Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

Riccardo Matteini Palmerini

Graph theoretical approach to sexual predator detection

Master's thesis in Information Security
Supervisor: Professor Patrick Bours
Co-supervisor: Researcher Jan William Johnsen
June 2021

NTNU
Kunnskap for en bedre verden

Riccardo Matteini Palmerini

Graph theoretical approach to sexual predator detection

Master's thesis in Information Security
Supervisor: Professor Patrick Bours
Co-supervisor: Researcher Jan William Johnsen
June 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of
Science and Technology

NTNU – Trondheim
Norwegian University of
Science and Technology

Graph theoretical approach to sexual predator detection

Riccardo Matteini Palmerini

Submission date: June 2021
Supervisor: Professor Patrick Bours
Co-supervisor: Researcher Jan William Johnsen

Norwegian University of Science and Technology
Department of Information Security and Communication Technology

Title:	Graph theoretical approach to sexualpredator detection
Student:	Riccardo Matteini Palmerini

Problem description:

In this project we will apply Social Network Analysis, which is a method to model the chat behaviour as a graph. The vertices will represent the chatters, and the edges will represent the conversations between chatters (where we restrict to one-to-one conversations). Density, centrality, and in- and out-degree of the nodes must be tested for being indicators of sexual predators. The tests should be run on various graph configurations, e.g. undirected graphs, weighted graphs (weight could be the number of messages in a conversation), directed graphs (directions could be an indication of the first message in a conversation, or there could be a (weighted) edge from user A to user B and another from user B back to user A). The goal is to find indicators of sexual predators based on training data and then test these indicators on test data with ongoing conversations, i.e. the graph is changing over time when new messages are added to existing conversations or new conversations are started.

Date approved:	2021-02-15
Supervisor:	Professor Patrick Bours
Cosupervisor:	Researcher Jan William Johnsen

Abstract

With the spreading of the technology, we are able to get in touch with distant friends as well as unknown persons. Behind a monitor, evil presences can hide their identity and act in freedom. Sexual predators in the Internet are able to communicate with minors by risking less than in the real life. Their activity is supported by the advent of online platforms characterized by public chat rooms.

The thesis seeks to use graph theory features and centrality measures in order to highlight possible indicators for the cyber sexual predators detection. More specifically, the study is not interested on mining chat logs, but rather it focuses on testing features like the number of edges and the number of messages exchanged to find out abnormal behaviours within the network.

The thesis analyses how the centrality measures can be used to detect abnormal nodes within a network. The thesis contributions is trying to determine a way for highlight strange behaviours of users in public chat rooms without looking at content of the conversations.

Sammendrag

Med teknologispredningen er vi i stand til å komme i kontakt med venner langt unna så vel som ukjente personer. Bak skjermer kan personer med onde hensikter skjule identiteten deres og handle i frihet. Overgripere tar mindre risiko når de kommuniserer med mindreårige via Internettet enn om de samme handlingene ble gjort i det virkelige liv. Den kriminelle aktiviteten fra overgripere støttes av fremveksten av elektroniske plattformer som bruk av offentlige samtalerom.

Denne avhandlingen søker å bruke grafteoretiske funksjoner og sentralitetsmålinger for å fremheve mulige indikatorer som detekterer seksulle overgripere. Spesifikt er denne avhandlingen ikke interessert i datautvinning av samtale logger, men fokuserer på testing av funksjoner som antall kanter og antall utvekslede meldinger for å finne unormal oppførsel i nettverk.

Avhandlingen analyserer hvordan sentralitet kan brukes til å oppdage unormale noder i nettverk. Avhandlingen bidrar til å fremheve merkelig oppførsel hos brukere i offentlige samtalerom uten å se innholdet i samtalene.

Preface

This thesis concludes my Master as student enrolled in a Double Degree program between the Alma Matter Studiorum, The University of Bologna and The Norwegian University of Science and Technology in Trondheim. My project's supervisor has been the Professor Patrick Bours and the Researcher Jan William Johnsen.
The topic idea was proposed by the Professor Patrick Bours. I hope our research in this field will contribute for further works as well as I hope it will open the doors for reaching my dream job.
To all of you, a good read.

Riccardo Matteini Palmerini

Acknowledgment

I would like to thank my supervisor, Professor Patrick Bours, and my co-supervisor, Researcher Jan William Johnsen, for guiding me throughout the project and also for the patience and availability towards me.
I would also spend some words in order to thank my parents, Maurizio and Cristina, for motivation and support. Thank to the support of my parents I have been able to join this study opportunity.
Finally, I would like to thank my girlfriend, Chiara, she has always been close to me, even if far away.

R.M.P.

Contents

List of Figures

List of Tables

List of Acronyms

HITS Hyperlink-Induced Topic Search.

JSON JavaScript Object Notation.

LCT Luring Communication Theory.

LIWC Linguistic Inquiry and Word Count.

PJ Perverted Justice.

Chapter 1
Introduction

Online platforms in which users can freely communicate in public chat rooms are getting even more present in everyday life. Public chat rooms are used by cyber sexual predators to lure minors.
The thesis aims to analyse the phenomenon of cyber grooming and to determine key indicators for cyber sexual predators detection. Graph theory and centrality measures are the fundamental tools which are utilized.

1.1 Problem Description

Cyber grooming is a spreading phenomenon, as reported [Chia] and [Chib] in their statistics.
The purpose of the thesis is to test graph theory features in order to find out indicators for cyber sexual predators detection. In this project we will apply Social Network Analysis, which is a method to model the chat behaviour as graph. We will deal with chat logs from an online platform of games, based on public chat rooms. The chat rooms will be managed in order to be able to work with one-to-one conversations.
We will focus on the number of edges and the number of messages exchanged between users to find abnormal behaviours. Centrality measures will help to determine the role of the users. We will provide a local point of view (by looking at the neighbors of a node) and a global point of view (by highlighting the nodes which take part in more than one chat room).

1.2 Thesis Outline

First, background knowledgment and related work are provided. Then, the choice of methods are presented, followed by the experiments and the results. At this point, discussion about the results and the limitations we met during the work are shown. Finally, a recapitulation as conclusion is provided as well as suggestions for further work.

- **Chapter 2** provides a description of the phenomenons of the pedophilia and the cyber grooming. It also includes an overview of the graph theory and the centrality measures as well as related work;
- **Chapter 3** presents the methodology we used through the thesis;
- **Chapter 4** gives a description of the dataset we utilized, providing basic information;
- **Chapter 5** presents the experiments we performed as well as the result obtained;
- **Chapter 6** provides explanation about the results we got and the limitation raised up during the study;
- **Chapter 7** provides a recapitulation of the work and interesting features to be analyzed as future work.

Chapter 2
Background theory selection

In this section, the definition of pedophilia and cyber grooming will be provided and then the analyses of this phenomenon will be described. Therefore, it will be introduced the fundamental mathematical tool which will be useful for achieving the final goal of the thesis: using graph theory to detect sexual predators.

2.1 Pedophilia

2.1.1 Definition of pedophilia and cyber grooming

Merriam-Webster states: *"pedophilia is a sexual perversion in which children are the preferred sexual objects. Specifically, a psychiatric disorder in which an adult has sexual fantasies about or engages in sexual acts with a prepubescent child"* [Mer].

