n-2)}{2}$. We normalize betweenness and closeness centrality so we can compare it between data sets. PageRank is an algorithm developed by Google that ranks linked elements in a set based on their importance [17]. The computations require several iterations to adjust PageRank value to more closely reflect the theoretical true value. In our social network the PageRank algorithm outputs a probability distribution used to represent the likelihood that when you take random steps in the network you will arrive at any particual person. These metrics give us a clear ranking of individuals and help us understand the network structure on a micro level.

$$c(v) = \frac{2\left|E_{N(v)}\right|}{k(v)(k(v)-1)} \tag{2.4.1}$$

$$C_D(v) = \frac{1}{n-1}k(v) \tag{2.4.2}$$

$$C_C(v) = \frac{1}{\sum_u d(u,v)} \tag{2.4.3}$$

$$C_B(v) = \sum_{\substack{s \neq v \neq t \in V \\ s \neq t}} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.4.4}$$

–Perfecting oneself is as much unlearning as it is learning.

Edsger Dijkstra

3

Related Work

Criminal networks have been studied by criminologists for decades but with the rise of the internet a new medium has erupted for criminals. It gives individuals the opportunity to sell illegal products and services directly to the customer from their home. The dark net has further made this interaction completely anonymous. Van der Bruggen & Blokland [5] did a literature study about child pornhography on the internet. First, they argued that child exploitation networks could be classified as criminal organizations because they firmly advertise and inspire others into the commission of crimes. On top of that there is a strong division of roles and these networks operate their criminal activities in an organized manner through a management hierarchy. Second, they concluded that child exploitation forums offer pedophiles a platform to meet like-minded people. Due to this community aspect they advocate for a network driven approach to investigate these communities. Third, they highlight that looking at the field of online child exploitation from an organized crime perspective, tracking down and focusing on the most important members of these communities could professionalize the combat against this type of crime and boost law enforcement efforts.

SNA has been applied to criminal networks. Duijn et al. [8] did research on the field of detecting and disrupting criminal networks with SNA. Going greatly in-depth into the possibilities and limitations of SNA in engaging criminal networks. They also examined 34 papers in this field. Themes of these papers are: organized cannabis cultivation, cocaine import, synthetic drugs, human trafficking, money laundering, youth groups, cold cases, cor-

ruption, illegal fireworks trafficking and vehicle theft. Since none of these themes focus on child exploitation we consider further research of SNA on child exploitation networks to be important.

Westlake et al. [18] studied child exploitation networks, defining a network as a connection between websites. In their research they identified the most important website to target for law enforcement. In order to do this they built a custom crawler called Child Exploitation Network Extractor (CENE) and defined a metric called network capital. Network capital is a combination of connectivity to other websites and the severity of the content on a website. However in this thesis we will zoom in on a website and study the connections between people active on this website.

Latapy et al. [10] introduce a set of metrics to capture properties of interest in two-mode networks. They provide an alternative to the projection approach. However they emphasize that (weighted) projection approaches also produce compelling insight and that the two approaches should be used interdependent to thoroughly understand the properties of two-mode networks. Therefore in this thesis we will focus on the projection approach.

Borgatti [19] proposes two algorithms to find sets of key players in a social network. He also demonstrated why existing graph-theoretic methods along with the naïve centrality-based heuristic fail to solve these two problems. The following two algorithms were proposed: KKP-POS and KKP-NEG. KKP-POS identifies key players for the purpose of optimally transmitting something through the network by using these key players as sources. KKP-NEG identifies key players for the purpose of disrupting or fragmenting the network by eliminating the key players. The algorithm takes as input n and finds the set of n players that optimally solves the problem. This seems to be an interesting metric if resources are limited. For example, if we have a team of five detectives that each will be assigned one person to investigate then this algorithm could be useful to efficiently assign tasks. However in our thesis we want to rank the entire user base and will therefor exclude this from the scope.

–Science is organized knowledge. Wisdom is organized life.

Immanuel Kant

4

Data

In this thesis we will use two different data sets which we will call data set A and data set B due to being law enforcement sensitive data. The data originates from two distinct child exploitation forums on the dark net.

4.1 Forums

Data set A is a forum that was crawled from the 8th of December 2010 till the 12th of September 2014 and consists of 14659 users. In order to get access to this forum users had to provide content that had to be verified by admins. It also had a tiered system which means that users were given access to special boards if they presented more unique or self produced material. This allowed them to gain prestige in the network by actively contributing. Unfortunately the tier that a user is in was not available to us in the metadata. Data set B is a forum that was crawled from the 1st of July 2015 till the 12th of October 2017 and consists of 21257 users. This forum had a standard approach to user registration. Users had to fill in an username and password which then gave them access to all boards. There were just a few protected boards for producers and administrators. In Section 6.2 we will explore these two-mode networks and their network metrics.

	Data set A	Data set B
Managers	0.49%	3.52%
Abusers	0.59%	4.1%
Technical	0.4%	3.14%
Embedded	98.3%	89.24%

Table 4.2.1: Distribution of assigned roles in PIM analysis

4.2 Membership roles

Researchers at the Dutch National Police studied these two data sets and the study is called the Program Identifying Main Targets (PIM) analysis. This analysis was partially based on research by Nolker et al. [20] and combines SNA with TF*IDF weighting to determine membership roles of communities in a network. The researchers at the Dutch National Police added an extra layer based on text analysis to strengthen the membership role classification. Users were divided into four groups: Managers, Abusers, Technical and Embedded. Managers are responsible for organizing the forum, recruiting and welcoming new members and enforcing rules. Abusers are producing material and are fanatic about their work. They share experiences and fantasies with the community while also encouraging others to commit criminal activities. Technical users focus on developing software and providing technical support to other users in the network. Embedded users are those that do not fall into any of the first three groups and are therefore considered to be not a key player by the PIM analysis. This data allows us to label users based on their role and study characteristics of these roles. Table 4.2.1 shows the fraction of the total number of users that are classified as a certain role. Besides the classification there is also a PIM ranking. This ranking was based on the highest value of the TF*IDF and direct two-way conversations metrics. The TF*IDF value determines the importance that a comment by a user has in a certain topic. A direct two-way conversation means that both players reply to each other in a topic. High value for direct two-way conversations indicates that a user has relatively more one-on-one conversations with other users on the forum.

–Failure is just practice for success.

Christopher Hitchens

5

Methodology

In this chapter we will discuss our approach to answering the four research questions. To validate our research we focus on centrality metrics and the distribution of these metrics. Besides this we also look at a rank-order correlation coefficient to compare different results.

5.1 Important Actors

Actors that play a significant role in a network supply human capital. Human capital consists of services that are of importance to the survival of the network [8]. For example moderating topics, recruiting new users, distributing resources or helping people with technical questions. To identify these users we will focus on the individual level of the largest connected component (see Section 2.3) of the *person-to-person* network. We will look at the following centrality measurements: degree, closeness, betweenness, eigenvector and PageRank (defined in Section 2.4). Each measurements will result in a ranking and we consider an actor significant if he/she scores higher than 3σ above the mean of all the users. We choose 3σ as our threshold to keep the resulting set a reasonable size to analyze. Our final set of important actors consists of the union of these five sets. Through this method we can gain more insight in the applicability of these metrics and more complex models could be trained in future work.

5.2 Comparing projections

In order to understand which type of projection is best suited to replicate the underlying social structure of the data we want to compare the impact of each projection method discussed in Section 2.2.1. The type of projection has significant influence on the metrics that we use to determine important actors because these metrics use weight as input. If we want to compare two graphs where the weights are computed with a different projection we should only look at metrics that use weight as a parameter such as closeness and betweenness centrality. Comparing other metrics such as the clustering coefficient would not make sense since they should be identical. In addition comparing the topology between different projections is not possible because the definition of weight is relatively defined in the system it is computed in. Rather we will focus on the distribution and ranking of metrics on an individual level since these measurements are comparable. First, we will look at the weighted degree distribution and average weighted degree connectivity. The weighted degree of a node v is the sum of edge weights adjacent to node v. A weighted degree distribution is a distribution of the frequency of each weighted degree k. The average weighted degree connectivity is the average nearest neighbor weighted degree of nodes with weighted degree k. By fitting a distribution we can compare the parameters to determine the effect of different weights on these distributions.

Second, we are interested in the different rankings as a result of each projection. We compute the Spearman rank-order correlation coefficient [21] for all users and the set of important actors defined in the previous section to see the difference between the effect of projection on both sets.

5.3 Filters

Filters such as removing a big linker (see Section 2.2.2) can reduce the number of edges and nodes in the unfiltered graph. One benefit of filtering graphs is the fact that computations are faster on smaller graphs. The main goal however is to reduce the number of edges in the graph that have no effect on the underlying social structure. We will focus on a filter that removes $1, 2, \ldots, i$ biggest linkers in the two-mode graph. To determine i we fit a power law function on the degree distribution of the $\top$ node set. Big linkers that are more than 1σ above the mean are filtered out. The threshold of 1σ is based on empirical findings (see Section 6.5), if we raise the threshold the number of big linkers is too small to make an analysis. After applying a filter it is possible that nodes get disconnected from the giant component and form a new connected component. By analyzing the size of each component we verify that the largest connected component is the only significant component and remove the additional components. To analyze the effect of a filter we

need a graph as baseline. Our baseline is an unfiltered one-mode graph that is constructed with the same method of projection. Then we can examine the topology measurements and ranking correlation of centrality metrics between each graph.

5.4 Groups

Another point of interest is identifying groups in the network. Thanks to the PIM research we can label users that are identified as having a certain role. This allows us to analyze the centrality measures between the different roles and users without a role in the network. This information could help us automatically classify certain roles based on their characteristics. These characteristics can give us more insights into how these users position themselves in the network and as a result be able to better detect them. We will try to rank these users based on centrality measurements defined in Section 5.1. In order to validate our ranking we will use the PIM ranking (see Section 4.2). The PIM ranking serves as an indicator of a correct ranking and we study different types of projection and filters to replicate this ranking with only the relational data. We will compute the Spearman rank-order correlation coefficient [21] between the PIM ranking and the ranking of each centrality metric defined in Section 5.1.

–No amount of experimentation can ever prove me right; a single experiment can prove me wrong.

Albert Einstein

6

Experiments

In the first experiment we will explore the two-mode network and study the properties of these two forums. In the second experiment we will go over the identified important actors that play a significant role in the network. The third experiment we examine the three different graphs that are a result of distinct types of projection. The fourth experiment focuses on the effect of reducing noise by removing the biggest linkers in the bipartite network. In the final experiment we evaluate the properties of users that have roles assigned in the PIM analysis.

6.1 Experimental setup

In order to complete our experiments we need to build a research pipeline. This pipeline allows us to parse, filter, project and compute metrics for multiple graphs in parallel (see Figure 6.1.1 for a conceptual diagram). Parallelism is accomplished by gathering all tasks and distributing these across processes. A task consists of a graph and the function with arguments we want to compute on this graph. Each result is stored in cache to prevent loss of data when the pipeline is interrupted during run time. This efficiency is greatly needed because of the size of each graph and the number of parameters we want to tune. We build a modular pipeline to allow customization of parameters. It is possible to change how we define a relationship, which filter or projection we apply and which metrics we compute. This can be configured by simply changing the environment variables of the pipeline which allows it to run

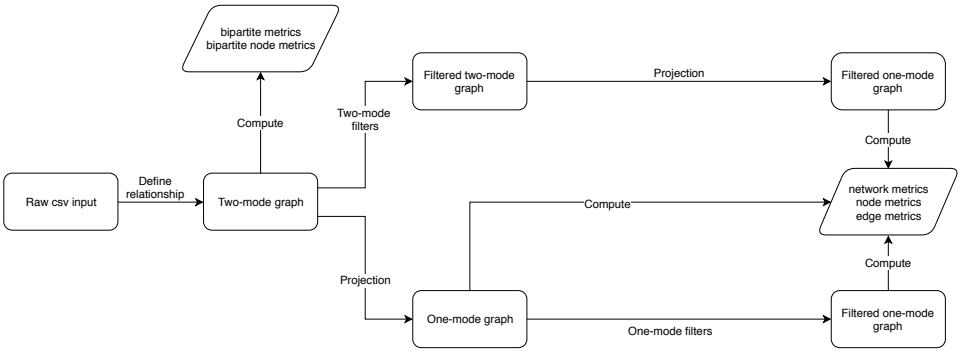

Figure 6.1.1: Pipeline diagram

on multiple machines that are linked to the same code base. A graph object can be one of the following three types: two-mode graph, one-mode graph and a filtered one-mode graph. We can define which metrics we want to compute for each type and compute these parallel. To parse the input to a graph object and project it to a one-mode network we use `NetworkX` [22]. Then to compute algorithms parallel we convert it to a graph object in `Graph-tool` [23] which utilizes Cython to boost performance. We use the default parameter for each algorithm as defined by `Graph-tool`. Implementation of all above was done in `Python` [24] and can be found on the public repository: `https://git.liacs.nl/s1437690/key_players`.

6.2 Exploring the two-mode network

Table 6.2.1 shows basic statistics about the topology of both two-mode networks. Most of the metrics seem quite similar but one of the interesting discrepancies is the relatively large number of top nodes ($n_\top$) and the high average degree for the bottom nodes ($k_\bot$) in data set A. In order to interpret this discrepancy we will look at the degree distribution and average degree connectivity. Figure 6.2.1 shows the degree distribution which is the number of nodes with the same degree k for each value of k. In order to study the decline in number of nodes with the same degree k we will fit a power law (see Equation 2.3.1) function on this distribution. The top nodes of data set A have a higher value of γ than in data set B. This implies that there is more emphasize on smaller topics in data set A. The bottom nodes show the reverse since data set A has a lower value of γ than in data set B. This indicates that a few users in data set A comment on almost all topics, which explains the value of average degree $k_\bot = 21.1$.

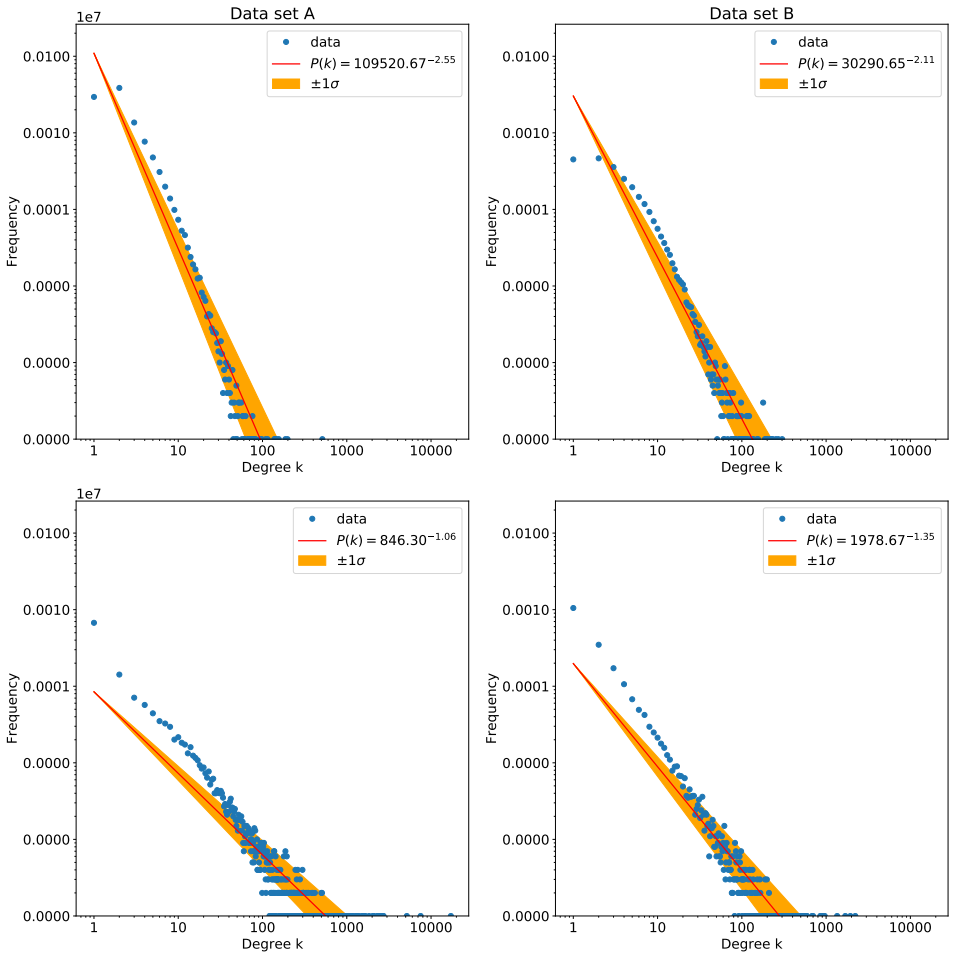

Figure 6.2.1: Degree distribution of the two forums. The first row shows the top nodes and the second row shows the bottom nodes.

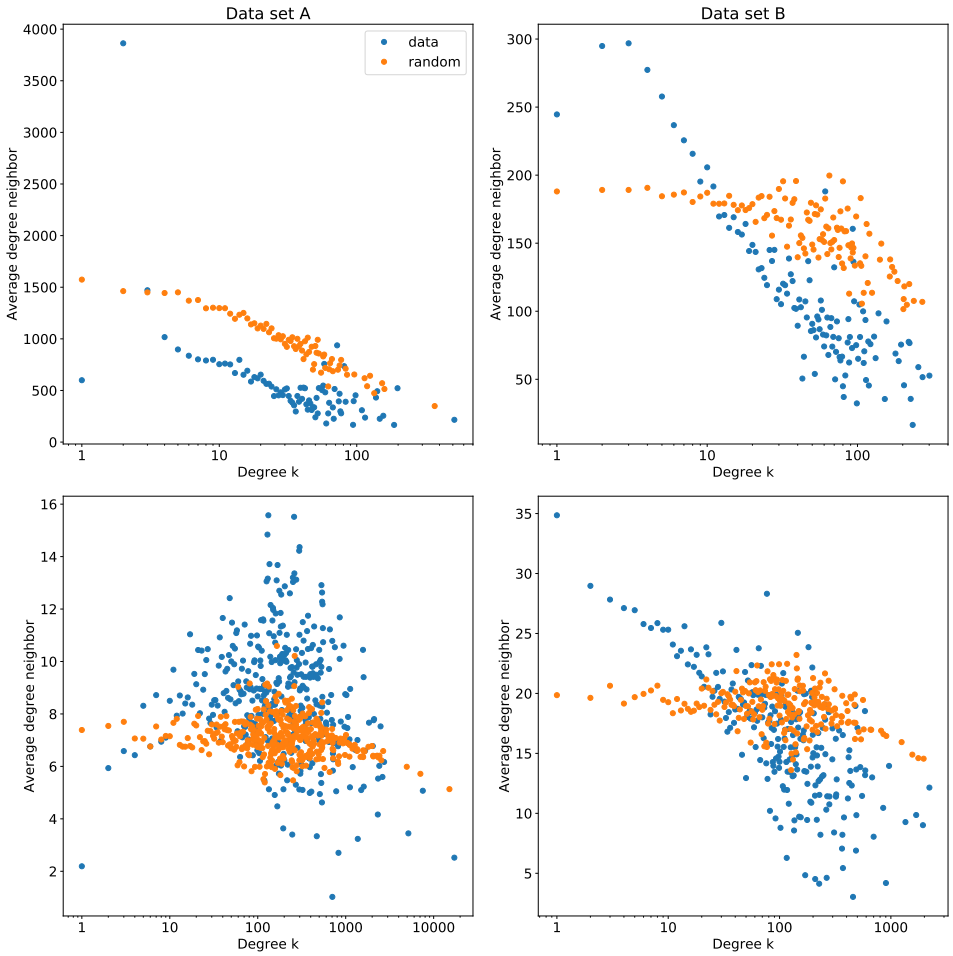

Figure 6.2.2: Average degree connectivity of the two forums. The first row shows the top nodes and the second row shows the bottom nodes.

Figure 6.2.2 shows the average degree connectivity, which is the average nearest neighbor degree of nodes with degree k. This highlights the correlation between degree and the degree of a node's neighbors. The top nodes display a negative correlation which implies that larger topics are commented on by users whom, on average, are less active. Which could indicate that larger topics include, based on degree, relatively fewer people that may be of significance for the network. However the bottom nodes do not show this clear negative correlation. It does seem that the most active users comment on topics with relatively few comments. Another interesting matter is the average degree connectivity of the top nodes in data set A. This shows that topics with two comments are on average commented on by users with a degree of almost 4000. Though when we look at the degree distribution in Figure 6.2.1 there are only a handful of users with a higher degree than 4000. Through qualitative analysis with specialists on this case we concluded

that this was due to the fact that new users in data set A had to make an application to gain access. When a new user wants to join this forum he has to provide new content. Such an application is approved by the main admin and then closed. The main admin was later assisted by a few other admins in this process. This causes the main admin to have a degree of over $10,000$ and the spike in the top left plot of Figure 6.2.2 at degree $k = 2$. When we look at the bottom left plot we can see the same phenomenon occur at degree $k = 1$ with an average degree connectivity of 2, whereas data set B has an average degree connectivity value of 35. This implies that users who commented once did this on a topic with only 2 comments. This could be because most of the applicants are lurkers and their only comment is on the application topic with their own and the admins comment.

	Data set A	Data set B
n	119742	46313
$n_\top$	105083	25056
$n_\bot$	14659	21257
m	309716	145086
δ	0.000201	0.000272
$\langle k \rangle$	5.2	6.3
$k_\top$	3.0	5.8
$k_\bot$	21.1	6.8
$cc_N(G)$	0.239	0.153
$cc_\bullet(\top)$	0.271	0.158
$cc_\bullet(\bot)$	0.009	0.146
$cc_{\underline{\bullet}}(\top)$	0.608	0.465
$cc_{\underline{\bullet}}(\bot)$	0.598	0.702
$cc_{\bar{\bullet}}(\top)$	0.324	0.189
$cc_{\bar{\bullet}}(\bot)$	0.011	0.154

Table 6.2.1: Bipartite statistics about our two bipartite graphs.