Although the definition of pedophilia is a common knowledge nowadays, there are different kind of pedophiles. Therefore, it is not possible to draw a unique psychological profile.
Indeed, in [MB] the authors, prestigious psychologists, recognize pedophilia not only as a crime but rather as an antisocial behaviour. It is required to be studied from different angles. Factors as institutional aspect, sexual education and act of violence, contribute to build up the complex and perverted mind of a pedophile. The article describes and analyses the grounds of this disorder. That may range from a regression to the childhood due to an early trauma, to the need to feel superior and so the pedophilia is the way to express authority. As the authors draw, pedophiles act in different ways. They may build trust and love relationship with a child or, on the other hand, they may be aggressive and threat the victim. Moreover, in a complete view, they may alternate lures and threats, gifts and punishments. A common point in the behaviour of the pedophiles is that they know the prey. They know the habits and the customs of the child in order to envelops an aura of confidence before striking. The Mental Health Disorders warns that pedophiles typically are adults known by the children. They can be a family member, like a step-parent, attracted only to

children within their own family. Or pedophiles can be autorithy persons, like teacher or coach [GR]. It doesn't mean that an external person is excluded in advance, but it was less frequent. Therefore, prior the advent of Internet, sexual predators were usually inner family's members.

Due to the born of the World Wide Web, it is easier, simpler and faster to keep in touch with a far friend and also meet new people. Like all great powers, if they are left in the wrong hands, they bring nothing good.
The Child Safe Net defines the cyber grooming as *"the process of 'befriending' a young person online to facilitate online sexual contact and/or a physical meeting with them with the goal of committing sexual abuse"* [Chia].

In order to highlight the danger of cyber grooming, the Child Crime Prevention & Safety Center estimates 500'000 online predators active each day. This association estimates that 89% of sexual harassment toward children happen through the Internet, both in public chat rooms and via private messages. The 25% of them reported cyber predors which explicitally aimed to obtain naked photos of the children. While, the 4% of the cases, reported aggressive solicitations as attempts to meet in person [Chib]. Even though, these numbers are scary, many cases are not registered due to fact that not in all countries an exact definition of cyber grooming exist and, furthermore, not all parents are warned against cyber sexual predators.
Nowadays, cyber grooming is a real problem.

2.1.2 Predators' Behaviour

As well as in the physical world the behaviour of a sexual predator cannot be defined by an unique theory since pedophilia is a complex topic, it is the same also for the virtual world. Despite that, it is possible to underline a general scheme.
One of the most famous model which explains the different child grooming stages in the physical world is the well-known Olson's Luring Communication Theory (LCT) [LNJLBLTKK], in which the approach of a pedophile is subdivided into three steps (once the victim is chosen):(1)*Deceptive Trust Development*, (2)*Grooming Stage* and the final stage in which the predator seeks to (3)*Physically Approach* the minor. The LCT is not an absolute rule, predators can move into these steps. They may just skip one or simple go back to the previous stage. The important things to be stressed are the repeating words and expressions in the different stages and the starting point for which a pedophile achieves its main goal: a pedophile want to build a trust relationship.
[AEMH] describes not only the phases of this theory but also analyzes which expressions are used in the different stages and how.

Regarding LCT, the first stage is constituted precisely by the exchange of personal information (i.e. age, likes, dislikes) in order to build a common ground with the

child and to get its full confidence. Then, the second stage can begin: the predator starts to trigger the sexual curiosity of the minor. This stage can be identified by the use of sexual words. During the second stage, the cycle of entrapment begins. The trust relationship get even stronger and at the same time the minor is estranged by the real life, getting isolated from family and friends. Once the predator perceives that the victim will not betray its "new friend", it is ready to enter in the final stage. Trough the third stage, the predator is able to sexually harass the child, but only after requesting information regarding the scheduling of the parents.

Starting from LCT, Rachel O'Connel extended and improved this theory in order to be directly applicable also in the Internet world [O'C]. As for the theory of Olson, it is not an always true scheme followed step by step by the predators in which the stages have the same time duration, it is a general overview over the cyber predator's behaviour which highlights the importance of a trust relationship with the minor. O'Connel work starts with the description of the victim selection. Before sending private messages, usually the online predators choose ad hoc nickname and profile picture to deceive and bait the minors, by acting as teenagers. Then, they may present themselves in the public chat room. But there are also pedophiles who pose as teenagers hoping to attract an equivalent age. Or they may read all the public conversations acting as a viewer in order to choose the victim without being noticed.

Once the victim(s) is(are) identified, the predator will move toward a private chat and it will follow more or less the following scheme (which is very similar to LCT, it just splits in a more complete way the different steps):

- *Friendship Forming Stage*, in which the pedophile gets in touch with the child, it requests general information about the victim and a picture to be sure that the minor is satisfying its desires;
- *Relationship Forming Stage*, it is an extension of the previous stage, more details about the life of the victim are requested;
- *Risk Assessment Stage*, as the name's suggests, the pedophile ensure itself about for instances the location of the computer, the scheduling of the parents etc.. ;
- *Exclusivity Stage*, the pedophile explicitally make sure how much the child trusts him. Depending on the answer from the child, the pedophile decides to move to the next step, focused on more intimate and sexual issues;
- *Sexual Stage*, it is the final stage, in which the predator starts asking about the sexual life of the minor, proceeding carefully toward exchange of erotic pictures and in the worst scenario a physical meeting between the two.

Although predators act in various ways in order to achieve different goals and uniquely defining category of them it is not possible, ChildSafeNet provides a list of three distinct types of groomers.
The first type described is the *Distorted Attachment*: it reveals its identity, behind its actions there is the need to be loved by the child. That's the reason why it spends a great amount of time creating a friendship with the victim.
On the other hand, the *Adaptable Offenders* and the *Hyper-Sexual Offenders*, as the names suggest, adapt them grooming style to make fast contact with a multiplicity of young victims. The main difference between these kinds of predator is how and how often they use indecent and sexual images with the children. Moreover, the *Hyper-Sexual Offenders* are part of pedophiles' communities to share within the "trophies" earned by them hunts [Chia].

2.1.3 State of Arts

In this subsection, literature and state of arts, which are related to draw the behaviour of a cyber predator, are introduced. Detection of cyber sexual predators is a raising research quest, in order to be able to protect minors.

First of all, as mentioned in the introductory chapter, the goal of the thesis is to detect the presence of cyber predators due to the analysis of public chat in online platforms. In order to achieve the goal, the behaviour of a cyber sexual predator (or, in a more general view, the one of a cyber criminal) is studied by the starting idea that on a public chat room a predator will seek for the victims, subsequently bait them the hooks and reach the abuse going through the stages described by [O'C].

Therefore, what we expect to find studying cyber predators chat logs is a one-to-multi communication, characterized with a predator who tries to get in touch with multiple minors.
As far as we know, a one-to-n approach has not been studied yet, due to the difficulties in finding chat logs. Most of the studies are based on datasets provided by Perverted Justice (PJ), a non-profit foundation in which volunteers pose as children and exchange messages with possible cyber sexual predators. [GKS12] confirmed the limitation of this approach: the lack of actual and real-world dataset. However, PJ's chatlogs are still good enough dataset to draw guidelines on the behaviour of cyber predators. Indeed, the authors in [GKS12] used the dataset provided by PJ to collect important information about the linguistic styles through the stages of a cyber grooming. A similar study is performed in [CFA] and [UPdV14].
In [GKS12] the chatlogs are manually classified into the different stages and they used a word counting program (Linguistic Inquiry and Word Count (LIWC)) to obtain the correlations between the word categories and online grooming steps.
[CFA] takes advantage of LIWC to compute the psycho-linguistic patterns and a

system based on frame and semantic label (achieved respectively with FrameNet and SEMAFOR) to characterise the cyber grooming stages.
[UPdV14] instead, is mainly focused on the emotional side of a predator and the work aims to underline in which stages the predator acts in a positive way and which are the events that make it go angry; they utilized to extract positive and negative words SentiWordNet.

Although [GKS12], [CFA] and [UPdV14] are based on different approach, they stress the importance of the relationship forming stage, which is the most prominent stage. Moreover, the significance of working on a dataset with a low false positive rate is highlighted.

Moving toward the detection of cyber predators, [BK19] and [CFA] are meaningful examples. Before going in depth with the methodology they used, it is important how the storage of the database is done. Chat messages are not always written in a correct way, such that persons like to use slang expression, emoticons, exaggerate with the use of letters and dots. All these can be interpreted as noise and and should be removed for a depth analyses.
It is worthy to notice the technics used by [BK19] and [AA14]. In [BK19] authors take advantage of the combination between three approaches (message-base, author-based and conversation-based), five classification algorithms and two features sets. As main achievement, they prove that early detection is possible. On the other hand, [AA14] uses a graph approach for text mining framework, based on the extraction of keys (vocabulary and users) in order to build a bipartite graph by a self-customized Hyperlink-Induced Topic Search (HITS).

Other important features related to the study of the behaviour of such predators are described in [SS], [EC10] and [SYSC06]. In these articles the authors describe how it is possible to highlight the subject of a conversation ([SS]) by focusing on the break within two topics and the vocabulary utilized ([EC10] and [SYSC06]).

It is also useful the work [KCAC08], in which the authors analyse chat logs in order to predict the sex and the age of an user, based on the stylistic preferences such as word frequencies and the length of the sentence, use of syllables, punctuation marks and the use of function words [Rud21]. They demonstrate that a child tends to have a smaller vocabulary and prefers using emoticons.

These articles are mostly based on the detection of cyber predators trough the analyses of the content of the conversations. Instead of the thesis will mainly focus on the frequencies of the messages.

2.2 Mathematical Tool

In the thesis we will focus on finding abnormal behaviours.
A simple tool is the **z-score**. Mathematically, the z-score for a value is defined as the value minus the mean of the population in question, divided by the standard deviation.

$$z = \frac{x - mean}{standard\ deviation}$$

Actually, it represents how many standard deviation a value is far by the mean. A positive value means that the value is above the mean, while a negative one is below the mean.

2.3 Graph Theory

Graph theory will be the main tools in this thesis to analyse dataset which will be described in the next chapters (Chapter 3 and Chapter 6). The following sections provide an overview of the graph theory and the features/measurements that can be achieved through the study of a graph.

2.3.1 Overview

Graph theory is a branch of mathematic concerned with networks of points connected by lines. A graph is a mathematical abstraction, it can represent various scenarios with application in chemistry, operations research, social sciences and computer science. Since the graphical representation, graph theory is widely used to illustrate social networks and the relationships within them.

The first time the word "graph" was utilized dates back to the 1878, when Sylvester James Joseph published a paper in Nature, speaking about analogies between quantic invariants and co-invariants of algebra and molecular diagrams [Syl]. Despite that, the ideas behind graph theory can be traced back to the 1735, when a Swiss mathematician called Leonhard Euler proposed a solution for the seven bridges problem Konigsberg [Teo]. However, the first textbook goes back to the 1936, published by Dénis Konig [Kon]. The main book, the one which is still considered the world over to be the definitive textbook on the graph theory was written by Frank Harary in the 1969 [Har] [Wil].

2.3.2 Features and Measurements

Since graph theory is utilized in different fields, there co-exist various terminology for this matter, in this thesis a mathematical approach will be followed.

A graph G is composed by an ordered triple, N(G), E(G) and f(G). That are non-empty set. N(G) is the set of nodes, E(G) is the set of edges, disjointed by N(G)

and f(G) collects the incidence functions which associate each edge e to an unordered pair of nodes [BM].
If the edge has a direction, so a starting point and an ending node, we are dealing with **directed graph**. Otherwise, we are talking about **undirected graph**. If more than one edge links two nodes, the graph is called **multi graph**. Figure 2.1 shows respectively an undirected graph and a directed graph, commonly called **digraph**.

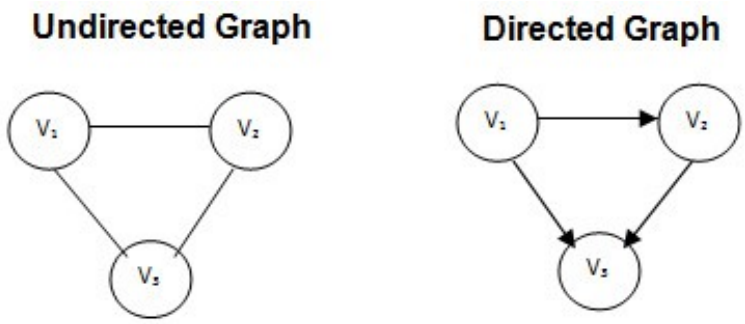


Figure 2.1: A graph and a digraph [Ind]

Despite the definition by [BM] sounds demanding, most of the concepts in graph theory are suggested by the graphical representation. From now on, the basic concepts and the more interesting features for this thesis will be described.
In order to introduce more vocabulary, two nodes are **adjacent** when they are incident with a common edge.
Considering a portion of the set of the nodes and the corresponding subset of edges, it is forming a **subgraph**(Figure 2.2). While if an undirected graph is considered, a subset of nodes, such that every two distinct nodes are adjacent (hence the composed subgraph is **complete**), it is called a **clique**.

Another important concept is the **node degree**, that is, considering a node in a graph, the number of edges incident with the above node.
Moreover, a simple tool on which the work of the thesis will be based, is the **weight**, that actually consists on assigning a value to an edge to give it more or less importance. To make more clear the concept, in a network analysis the weight can represent the amount of traffic between two nodes or in chat log, as in this thesis, the number of exchanged messages within two persons. The weights of edges are used in general to calculate the **shortest path** between two nodes, such that the path in which the total length is minimized.

Incident matrix, usually indicated with $\mathbf{M}(\mathrm{G}) = [m_{ij}]$ and **adjancency matrix**, $\mathbf{A}(\mathrm{G}) = [a_{ij}]$, are the mathematical representations of the graphs. The first one

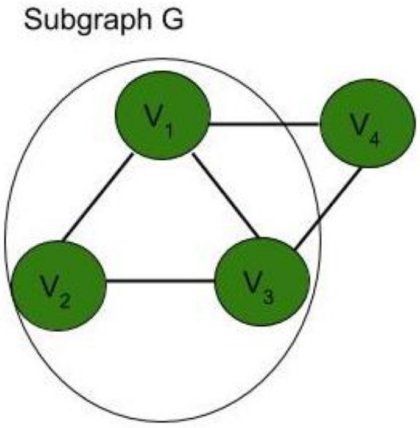


Figure 2.2: Subgraph [Esh]

is a nodes$x$edges matrix. Therefore an element of the matrix, m_{ij}, takes in account how many times $node_i$ and $edge_j$ are incident. m_{ij} can be equal to 0, 1 or 2 (in case of loop).
While the adjacency matrix represents the relation within the nodes of a graph, it is a nodes$x$nodes matrix and the element a_{ij} is the number of edges joining $node_i$ and $node_j$. Figure 2.3 shows a graph and its incident and adjancency matrices.