6.3 Important actors

Table 6.3.1 and 6.3.2 shows the number of key players for each type of projection and the percentage of each role within these key players for data set A and B. It also shows the percentage of verified admins that were discovered. In data set A we identified 43, 56 and 56 users as key players with the unweighted, weighted and newman collaboration projection respectively with the method specified in Section 5.1. The police provided us with a complete list of admins on this forum that were verified after a case. A total of 16 users were verified to have an admin status and of these 16 we found 12 (75%) with the unweighted projection and 13 (81.25%) with the weighted & Newman collaboration projection. We did the same experiment for data

	Unweighted	Weighted	Newman
Total	43	56	56
Managers	48.8%	50%	51.8%
Abusers	30.3%	25%	21.3%
Technical	6.9%	9%	9%
Embedded	14%	16%	17.9%
Discovered admins	75%	81.25%	81.25%

Table 6.3.1: Identified key players in data set A

	Unweighted	Weighted	Newman
Total	57	66	75
Managers	42.1%	42.4%	49.3%
Abusers	35.1%	30.3%	28%
Technical	12.3%	12.1%	10.7%
Embedded	10.5%	15.2%	12%
Discovered admins	43%	43%	43%

Table 6.3.2: Identified key players in data set B

set B and identified 57, 66 and 75 users as key players with the unweighted, weighted and newman collaboration projection respectively as important actors. In this case the police provided us with a list of known and suspected admins so this list was incomplete. Nevertheless of these admins we found 14 (43%) admins with each form of projection. We also classified around 6-10 actors in both data sets as important even though they were not labeled by the PIM analysis. These users could have been dropped of the list by the PIM analysis because they did not use relevant vocabulary.

6.4 Projection

We applied three different types of projection on two data sets which results in six different networks. Table 6.4.1 shows the network metrics of each individual graph. The degree assortativity coefficient increases in both data sets with weighted projection and Newman collaboration projection. An increase in r (see Equation 2.3.4) indicates that the variance between the weighted degree of neighbors decreases. The weighted degree of a node is now scaled by the strength of its relationships and this could imply that using weights better encapsulates the phenomena that people connect with people who are also well connected.

To get a better understanding of the effect on weighted degree of a node we plot the weighted degree distribution (see Figure 6.4.1). We can compare these distributions by fitting a power law function (see Equation 2.3.1). In

both data sets it shows that Newman collaboration model has the highest value of γ. Which implies that there are relatively less nodes with a higher weighted degree and therefore these node become relatively more important. An interesting aspect of the Newman collaboration model is that the weighted degree of a person is equal to the number of topics with two or more neighbors he has commented on. Therefore it is almost similar to the degree distribution in the bottom plots in Figure 6.2.1 and would be exactly similar if we excluded topics with an degree of one in these plots.

However this does not show us why r is increasing. We want to examine this by plotting the average weighted degree connectivity (see Figure 6.4.2). The left plot of both data sets shows a negative correlation which implies that the more people you are connected with the less well connected your neighbors are. However when we add weights to the edges this negative correlation seems to disappear. This explains the increasing r because it emphasizes if you are talking to other important actors. Through discussions with domain experts we determined that well connected actors are more inclined to chat with other well connected actors. This would suggest that adding weights to the edges gives a better representation of the underlying social structure and connectivity of actors.

The last subject matter we want to observe is the rank-order of individuals based on centrality metrics. We will compare the rank-order of all actors, and only the 99.8 percentile (which equals those 3σ above the mean, see Section 5.1). Figure 6.4.3 and 6.4.4 shows all actors and we can clearly see that Newman collaboration projection is the most disparate from unweighted projection for each metric. Figure 6.4.5 and 6.4.6 zoom in on the 99.8 percentile and shows that the closeness centrality in data set B increases the most. This implies that most of the embedded network was effected by the change of projection. PageRank appears to be the least affected by changing the form of projection.

6.5 Network derivation

Through removing the biggest linkers in the bipartite network we can reduce the number of edges that are created in the projected graph and accordingly reduce clutter. As we discussed in Section 5.3 we want to remove big linkers that are 1σ above the mean. When we look at the top plots in Figure 6.2.1 it shows that the top four and five topics in data set A and data set B respectively will be removed. Since we are comparing these to an unfiltered baseline graph and also want to look at the three different types of projection this results in 33 unique graphs. Table 6.5.1 and 6.5.2 present the resulting network metrics of each graph. Data set A loses $\frac{427097}{566834} \approx 25\%$ of its edges by removing only four big linkers. While data set B drops $\frac{996224}{1150375} \approx 15\%$ of its edges by removing five big linkers. It is interesting to note that the average

	Data set A			Data set B		
	Unweighted projection	Weighted projection	Newman collaboration projection	Unweighted projection	Weighted projection	Newman collaboration projection
n	14659	14659	14659	21280	21280	21280
m	566834	566834	566834	1150375	1150375	1150375
δ	0.005276	0.005276	0.005276	0.005081	0.005081	0.005081
$\langle k \rangle$	77.3	143.4	19.1	108.1	132.7	6.6
$\langle d \rangle$	2.3	1.5	2.0	2.5	1.6	39.4
d_{max}	5	3.01	567	6	4.2	570
r	-0.137	-0.123	-0.038	-0.109	-0.076	-0.049
c	0.321	0.321	0.321	0.218	0.218	0.218

Table 6.4.1: Topology measurements of the projected graphs

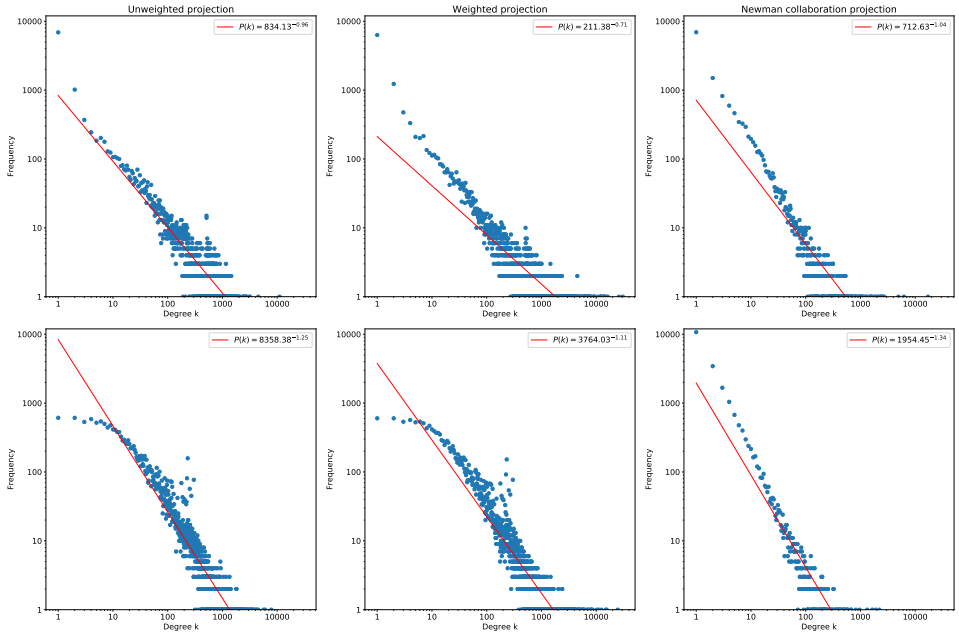

Figure 6.4.1: Weighted degree distribution of the three projected graphs. The first row shows data set A and the second row shows data set B.

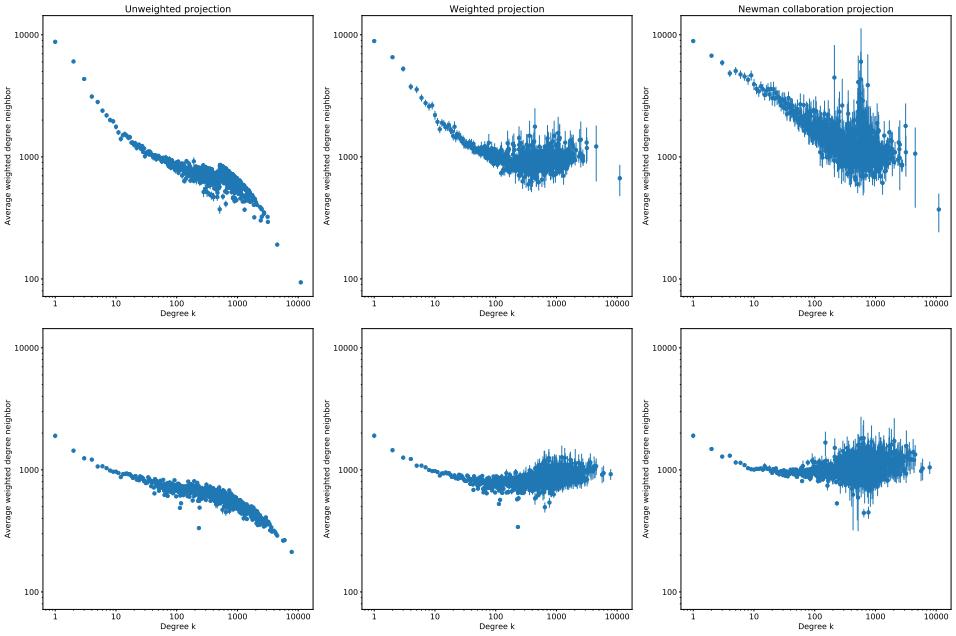

Figure 6.4.2: Average weighted degree connectivity of the three projected graphs. The first row shows data set A and the second row shows data set B.

Figure 6.4.3: Spearman rank-order correlation of centrality metrics with different projections in data set A.

Figure 6.4.4: Spearman rank-order correlation of centrality metrics with different projections in data set B.

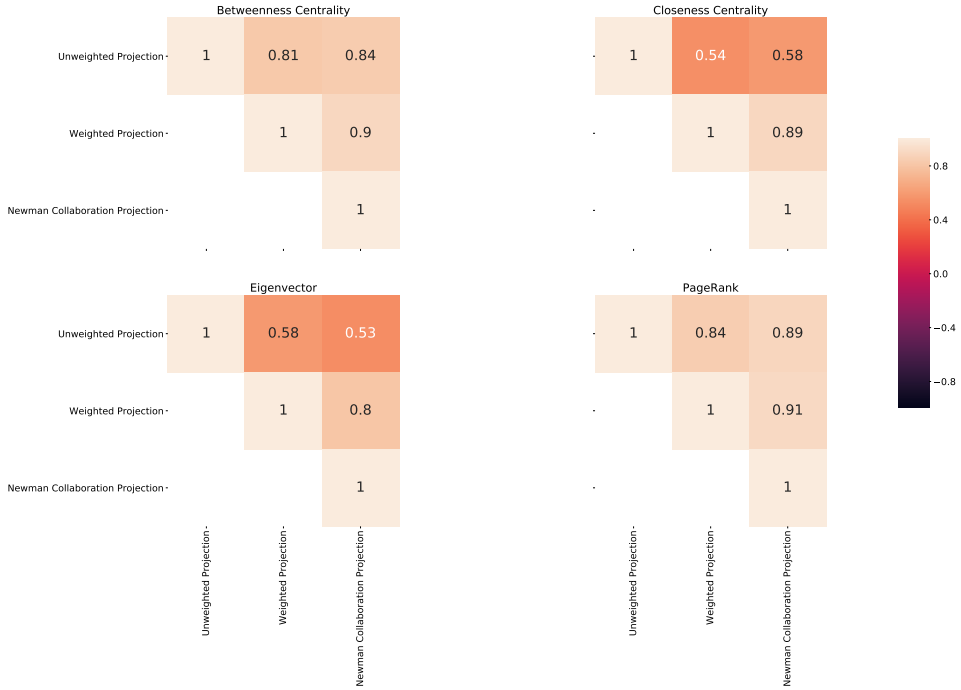

Figure 6.4.5: Spearman rank-order correlation of centrality metrics of the users in the top 99.8 percentile with different projections in data set A.

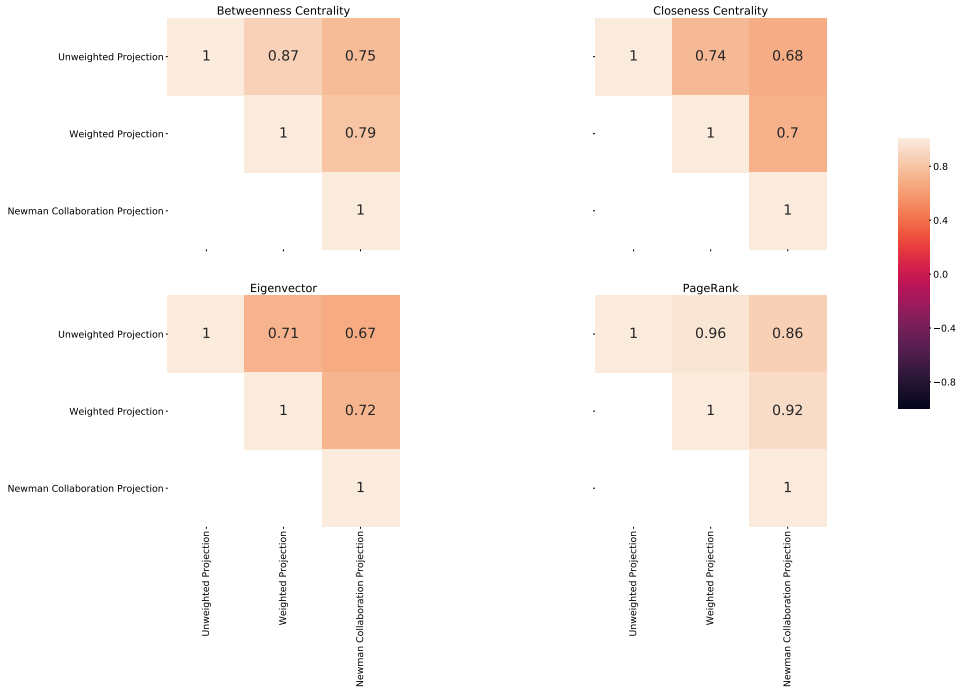

Figure 6.4.6: Spearman rank-order correlation of centrality metrics of the users in the top 99.8 percentile with different projections in data set B.

	Filter	n	m	δ	$\langle k \rangle$	$\langle d \rangle$	d_{max}	r	c
Unweighted projection	Baseline Graph	14659	566834	0.005276	77.3	2.3	5.00	-0.137	0.321
	Filtered 1 big linker	14658	461452	0.004296	63.0	2.3	5.00	-0.128	0.211
	Filtered 2 big linker	14658	450607	0.004195	61.5	2.3	5.00	-0.125	0.203
	Filtered 3 big linker	14658	435776	0.004057	59.5	2.3	5.00	-0.125	0.196
	Filtered 4 big linker	14658	427097	0.003976	58.3	2.3	5.00	-0.124	0.191
Weighted projection	Baseline Graph	14659	566834	0.005276	143.4	1.5	3.01	-0.123	0.321
	Filtered 1 big linker	14658	461452	0.004296	125.5	1.5	3.01	-0.127	0.211
	Filtered 2 big linker	14658	450607	0.004195	122.8	1.5	3.01	-0.126	0.203
	Filtered 3 big linker	14658	435776	0.004057	120.5	1.5	3.01	-0.128	0.196
	Filtered 4 big linker	14658	427097	0.003976	118.8	1.5	3.01	-0.128	0.191
Newman collaboration projection	Baseline Graph	14659	566834	0.005276	19.1	2.0	567.00	-0.038	0.321
	Filtered 1 big linker	14658	461452	0.004296	19.1	1.9	91.00	-0.044	0.211
	Filtered 2 big linker	14658	450607	0.004195	19.1	1.9	91.00	-0.044	0.203
	Filtered 3 big linker	14658	435776	0.004057	19.1	1.9	91.00	-0.045	0.196
	Filtered 4 big linker	14658	427097	0.003976	19.0	1.9	91.00	-0.046	0.191

Table 6.5.1: Topology measurements of filtered graph in data set A

shortest distance and the diameter does not change in the unweighted and weighted projection in both data sets. This could imply that the removed edges were irrelevant for the information flow of the network.

To further study the insignificance of these removed edges we will again look at the rank-order coefficient across all graphs. Figure A.0.1 till A.0.6 shows rank-order coefficient between the 33 graphs for all users. Notice that the lowest value is 0.95 which is still high. Figure A.0.7 till A.0.12 does the same but only for the users in the 99.8 percentile. It is interesting to note that using either weighted or Newman collaboration projected has an increase in the coefficient to 1. Which means that removing $\approx 25\%$ and $\approx 15\%$ edges in data set A and data set B respectively did not have any effect on the rank-order of the users in the 99.8 percentile since the correlation coefficient stays 1. This results is faster computations and less clutter in the visualization by removing the redundant edges while maintaining the same outcome.

6.6 Role and ranking analysis

The goal of the role analysis experiment is to find certain characteristics that define a role. We will first look at the differences between groups based on the distribution of centrality metrics in Figure 6.6.1 and 6.6.2. In these plots we have decided to exclude outliers in the visualization because we are interested in the general characteristics. We observe that the technical people seem to score lower in betweenness, degree, eigenvector centrality and PageRank in both data sets. Through discussion with domain experts we concluded that this could be due to the fact that technical users tend to be more individualistic on the forums. They also tend to focus more on building applications and mainly talk within their group. Managers on the other hand have to communicate with newbies about questions, new rules

	Filter	n	m	δ	$\langle k \rangle$	$\langle d \rangle$	d_{max}	r	c
Unweighted projection	Baseline Graph	21280	1150375	0.005081	108.1	2.5	6.00	-0.109	0.218
	Filtered 1 big linker	21212	1109150	0.004930	104.6	2.5	6.00	-0.108	0.209
	Filtered 2 big linker	21174	1076255	0.004801	101.7	2.6	6.00	-0.108	0.202
	Filtered 3 big linker	21128	1047211	0.004692	99.1	2.6	6.00	-0.108	0.196
	Filtered 4 big linker	20987	1020382	0.004634	97.2	2.6	6.00	-0.112	0.189
	Filtered 5 big linker	20915	996224	0.004555	95.3	2.6	6.00	-0.113	0.184
Weighted projection	Baseline Graph	21280	1150375	0.005081	132.7	1.6	4.17	-0.076	0.218
	Filtered 1 big linker	21212	1109150	0.004930	128.8	1.6	4.17	-0.075	0.209
	Filtered 2 big linker	21174	1076255	0.004801	125.6	1.6	4.17	-0.075	0.202
	Filtered 3 big linker	21128	1047211	0.004692	122.8	1.6	4.17	-0.075	0.196
	Filtered 4 big linker	20987	1020382	0.004634	121.1	1.6	4.27	-0.078	0.189
	Filtered 5 big linker	20915	996224	0.004555	119.1	1.6	4.27	-0.079	0.184
Newman collaboration projection	Baseline Graph	21280	1150375	0.005081	6.6	39.4	570.00	-0.049	0.218
	Filtered 1 big linker	21212	1109150	0.004930	6.6	37.6	523.00	-0.050	0.209
	Filtered 2 big linker	21174	1076255	0.004801	6.6	36.7	485.02	-0.051	0.202
	Filtered 3 big linker	21128	1047211	0.004692	6.6	35.7	457.00	-0.051	0.196
	Filtered 4 big linker	20987	1020382	0.004634	6.6	32.8	446.00	-0.053	0.189
	Filtered 5 big linker	20915	996224	0.004555	6.7	31.4	439.00	-0.055	0.184

Table 6.5.2: Topology measurements of filtered graphs in data set B

and advertise about other forums. In data set A this is highlighted by their higher score for eigenvector and PageRank because this means that they are important for their neighbors. To further support this claim we looked at the set of neighbors of nodes in a certain role. Table 6.6.1 shows the percentage of nodes in the neighbor set relative to the total number of nodes in the network. The neighbor set of the managers seems to be the largest while the neighbor set of the technical group seems to be the smallest in both data sets. Lastly, the abusers are fanatic about their work and like to share their experience with the community. They also influence others to do the same. This could be indicated by that they seem to have the same degree centrality as the managers.

Another point of interest was finding a configuration that had the most resemblance to the ranking of the PIM analysis. Figure 6.6.3 and 6.6.4 show the rank-order correlation coefficient between the PIM ranking and the different centrality measures. In data set A when we use weighted or Newman collaboration projection this has a positive effect on the rank-order coefficient for the metrics: betweenness centrality, eigenvector and PageRank. We can get up to a rank-order correlation of 0.79 when we use closeness centrality with a weighted projection. In data set B this positive effect only counts for weighted projection. In Section 6.5 we concluded that big linker filters have no effect on the rank-order so this means that removing topics 1σ above the mean had no positive or negative effect on the correlation with the PIM ranking.

	Data set A	Data set B
Managers (MA)	99.3%	93%
Abusers (AB)	38.95%	87.98%
Technical (TE)	34.41%	78.79%

Table 6.6.1: Percentage of nodes in the set of neighbors of nodes in a role.

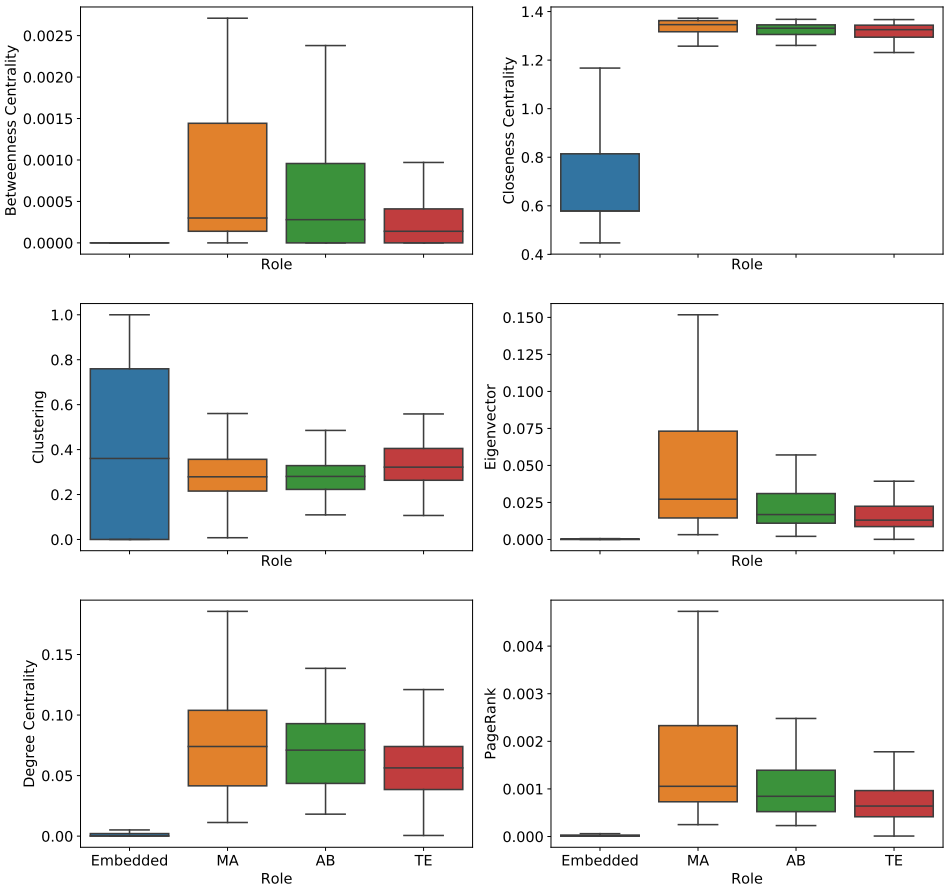

Figure 6.6.1: Distribution of node metrics between different groups in data set A.