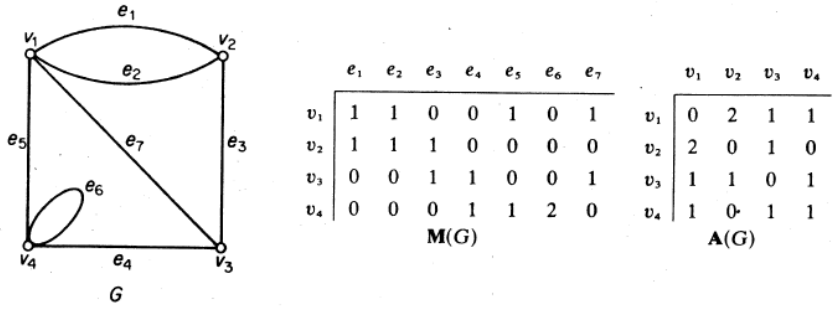


Figure 2.3: A graph and its matrix representations [BM]

In general, an adjacency matrix of a graph is smaller than its incidence matrix, for that reason graph are usually stored by using the adjacency matrix. Moreover, the **centrality** and its measurements are based on this kind of matrix.

Centrality identifies the most important nodes in a graph. The first one who

summarized the concept and the measurements behind centrality was Linton Freeman [Fre]. With reference to a network, the basic idea of centrality's studies is that a person, who knows many people and/or is the most relevant in that network, will tend to be in a central position. In the purpose of this thesis, predators can be considered as central node in a graph. For instance, predators classified as Adaptable or Hyper-Sexual offender before building a relationship with a child, they "cast the hooks" to several potential victims.

Degree centrality of a node, $C_D(i)$, that can be split in *indegree* and *outdegree*, is actually the node degree:

$$C_D(i) = \sum_{j=1}^{n} x_{ij} = \sum_{i=1}^{n} x_{ji}$$

Where x_{ij} is 1 or 0 whether and edge between two nodes is present or not, and n is the total number of nodes. The degree centrality can be seen as the level of connections a node has, without considering the direction and neither the weight of the edges. Since this kind of measurement is strictly dependent on the size of a graph, in order to put in relation two different graphs the normalized degree is used:

$$C'_D(i) = \frac{\sum_{j=1}^{n} x_{ij}}{n-1} = \frac{C_D(i)}{n-1}$$

However, degree centrality doesn't take in account the rest of a network, just the number of nodes directly tied to an other one. In order to study the cenetrality with respect to an overview of all the graph, other tools exist: **eigenvector centrality**, **betweenness centrality**, and **closeness centrality**. . The eigenvector centrality of a node expands the concept of degree centrality by analysing adjacent nodes connections in their turn. It deals with how many nodes are connected to the designated node, but also how many adjacency nodes each of its neighbours have. This value is obtained through an algorithm based on the research of the largest eigenvalue of the adjacency matrix in an iterative way. It can be also performed by taking into account the weight of the edges.

On the other hand, the betweenness centrality and the closeness centrality don't take into consideration only the number of nodes but rather the position in which the candidate node is placed. In general, the betweenness centrality calculates the potential control over the network of a node, while the closeness one the independence of a node. The betweenness centrality is based on how often a node rests between two other nodes:

$$C_B(k) = \sum \frac{d_{ikj}}{d_{ij}} \text{ with } i \neq j \neq k$$

where k is the candidate node, d_{ijk} represents how many time the shortest path from node i to j passes through node k and d_{ij} is the number of total shortest paths.

While the closeness centrality deal directly with the shortest path:

$$C_C(i) = \frac{1}{\sum_{j=1}^{n} d_{ij}}$$

and d_{ij} is the distance connecting node i to node j.

Figure 2.4 presents a graph in which the nodes characterized by the highest value of a certain kind of centrality are indicated.

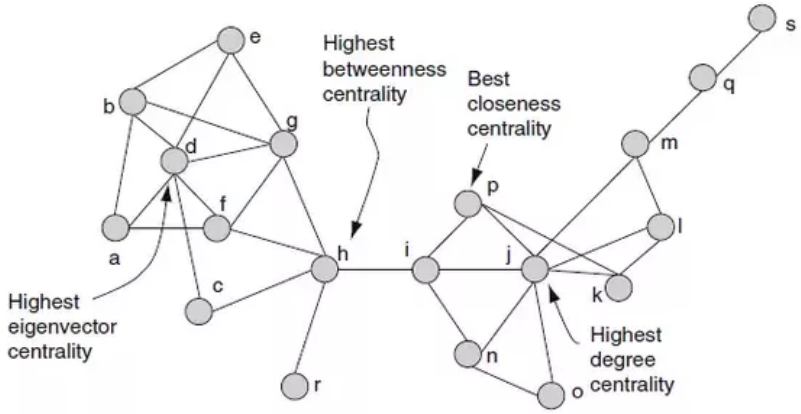


Figure 2.4: A graph and the centrality measures [BM]

2.4 Related Works

In the following section related works, which will properly help to achieve the goal of the thesis, are considered.

First of all [POC09] describes and analyzes the structure and evolution of an online community which is aimed to sustain social interaction and help members enlarge their circles of friends. The authors focused on the reachability, robustness, average distance and clustering studies. The main overcomes are the "small-world" and fat-tail degree distribution properties. Small-world means that most nodes in the network can be reached from every node by a small number of steps. Fat-tail degree distribution references to the fact that some nodes are highly connected, hence the network load distribution becomes highly unbalanced. Therefore, the central role of some nodes can be highlighted.
Figure 2.5 shows an example of a network following a fat-tail degree distribution. Mathematically, the curve of the degree distribution can be seen similar to a normal distribution but with the curve-bell more close.

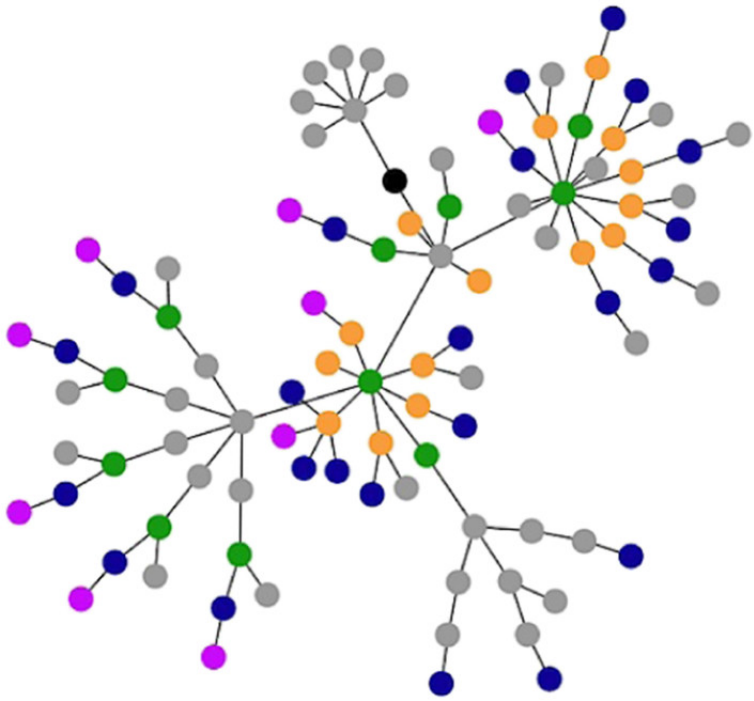


Figure 2.5: A network with a fat-tail degree distribution [TL]

On the other hand, [IFD12], [LPLC10] and [Joh] concentrate in digital forensic for malware analysis and network forensics. [IFD12] and [LPLC10] use real hacker dataset, while [Joh] used the Enron corpus dataset.