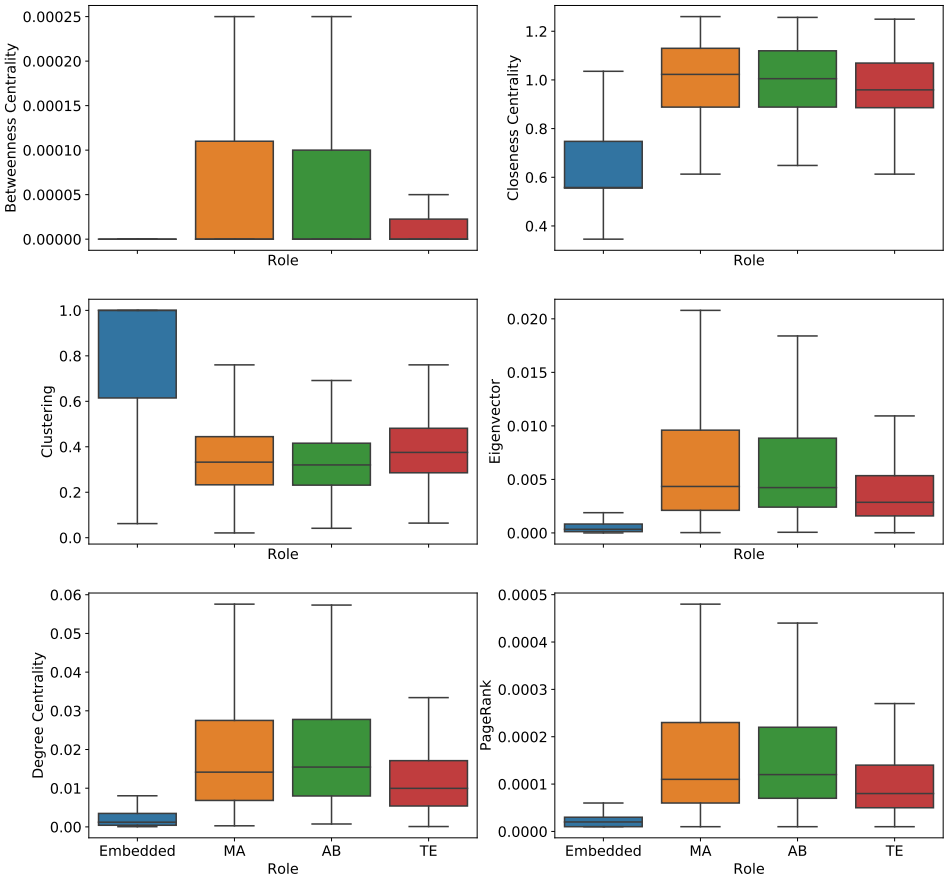

Figure 6.6.2: Distribution of node metrics between different groups in data set B.

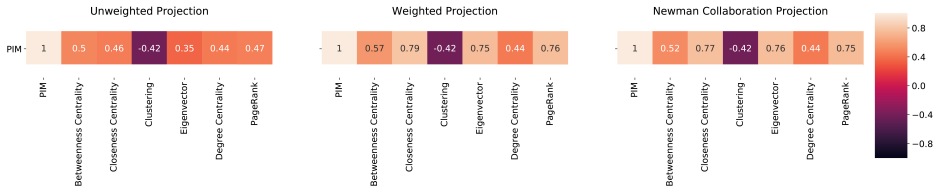

Figure 6.6.3: Spearman rank-order correlation between PIM and node metrics ranking of the baseline graph in data set A.

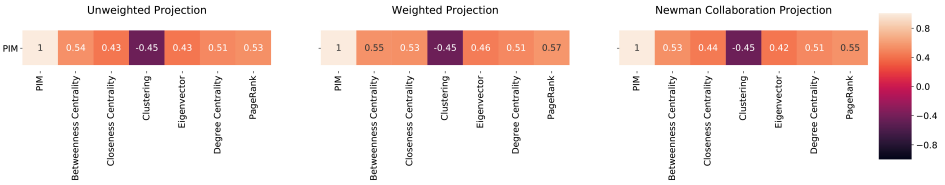

Figure 6.6.4: Spearman rank-order correlation between PIM and node metrics of the baseline graph in data set B.

6.7 Discussion

In Section 6.2 we examined that larger topics are commented on by relatively less active users. This could imply that larger topics include, based on degree, relatively fewer people that are of significance for the network. Which could be due to new users finding big topics more easily. We also noticed that topics with few comments are commented on by relatively more active users. Another interesting finding was that we found a forum rule in the average degree connectivity of data set A. It shows the rule that users had to supply content before getting accepted to the forum by one of the admins. These three findings can help identify forum behaviour and rules with only the relational data of messages.

Section 6.3 shows that in data set A 81.25% and in data set B 43% of the verified admins were detected by centrality metrics. They were discovered in the set of users discussed in Section 5.1. Admins are active users since they communicate with new users about questions, discuss new rules and advertise for other forums. Due to these activities they score high in our metrics and are important to the network.

It is important that the *person-to-person* network still represents the underlying social structure after the projection. Therefore in Section 6.4 and 6.5 we examined different types of projection and a filter that have an impact on the *person-to-person* network. We concluded that adding weight to the edges was beneficial because it captures the phenomena that well connected actors communicate with other important actors. We also found that filtering big linkers that are 1σ above the mean resulted in 25% and 15% less edges in data set A and B respectively. Even though we removed these

edges the rank-order correlation coefficient stayed close to 1 which implies that we did not lose information. Removing these edges further benefits the computational speed of algorithms and reduces the clutter of the *person-to-person* network.

Lastly, we explored the three groups defined by the PIM analysis. We discovered that managers are most connected while the technical users are the least connected of the groups (see Table 6.6.1). This could validate the empirical findings of the domain experts that managers focus more on interacting with the community while technical users concentrate on building applications and specific technical questions. During our experiment we also examined which parameters result in the highest rank-order correlation coefficient between the PIM ranking and centrality measurement rankings (see Figure 6.6.3 and 6.6.4). Using weighted projection and closeness centrality results in a 79% correlation coefficient in data set A. In data set B weighted projection and PageRank results in a 57% correlation coefficient. Therefore we conclude that weighted projection is best suited to represent the underlying social structure.

It is also important to note that this approach does not need text and can use relational data of messages and thus also messages of an unknown language as input.

–Use what talents you possess; the woods would be very silent if no birds sang there except those that sang best.

Henry van Dyke

7

Conclusion and Future Work

In this thesis we studied two online child exploitation forums and observed certain forum rules and phenomena in the data. We identified 81.25% and 43% of the verified admins in data set A and data set B respectively while limiting the scope of important actors to 0.2% of the total users. Besides this we researched three types of projection that transform our bipartite network to an one-mode network. We conclude that a weighted projection is best suited to capture the underlying social structure. Another point of research regarding projections was the effect of filters to reduce the clutter. We removed four and five big linkers which removed 25% and 15% of the edges in data set A and data set B respectively. To our surprise this had no effect on the rank-order of the top 99.8 percentile of users based on centrality metrics. Finally we evaluated the groups that were labelled by the PIM analysis. Technical users scored lower on centrality metrics and are the least connected group presumably because they are more individualistic and focus on their own group. We also noticed that using a weighted projection resulted in the highest rank-order correlation with the ranking that the PIM analysis provided. The advantages of using a social network approach is that it can use the relational data of encrypted messages and messages of different languages. This implies that we can apply this technique to a foreign discussion forum on the dark net and gain insights into the organization without understanding the messages.

One of the biggest criticisms of applied network science is that it is too slow because gathering data and performing a qualitative analysis takes time. Therefore we deem it necessary to look into ways to make the network

analyzable through time. Which means that the network has to represent the current underlying social structure at each time unit. This could be achieved by making the weights of the edges dependent on time. Such a model would determine how edges lose weight over time. It is also vital to remember that a data-driven approach can be as good as the data presented. Unfortunately a person can have multiple usernames which impairs the models. Finding ways to connect two or more usernames to the same person would advance any data-driven approach. Finally, since users are often on multiple forums we consider research into ways to connect multiple child exploitation networks together into one big ecosysteem interesting. For future work we suggest studying the private messages between users in order to validate our findings with the actual underlying social structure. We also propose to train a model that predicts the role of a user through supervised learning with the centrality measurements as features and PIM labels as target.

A

Filter correlation coefficients

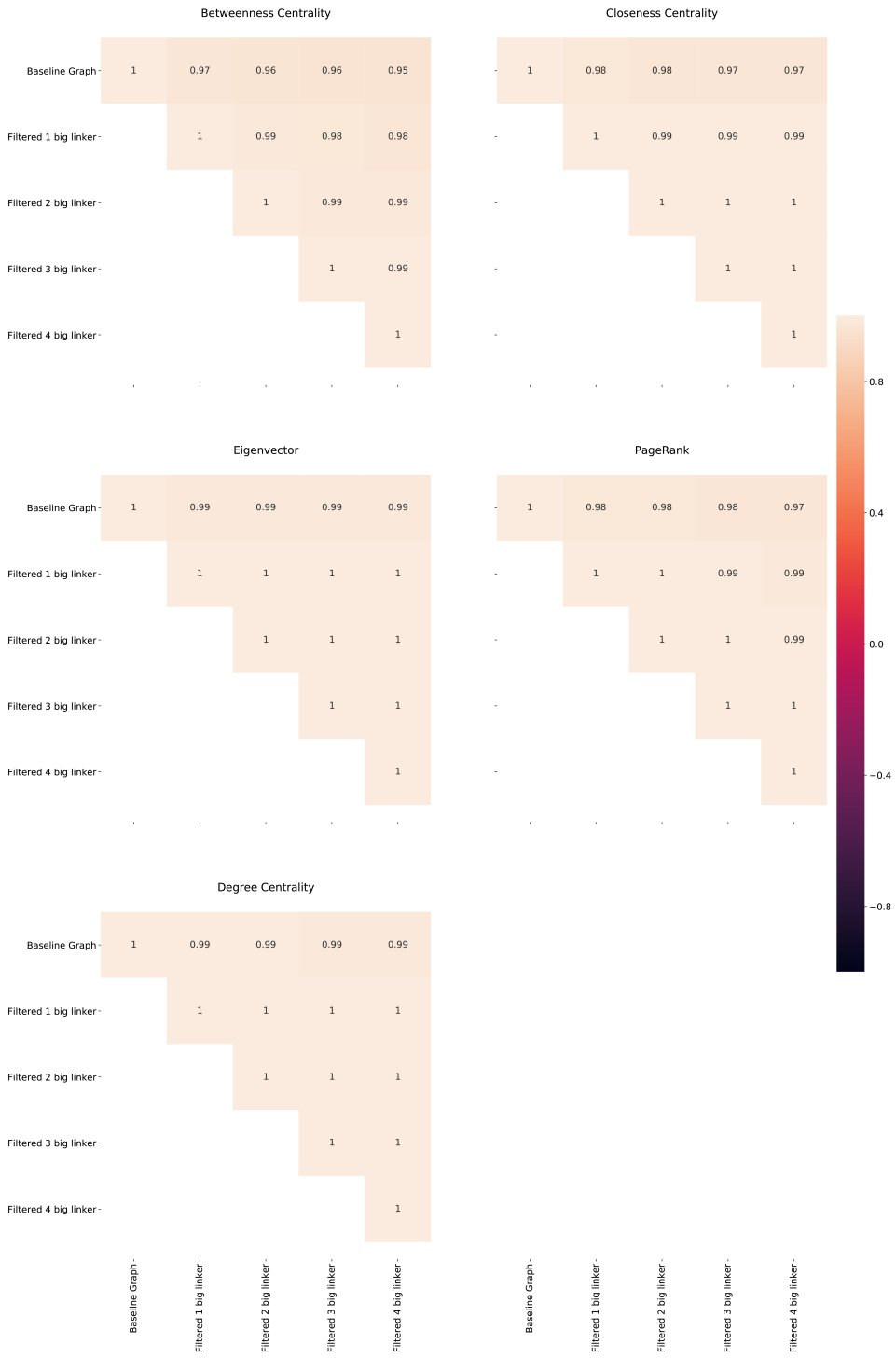

Figure A.0.1: Spearman rank-order correlation of centrality metrics with different filters on the unweighted projection in data set A.

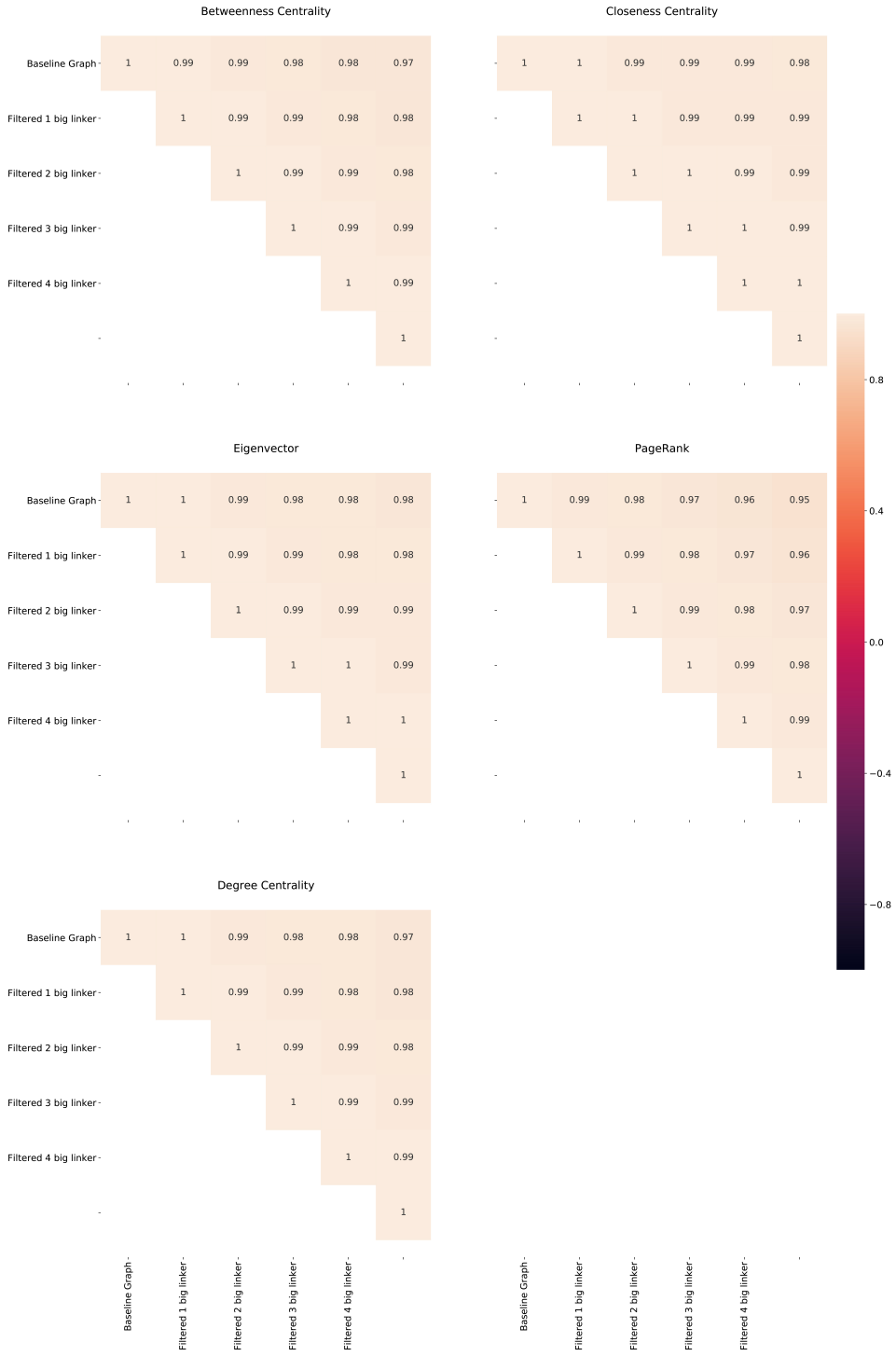

Figure A.0.2: Spearman rank-order correlation of centrality metrics with different filters on the unweighted projection in data set B.

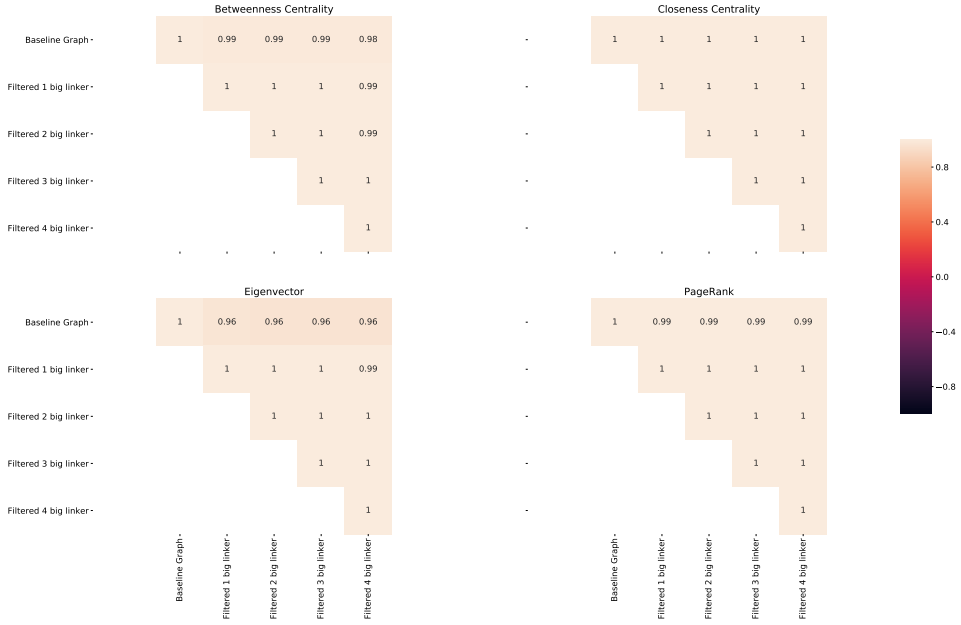

Figure A.0.3: Spearman rank-order correlation of centrality metrics with different filters on the weighted projection in data set A.

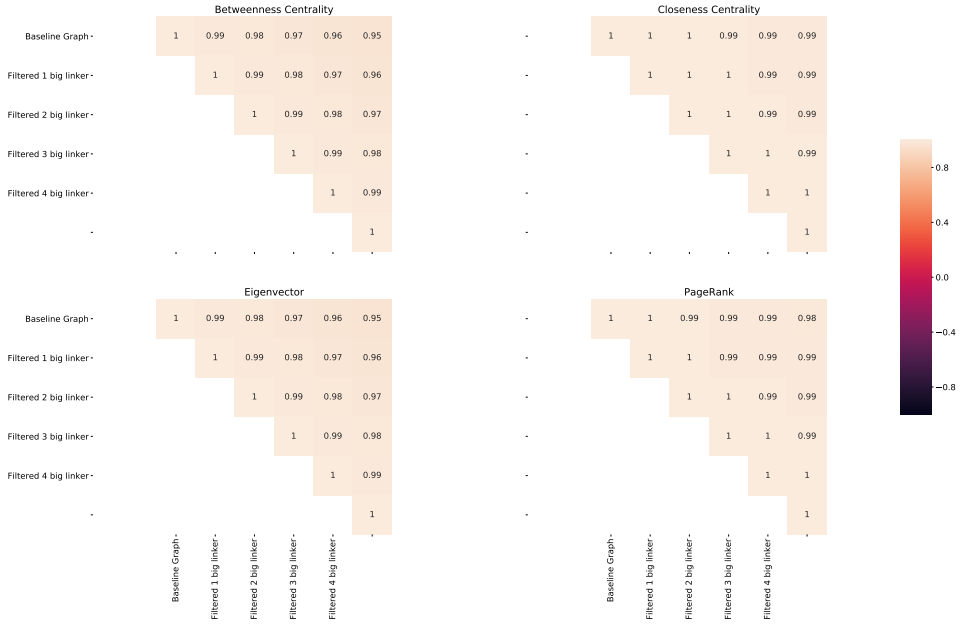

Figure A.0.4: Spearman rank-order correlation of centrality metrics with different filters on the weighted projection in data set B.