[IFD12] takes advantage of mining cliques technique by modifying the Apriori algorithm in order to efficiently extract frequent patterns and then analyses in depth the cliques through topic mining. In this way, they are able to identify a meaningful group of cyber individuals and even the connection between them.

[LPLC10] analyses also the network of a hacker organization by using centrality measurements and cliques. The results is that the considered hacker organization is not a centralized network, despite the fact that subgraphs with high centrality features can be found. They highlighted the inner organization of a malware community: different roles, different jobs and different power.

[Joh] generalises the studies of the previous one by considering e-mail addresses of a service company. In order to identify important actors, he utilized social network analyses centrality measures within a systematical evaluation of the methods by showing how reliable can be communicability betweenness and closeness centrality.

Moreover, it explains how graph theory can be applied in digital forensics for identifying targets. Through the neighbourhood approaches (e-neighbourhood and k-NN), it is possible to highlight the "starting" node in order to follow the edges and finding similar nodes.

Chapter 3
Choice of methods

Chapter three aims to analyze and motivate the choiches done in thesis in order to achieve its scope. First of all the dataset will be described and even the way it is managed. While in the next section, the tools and the decisions related to the technical side are provided in depth.

3.1 Dataset

As mentioned in Section 2.1.3, finding an appropriate dataset is a huge issue: most of the research on the pedophilia's field are based on the PJ's dataset, built up through conversations between volunteers and cyber sexual predators. Volunteers act as children. On the other hand, real-world conversations are protected by privacy policy.

To proceed with the thesis, we worked with anonymous data from an online platform of games. The game-play is based on different public rooms in which users can communicate or just be passive during the break between two matches. Users can participate in more than one room. Therefore, the conversations come from public group chats, they are not one-to-one private messages. Moreover, in these chat logs any **evil** user is not present. But it is still a good starting point to figure out the behaviour of online communities in order to be able to detect abnormal behaviours of users.

The anonymous data gathered chat logs within users of the game, collected in a relative short time of 31 days (between 9/2/2021 and 11/3/2021). The provided file was a JavaScript Object Notation (JSON), a format widely used to store and transmit simple data structures and also simple to work with through open source python's packages as for instance Pandas[1].

[1] *Pandas* is an open source, BSD-licensed library [Pan]

The nicknames and the content of the messages had been hidden. Unique identifiers replaced the nicknames of the users, while the conversations were summarized as the number of messages from a node to another node. Indeed, each row of the file is on the form From: X, To: Y, Weight: Z, where X is an unique number indicating a user who has sent Z messages to the unique user Y. In such a way Z is the **unilateral** number of messages from node X to node Y, so in just one direction.

How the dataset is utilized to create graph is discussed in the next section.

3.2 Technical Choices

First of all, it is important to highlight the role of the graph theory in such studies. By using nodes and edges, it is simple and intuitive to represent and study the relationship within a network.

Given the fact that in the provided dataset "evil" presences are not present, the object of the thesis is to study the patterns of the different nodes and underline users whose behaviour quite diverges from the mean. Such a way, having in mind the behaviours of a community, further studies can be done in order to detect anomalies. The measuraments and the achieved results are respectively analyzed in Chapter 5 and in Chapter 6, while further studies are discussed in Section 7.2.

We worked in a python environment, taking advantafe of the package **NetworkX**[2], which includes the possibility to use edges' labels as key values in some computation (i.e. for the centrality measures).

The provided data has been utilized to produce the graph within the relationship between the users of the game. Users are represented due to nodes with unique identifiers (numbers, from 1 to 196'489), while the edges act as relations between two users. There are 754'031 edges. Since the data is in the form From: X, To: Y, Weight: Z, the edges have the weight as label, that is the number of messages Z from user X to user Y. We have to underline that the weight is the **unilateral** number of messages within two nodes and there is no self-loop. Both directed and undirected graphs are created in order to perform different kind of analysis. Directed graph is needed to compute the ingress/egress node degree and centrality measures (as betweenness and closeness centrality). While undirected one is used to identify the cliques of the network and the degree centrality. With regards to the generation of the undirected graph, in order to be able to highlight the contributes of both two nodes in a conversation, the *MultiGraph* object is utilized, which is available within

[2] *NetworkX* is a free software, BSD-licensed library. The python package permits to create, manipulate and study the structure of complex networks [Net]

the NetworkX package. On the other hand, the directed one is a simple *DiGraph*. The weights of the chats are implemented as labels on the edges.

Since the size of the network is very large, one of the most intuitive choices to reduce the dimension of the network is by looking at the cliques. In this approach, we obtain a more manageable dataset. Subgraphs, generated starting by cliques, represent users that communicate between each other. In a realistic way, they can be interpreted as communities/rooms of the game and some type of symmetry can be looked at. In addition, cliques are not just useful to see in how many cliques a node is present, but also to check the nodes which are quiet isolated by the rest of the network.
Another way to reduce the complexity is by creating subsets of nodes, based on different attributes.

In order to analyze the role of the nodes, we took advantage of centrality measures. As discussed in Section 2.3.2, there are several kind of centrality measures available and the definitions are provided in the section mentioned above. Here, we focus on the meaning of these features. In a network, the degree centrality can be translated as how much an individual is more or less inclined to diffuse/know information. A high value means that a node has a lot of connections, removing this node can impair the network. Eigenvector centrality expands the previous concept by including in the computation also the neighbours of a node. A node with a high eigenvector centrality represents a subject which is most connected to most other significant people in a network. This node has an important role within the network but it could be removed without critical side effect. A node with a high betweenness centrality plays role of a gatekeeper, if it is removed, the flows of information trough a network changes radically. While closeness centrality shows how easy/hard a node is able to communicate with others in a network by minimizing the degree of separation. Removing a node with a high value of closeness centrality generates a "lack" of edges in the network, such as reaching nodes requires more effort.

Chapter 4
Data Description

In this chapter a deeper overview of the dataset will be provided, emphasizing the main features and the assumptions we have made.

4.1 Data Description

The dataset came from an online game platform in which users communicate in a limited amount of time through chat rooms between two matches. Therefore, the gathered conversations are not one-to-one communications in private chats but instead they are the exchange of messages in the public chat of a room. Users can join in more than one room, actively participate or being passive players (just reading the conversations).
The chatlogs are collected over a period of a month (9/2/2021 - 11/3/2021). Since the privacy issues, in order to proceed with the analysis of the dataset, the name of the users and the content of the messages had been hidden.