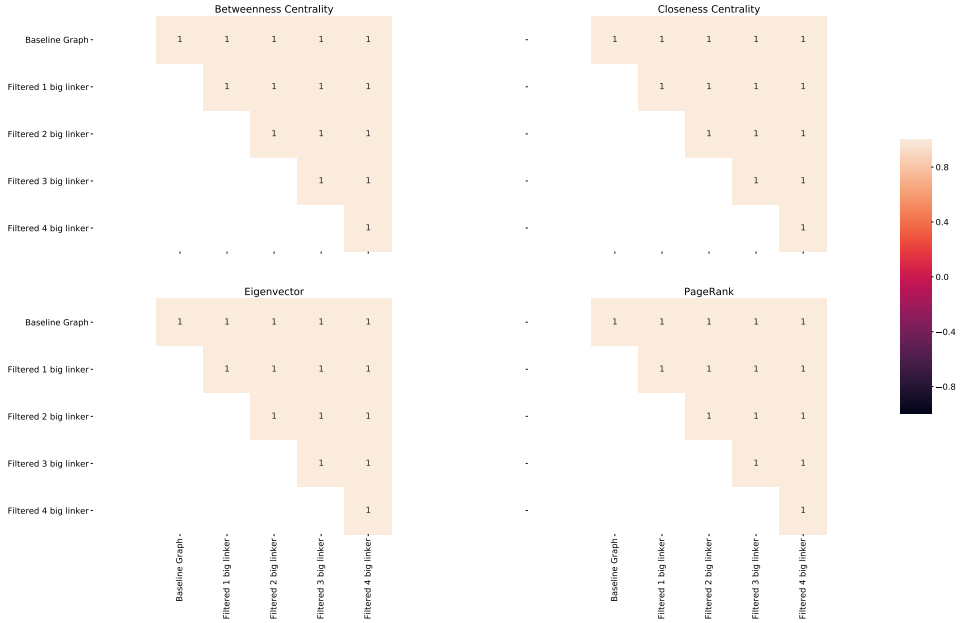

Figure A.0.5: Spearman rank-order correlation of centrality metrics with different filters on the newman collaboration projection in data set A.

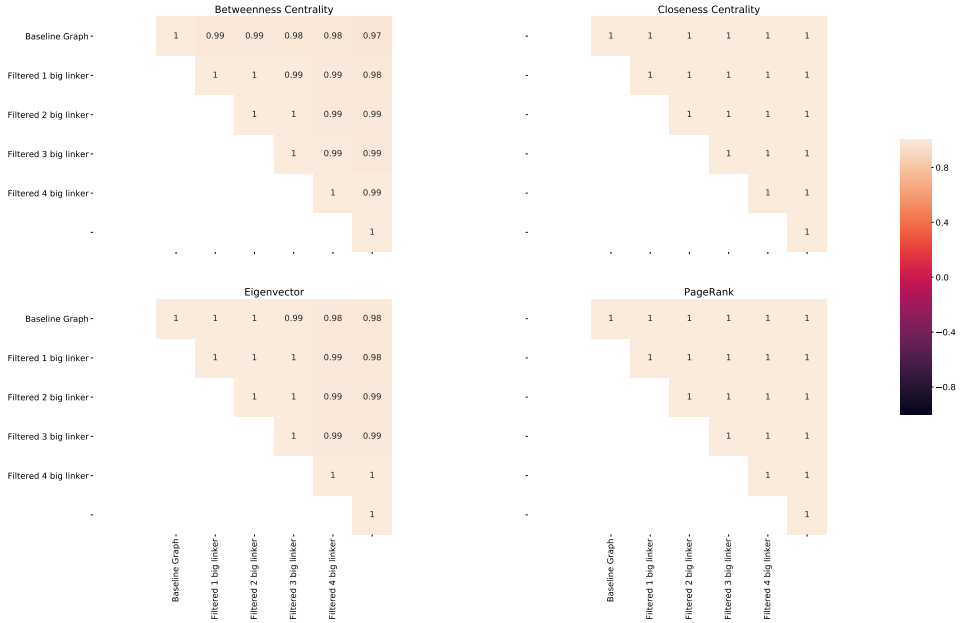

Figure A.0.6: Spearman rank-order correlation of centrality metrics with different filters on the newman collaboration projection in data set B.

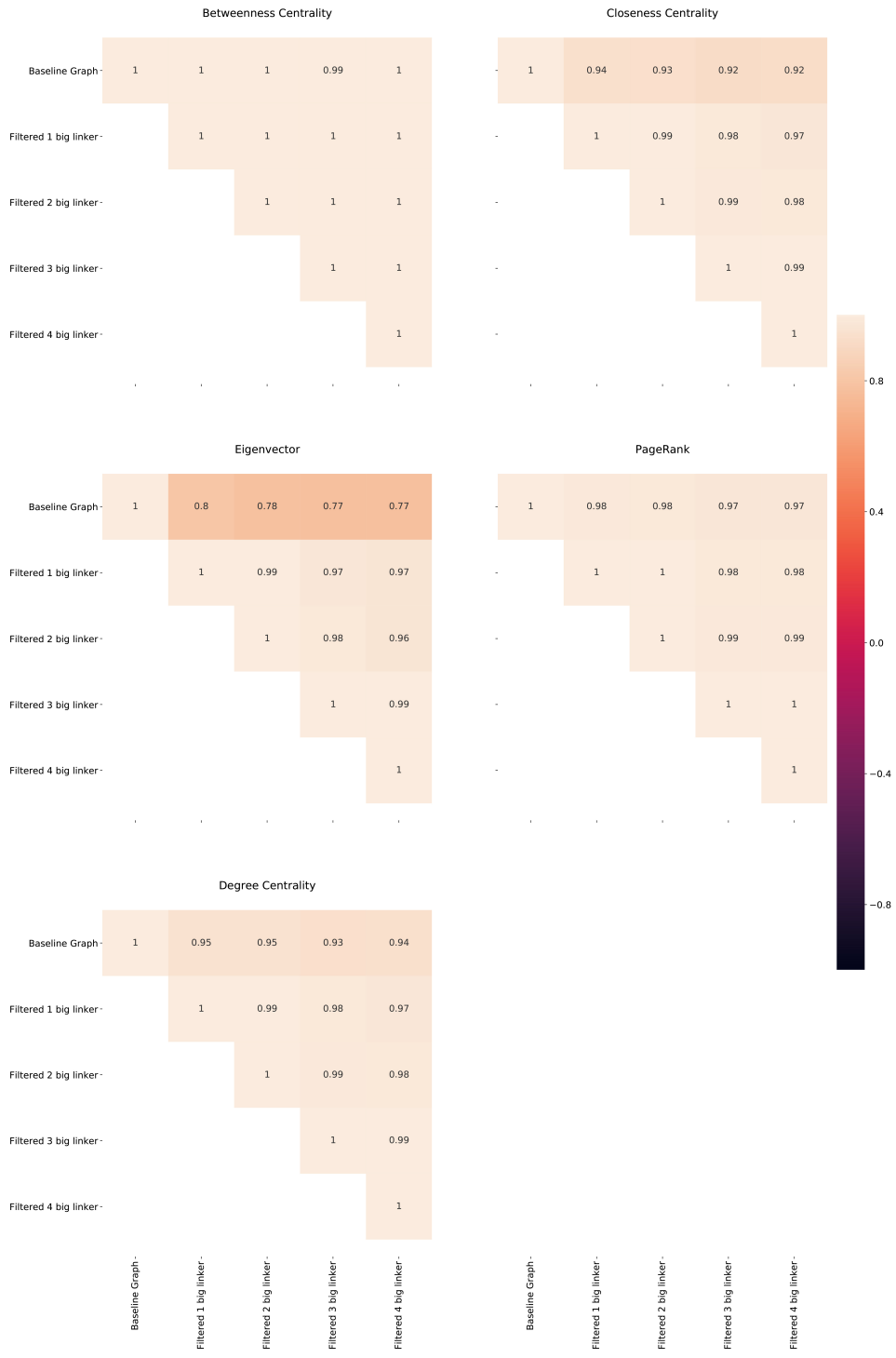

Figure A.0.7: Spearman rank-order correlation of centrality metrics of the users in the top 99.8 percentile with different filters on the unweighted projection in data set A.

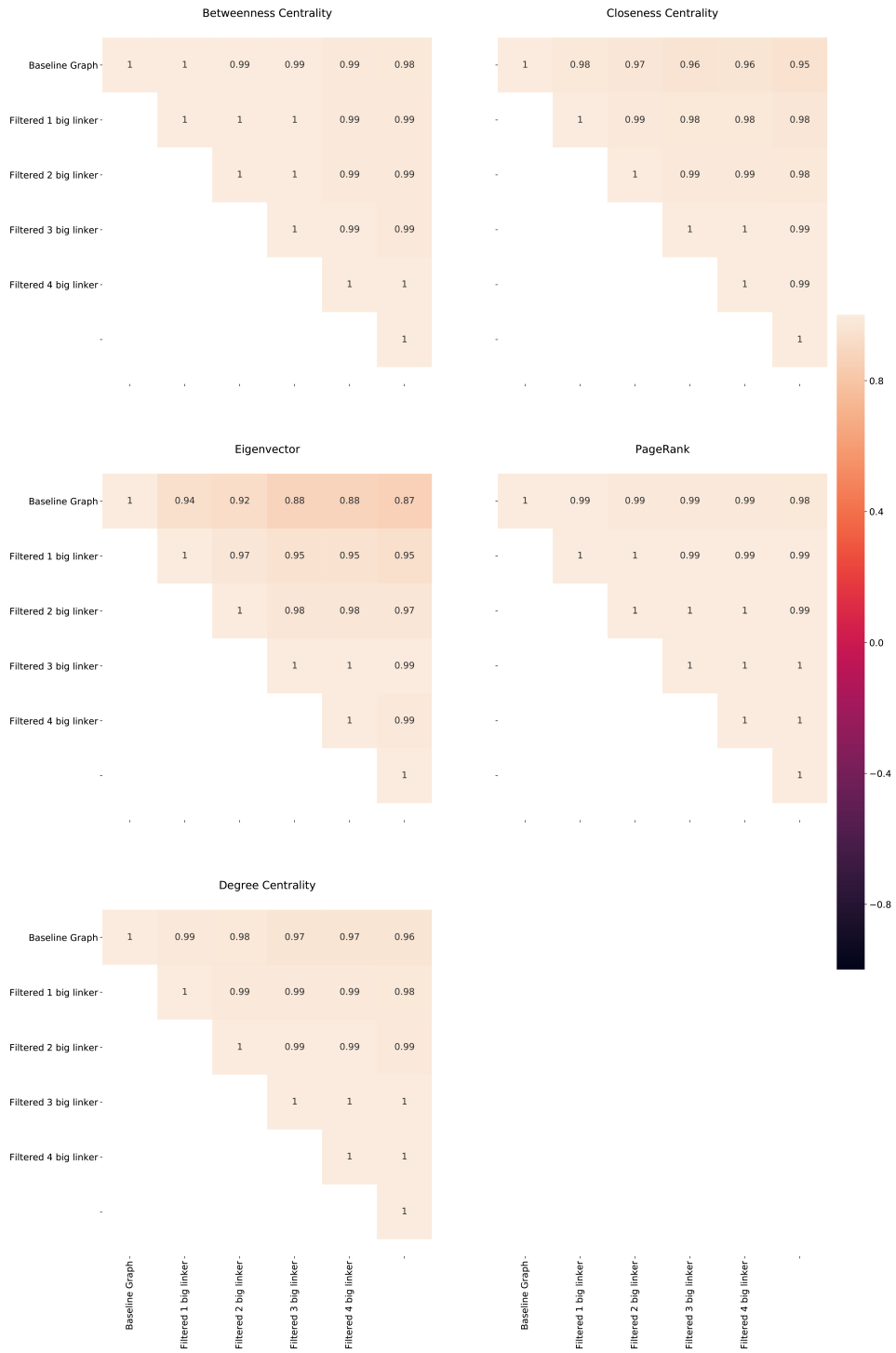

Figure A.0.8: Spearman rank-order correlation of centrality metrics of the users in the top 99.8 percentile with different filters on the unweighted projection in data set B.