In addition, we have the guarantee that all the information in the dataset represents users. Since they are talking in public rooms and the rows in the datset are in the form From: X, To: Y, Weight: Z, we supposed the data was recorded such that user in a public room broadcasts messages to all other users in a room and at the same time receives messages from all the users within the same room. Such a way, if user X sends a message in a room with four more users, all the weights of the edges outgoing from X to the other users in that room increase by one.
With the available chatlogs, we were able to create graphs in order to take advantage of the graph theory and to study the network. We built an undirected graph and a directed one. In both, there are present 196'489 nodes and 754'031 edges. Users are represented due to nodes with unique identifiers (numbers, from 1 to 196'489), while the edges act as relations between two users. Since the data is in the form From: X, To:

Y, Weight: Z, the edges have the weight as labels, that is the number of messages Z from user X to user Y. Thanks to simple functions present in the package **NetworkX**, we have checked the graphs are not connected, the directed graph is not connected and not even the undirected one. In the undirected graph, the most connected component is made by 99'599 nodes. While the largest connected component in the directed graph includes 75'393 nodes. The total number of connected components is 35'249 and analyzing the cliques, we obtained 325'338 cliques.

The nodes within the graph represent a huge variety of statistic samples due to the different behaviours: from the node which has one connection and sends a couple of messages, up to the one which plays a central role in the network with more than one thousands links exchanging hundreds messages. Visualizing such a graph in an understandable way requires a lot of resources. By exporting the graph as a Gephi[1] file, we have gotten a huge black blob. More immediate, it is looking to the degree

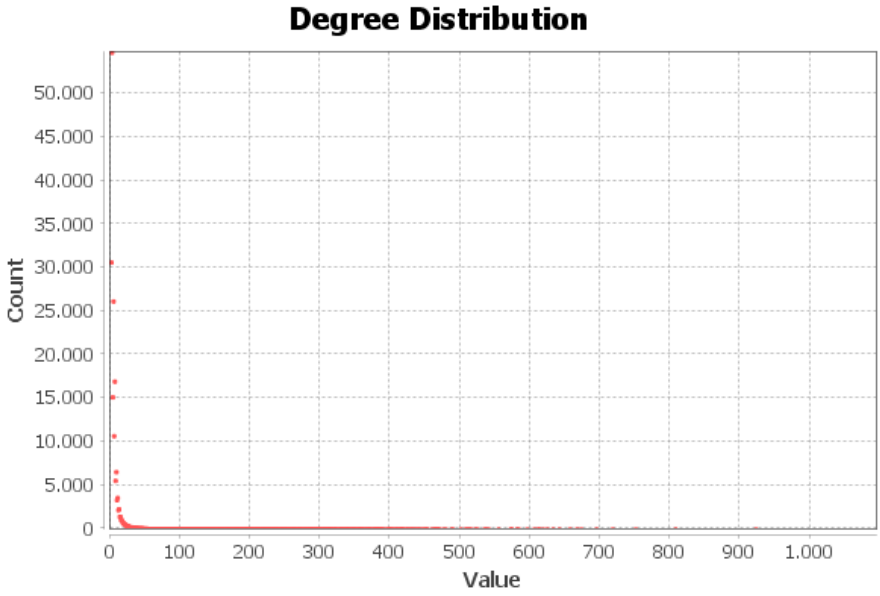


Figure 4.1: Degree distribution of the directed graph. On the Y axis the number of nodes, on the X axis the degree of the nodes

distribution. Figure 4.1 displays the degree of the nodes with respect to the number of nodes in the graph with that degree. We can see as the majority of the nodes are

[1] *Gephi* is an open-source and free software for the visualization and exploration of all kinds of graphs and networks [gep].

very close to the origin of the axes, with a degree value less than ten. The average degree value is 3.838. The mininum degree value is 1, while the maximum degree value is 1'094. Table 4.1 better shows how the nodes are distributed. A similar and

Degree range	Number of nodes
1-3	100'633
4-6	53'916
7-10	19'175
10-15	8'524
15-20	3'835
>20	10'406

Table 4.1: Degree of the nodes and the number of nodes with that characteristic

correlated features is the weighted degree distribution. The weighted degree of a node is not just the number of incident edges, but rather it is the sum of a certain attribute of the edges. In our case we utilized the number of messages between users as attribute, so the weight of edges, implemented in the graph as labels of the edges. Figure 4.2 display the weighted degree of the nodes with respect to the number of nodes with that characteristic. The average weighted degree is 745.285, with the majority of the population close to the origin of the axes. The minimum value is 1, while the maximum is 954'073.

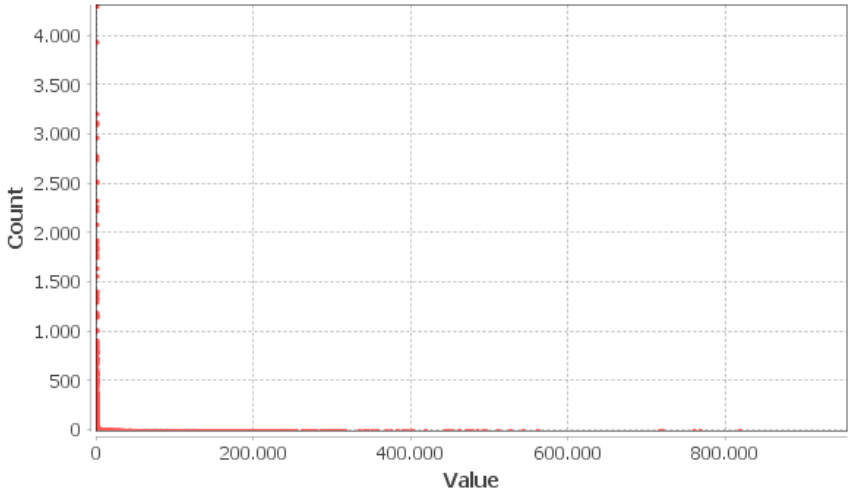


Figure 4.2: Weighted degree distribution of the directed graph. On the Y axis the number of nodes, on the X axis the value (the sum of the weights of the edges of a node)

We can see how the degree distribution and the weighted one rapidly decrease with respect to the the degree value. We have tried to apply different layouts for such huge graph. **ForceAtlas2** pulls strongly connected nodes together and pushes weakly connected nodes apart, it fits for graph where a node doesn't have many neighbours and the degree distribution is decreasing. **OpenOrd** is recommended for networks up to one million nodes and it is very useful to detect clusters. **RadialAxis** layout is also used for very large graph and interesting to study homophily (theory in network science, it states that similar nodes tends to be more attached to each other than dissimilar ones).
OpenOrd was the only layout we have tried that gave us a "nice" overview of the graph. Through figure 4.3 we can understand in depth the complexity of the network in question.

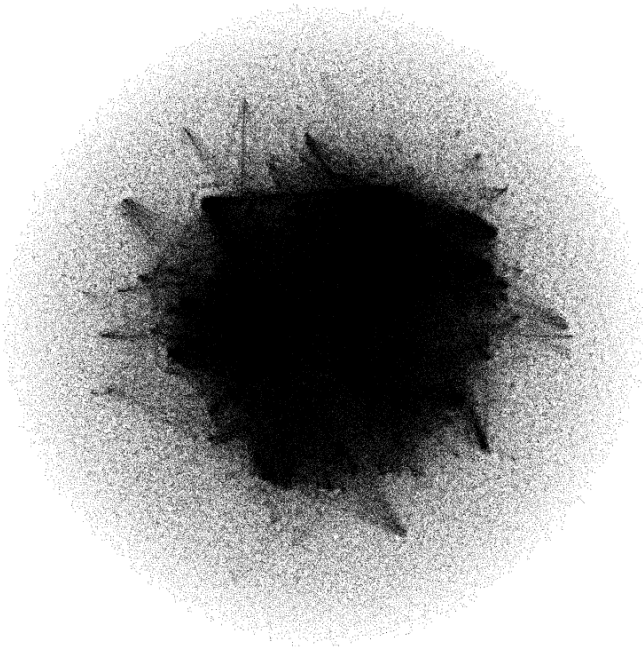


Figure 4.3: Visual representation of our network through OpenOrd layout

Chapter 5
Experiments and Results

We considered two approaches to classify the nodes: a neighbors approach and a cliques point of view. In this chapter they will be introduced and further investigated through the centrality measures.