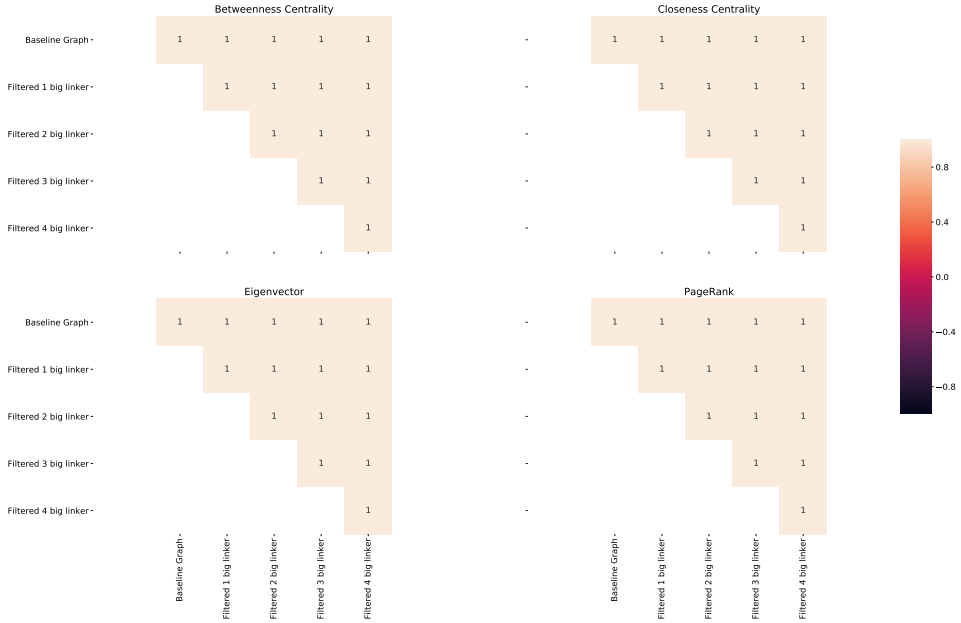

Figure A.0.9: Spearman rank-order correlation of centrality metrics of the users in the top 99.8 percentile with different filters on the weighted projection in data set A.

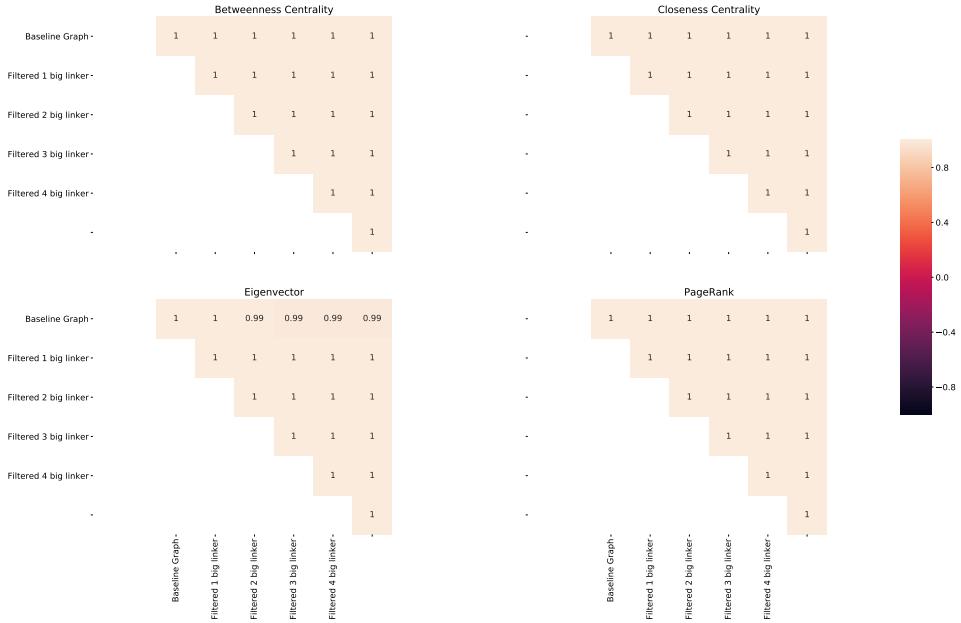

Figure A.0.10: Spearman rank-order correlation of centrality metrics of the users in the top 99.8 percentile with different filters on the weighted projection in data set B.

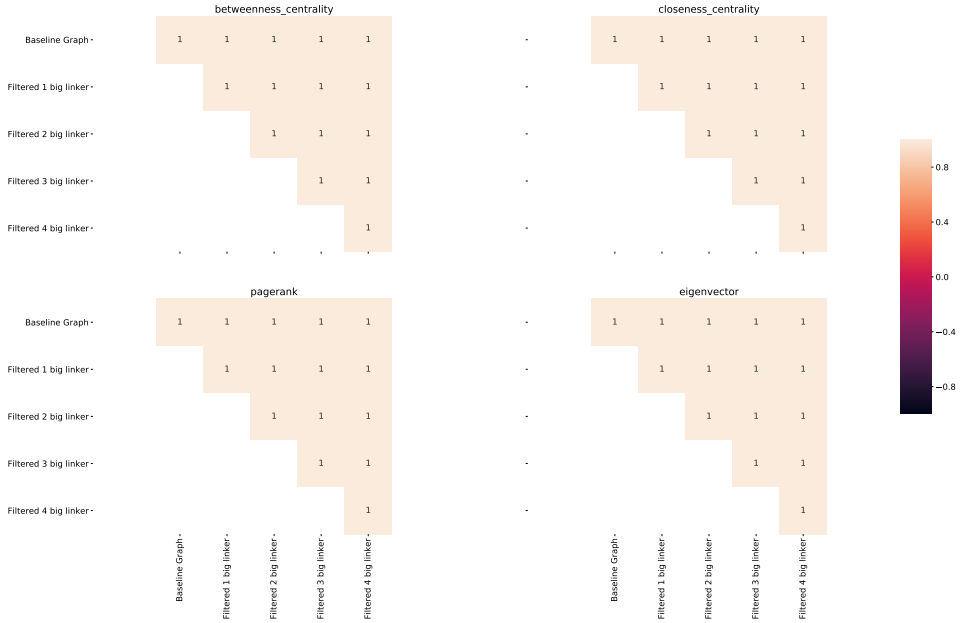

Figure A.0.11: Spearman rank-order correlation of centrality metrics of the users in the top 99.8 percentile with different filters on the newman collaboration projection in data set A.

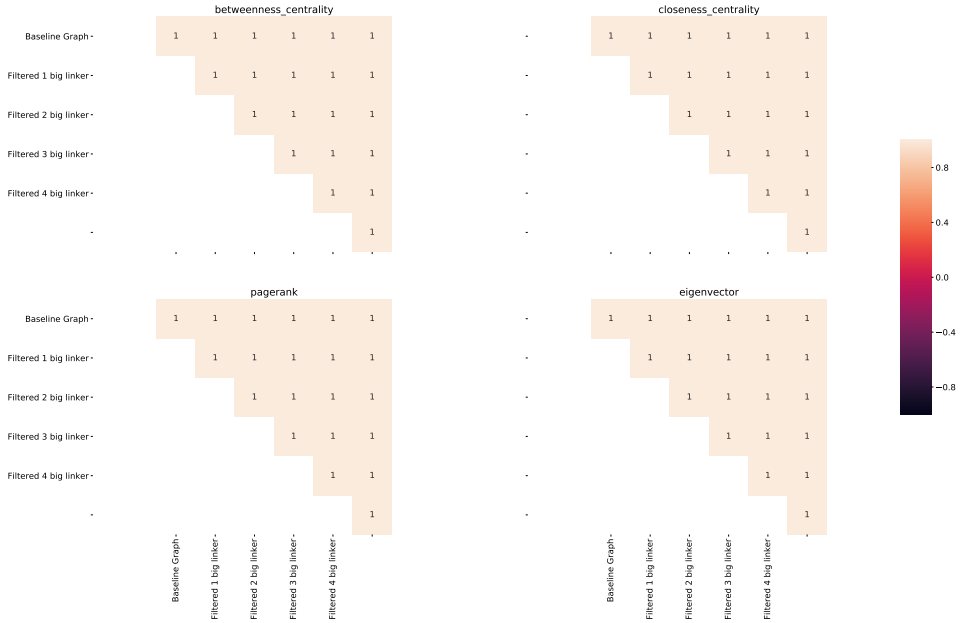

Figure A.0.12: Spearman rank-order correlation of centrality metrics of the users in the top 99.8 percentile with different filters on the newman collaboration projection in data set B.

Bibliography

[1] Ministerie van Justitie en Veiligheid. Veiligheidsagenda 2015 - 2018, Sep 2014. URL `https://www.rijksoverheid.nl/onderwerpen/kinderporno/documenten/rapporten/2014/09/17/bijlage-veiligheidsagenda-2015-2018`.

[2] Richard Wortley and Stephen Smallbone. *Child pornography on the Internet*. US Department of Justice. Internal report, 2006.

[3] Floor Bakker, Hanneke de Graaf, and Stans de Haas. *Seksuele gezondheid in Nederland 2009*. Rutgers Nisso Groep, 2009.

[4] Matt Egan. Thinking of venturing on to the dark web? You might want to change your mind. *Tech Advisor*, Jan 2018. URL `https://www.techadvisor.co.uk/how-to/internet/dark-web-3593569/`.

[5] Van der Bruggen and Blokland. Child pornography and the internet: a systematic literature review. *Under review*, 2018.

[6] Albert-László Barabási. *Network science*. Cambridge University Press, 2016.

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

[8] P.A.C. Duijn. *Detecting and disrupting criminal networks: A data driven approach*. PhD thesis, University of Amsterdam, 2016.

[9] Jean-Loup Guillaume and Matthieu Latapy. Bipartite structure of all complex networks. *Information processing letters*, 90(5):215–221, 2004.

[10] Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social networks*, 30(1):31–48, 2008.

[11] Stephen P. Borgatti and Martin G. Everett. Network analysis of 2-mode data. *Social networks*, 19(3):243–269, 1997.

[12] Mark E.J. Newman. Scientific collaboration networks. II. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132, 2001.

[13] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378, 2000.

[14] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440, 1998.

[15] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[16] Phillip Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182, 1987.

[17] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN systems*, 30 (1-7):107–117, 1998.

[18] Bryce G. Westlake, Martin Bouchard, and Richard Frank. Finding the key players in online child exploitation networks. *Policy & Internet*, 3 (2):1–32, 2011.

[19] Stephen P. Borgatti. Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory*, 12(1):21–34, 2006.

[20] Robert D Nolker and Lina Zhou. Social computing and weighting to identify member roles in online communities. In *Proceedings of the 2005 IEEE/WIC/ACM international conference on web intelligence*, pages 87–93. IEEE Computer Society, 2005.

[21] Eric R. Ziegel. Standard probability and statistics tables and formulae. *Technometrics*, 43(2):249, 2001.

[22] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, 2008.

[23] Tiago P. Peixoto. The graph-tool python library. *Figshare*, 2014. doi: 10.6084/m9.figshare.1164194. URL `http://figshare.com/articles/graph_tool/1164194`.

[24] G. van Rossum. Python tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995.