5.1 Introduction

In our study, we are interested to figure out nodes with an abnormal behaviour. It means nodes that behave in a different way with respect to normal nodes. In order to put in comparison measurements, normalized measures are recommended since the graph is not connected (indeed it is composed by various components). Working with normalized values in such a huge graph is quite difficult, because it corresponds to deal with very small number (of the order of $10^{-10}/10^{-15}$); moreover, the computations are demanding with respect to time and resources . Due to the complexity of the graph (almost 200'000 nodes and 800'000 edges), we focused on the creation of subsets. We followed two different path: a neighbors approach and a cliques point of view. In the neighbors approach, we classified the nodes on the base of attributes like the number of edges, the number of messages per link and the difference between the number of sent messages and the received ones. On the other hand, the starting point for the cliques approach is the splitting within cliques.

In order to find outliers, we took advantage of the centrality measures. In particular, we stressed our attention on the betweenness centrality and on the closeness centrality, due to the fact that they are based on the shortest path and in addition they can be configured to deal with attributes of the edges such as the weight. As regards other centrality measures available within **NetworkX** package, they were discarded due to limitation of our graphs. For instance, the eigenvector centrality fails to converge when there are multiple eigenvalues with the same magnitude.

5.2 Neighbors Approach

5.2.1 Experiments

In the neighbors approach, we considered the entire graph. In order ti obtain subsets of nodes, we performed a classification in two steps. First of all, we looked at the normalized ratio of the difference within the incoming edges and the outgoing edges over the total number of edges for each node:

$$c = \frac{\#incomingEdges - \#OutgoingEdges}{\#TotalEdges}$$

Since it is a normalized ratio, the values go from -1 to 1, where the value is negative for the nodes which have more outgoing edges than the incoming one, and it is positive for the dual situation. At the border, we have the extreme situations in which a user just sends messages or only receives messages.
We assumed a normal node has the same number of incoming and outgoing edges, so a ration equal to 0, or has pretty much the same number of ingress and egrees edges, so a ratio within 0.4 and 0.6. The table 5.1 collects the obtained results.

Classification	Ratio (c)	Number	Percentage
pure spammer	$c = -1$	32'592	16.59%
huge spammer	$-1 < c < -0.5$	726	0.37%
little spammer	$-0.5 <= c < 0$	19'992	10.17%
less requested	$0 < c < 0.4$	33'230	16.91%
normal	$c = 0$ or $(0.4 <= c <= 0.6)$	97'519	49.63%
more request	$0.6 < c < 1$	59	0.03%
only requested	$c = 1$	12'371	6.29%

Table 5.1: Classification based on the ratio c

The normal set represents the 49.16% of the nodes of the graph. We took into consideration the normal set. This first classification doesn't take into consideration the weight of the edges. Therefore, the nodes within the normal set are further splitted in 27 subsets, on the base of three attributes: the total number of the edges (simply the degree of a node, considering both incoming and outgoing edges), the message per link (computed as the ratio between the weighted degree and the degree of a node, where the weighted degree is the sum of the weights of each edges connected to a node) and the difference within the weight of the incoming edges and the weight of the outgoing edges. This classification is performed as a cascade: from the number of edges we obtained three subsets, each of them further splited into three more subsets on the base of the message per link and, in turn, splitted in three more subsets due to the value of the difference within the weight of the incoming

edges and the weight of the outgoing edges. The thresholds were chosen looking at the values of the three attributes and by trying to subdivide the nodes in the most possible equal way. Indeed, firstly we did an attempt by using the z-score (how many standard deviation a value is far from the mean). In this way, we computed the mean and the standard deviation for the three attributes and set the thresholds as multiple of the absolute value of the z-score (one, two and three). Following this choice, we obtained a subset containing 91'563 nodes over the 97'519 nodes of the normal set. That is the reason why we discarded this attempt. Therefore, we set the thresholds by testing different values and looking at how the nodes were splitting into the subsets. By reminding that the dataset are gathered over a period of a month and the chat rooms are open in the break within two games, we fixed the thresholds as described in the table 5.2 and it is also present the arbitrary names that we used.

	Threshold1	Threshold2	Threshold3
Number of edges	$x <= 6$	$x > 6$ and $x <= 12$	$x > 12$
-subset's name-	Small	Medium	Large
Messages per link	$x <= 20$	$x > 20$ and $x < 70$	$x >= 70$
-subset's name-	Casual	Normal	Active
Difference weighted I/O	$x <= 25$	$x > 25$ and $x < 50$	$x >= 50$
-subset's name-	Balanced	notsoBalanced	Unbalanced

Table 5.2: Classification based on the three attributes: number of edges, messages per link and differenced weighted I/O

For instance, a node with a degree of 14, messages per link equal to 25 and the difference within the weighted incoming edges and the outgoing equal to 24, will be present in the subset called *Small_Normal_Balanced*.

The table 5.3 shows how many nodes are present in each subsets.

We can see as the users tend to communicate with a restricted number of users. By increasing the number of edges, the participation of a user moves toward an unbalanced situation but at the same time he sends more messages. Since we don't have information regarding the login history in the rooms and the time when a message is sent, we cannot distinguish a user that joins often a room and sends few messages by a user that joins just one time a room and sends a lot of messages. This one can be an interesting feature to be investigated.

Subset's name	Number of nodes
Small_Casual_Balanced	26'236
Medium_Casual_Balanced	646
Large_Casual_Balanced	56
Small_Casual_notsoBalanced	1'713
Medium_Casual_notsoBalanced	269
Large_Casual_notsoBalanced	28
Small_Casual_Unbalanced	224
Medium_Casual_Unbalanced	140
Large_Casual_Unbalanced	48
Small_Normal_Balanced	16'303
Medium_Normal_Balanced	830
Large_Normal_Balanced	127
Small_Normal_notsoBalanced	6'460
Medium_Normal_notsoBalanced	696
Large_Normal_notsoBalanced	132
Small_Normal_Unbalanced	6'019
Medium_Normal_Unbalanced	1'838
Large_Normal_Unbalanced	700
Small_Active_Balanced	6'590
Medium_Active_Balanced	344
Large_Active_Balanced	112
Small_Active_notsoBalanced	4'948
Medium_Active_notsoBalanced	366
Large_Active_notsoBalanced	92
Small_Active_Unbalanced	16'052
Medium_Active_Unbalanced	4'292
Large_Active_Unbalanced	2'258

Table 5.3: Names of the subsets and the number of nodes within them

At this point, for each node in the subsets we obtained a subgraph by considering the neighbors of that node, so the adjacent nodes. Starting by the subgraphs we computed the betweenness centrality and closeness centrality and stored only the results corresponding to the nodes in our subsets. For both, we considered also the version with the weight of the edges as key value.

We organized the results of the centrality measures on the base of the length of the subgraphs (the number of nodes within them). The maximum length was 313 nodes, but we checked that the majority of the subgraphs was characterized by small size (less than 35 nodes). Therefore, we took into consideration the subgraphs with a length up to 43, in this way we discarded just 145 subgraphs over 52'720 and at the same time simplified the computation. We performed this classification in order to apply the z-score for finding the outliers, and so computing the mean and the standard deviation of the centrality measures with respect to the size of the subgraphs and always maintaining the subdivision illustrated in table 5.2. A node is supposed to be an outlier whether its value of a certain centrality is greater than three times the absolute value of the z-score. A node can be an outlier for more than one centrality measures. Within the normal nodes, few nodes resulted to be characterized by an abnormal behaviour from the mean. In addition, we distinguished the outliers through the category of centrality measures by considering all the combinations within betweenness, weighted betweenness, closeness and weighted closeness centrality. Such as, a node could have a highly different value of certain centrality measures while behaving normal with respect to the others, or, for instance, a node could be outliers for both closeness and weighted closeness centrality. Table 5.4 shows the results, highlighting for each subset the number of the outliers for each centrality measures, the total number of outliers within the subsets. Subsets without outliers are not present and only the centrality measures in which outliers were detected are recorded.

The subsets that disappeared from the study are the one characterized by a quite constant behaviour and/or an insufficient number of nodes to see appreciable behaviours far away by the mean. We can immediatly see that within the subsets, containing nodes characterized by similar attributes, outliers tend to behave different from the mean just in one category of centrality.

5.2.2 Result

Before looking at the subgraphs of the outliers, it is interesting to show the scatter plots of the centrality measures, underlining the outliers. Below, four scatter plots are provided. They show the values of the centrality measures computed for the nodes in the subsets, with respect to the size of the subgraph in which the node is present. In all of them, the blue squares represent all the values of a certain centrality, so the values from all the subsets. In order to keep the plots understandable, just the outliers are highlighted with a different color and a different shape on the base of the subset it comes from. For that reason, a blue square could be mistakenly seen by the reader as a "fake" outliers, but have in mind that the outliers are obtained with respect to the lenght of a subgraph and keeping the subsets' classification.

Subset's Name	onlyb	onlybW	onlyc	onlycW	c_cW	b_bW	cW_bW	Total
Small_Casual_Balanced	58	-	933	658	-	-	-	1'649
Small_Casual_notsoBalanced	33	-	39	49	1	-	-	122
Small_Casual_Unbalanced	11	-	4	6	-	-	-	21
Small_Normal_Balanced	-	-	263	299	-	-	-	562
Small_Normal_notsoBalanced	-	-	175	82	-	-	-	257
Small_Normal_Unbalanced	204	-	165	71	-	-	-	440
Small_Active_Balanced	-	-	36	34	1	-	-	78
Small_Active_notsoBalanced	-	-	29	25	-	-	-	54
Small_Active_Unbalanced	147	-	177	223	2	-	-	549
Medium_Casual_Balanced	-	-	17	16	-	-	-	33
Medium_Casual_notsoBalanced	-	-	10	6	-	-	-	16
Medium_Casual_Unbalanced	-	-	1	2	-	-	-	3
Medium_Normal_Balanced	-	-	19	17	-	-	-	36
Medium_Normal_notsoBalanced	-	-	17	12	-	-	-	29
Medium_Normal_Unbalanced	-	-	32	29	-	-	-	61
Medium_Active_Balanced	-	-	5	6	-	-	-	11
Medium_Active_notsoBalanced	-	-	18	10	-	-	-	18
Medium_Active_Unbalanced	-	-	159	75	1	-	-	235
Large_Casual_Balanced	-	-	1	1	-	-	-	2
Large_Normal_Balanced	-	-	-	3	-	-	-	3
Large_Normal_notsoBalanced	-	2	1	-	-	-	-	3
Large_Normal_Unbalanced	1	7	6	14	-	1	-	29
Large_Active_Balanced	-	-	2	1	-	-	-	3
Large_Active_notsoBalanced	-	-	2	1	-	-	-	3
Large_Active_Unbalanced	-	5	25	29	-	-	1	60

Table 5.4: Outliers of the subsets (b = outlier with respect only to betweenness centrality, bW = outlier with respect only to weighted betweenness centrality, c = outlier with respect only to closeness centrality, cW = outlier with respect only to weighted closeness centrality, c_cW = outlier with respect to closeness and weighted closeness centrality)

Figures 5.1 and 5.2 show the scatter plots of, respectively, the betweenness centrality and the weighted betweenness centrality. Since the values of these centrality measures vary a lot, we can easily understand the low number of outliers nodes (504 outliers with respect to the betweenness centrality and just 14 with respect the weighted betweenness within all the subsets). Despite the outliers' list of the betweeness centralities are sparsely populated, we still considered the threshold to be considered outliers (set as three standard deviations above/below by the mean) as a right compromise in order to detect abnormal behaviours. We can underline that the outliers with respect to the weighted betweenness are the nodes characterized by a low value.

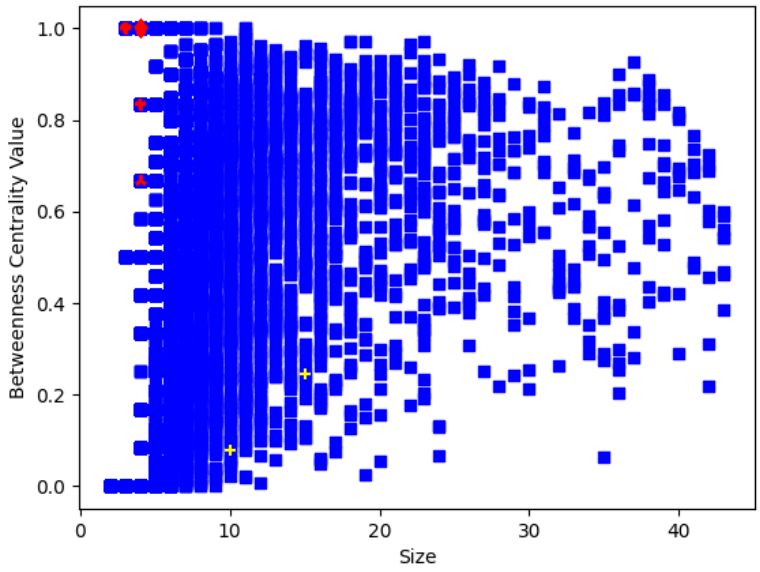


Figure 5.1: Scatter plot of the betweenness centrality. Blue squares represent all the values. Red symbols represent the outliers within the *small_x_x* subsets, yellow ones represent the outliers within the *large_x_x* subsets

Instead of, we can appreciate well-defined curves in the figures 5.3 and 5.4. They represent the closeness centrality and the weighted one obtained through the neighbors approach. They show a completely different behaviour, they almost seem to be dual. The reason is that closeness centrality is computed as the inverse of the sum of the shortest path that flows over a node, in contrast to the betweenness centrality which is based on a ratio. Therefore, having a high value of closeness centrality means that the node has close relationships with many other nodes, and we can see by the plot that the outliers are nodes with a low value. In case we consider the weight of the edges as key value, such as computing the weighted closeness centrality, the shortest path will be always considered in order to minimize the effort, so an high value of the shortest path (and so a low value for the centrality) could be interpreted in different ways in our graph: as a user that in the network doesn't have occasional conversations, so he speaks a lot with his close friends; otherwise as a stand alone quiet user. By considering the weighted closeness centrality, the outliers are the nodes with a good reachability.

For further investigation, we looked at the subgraphs. We created 25 subgraphs:

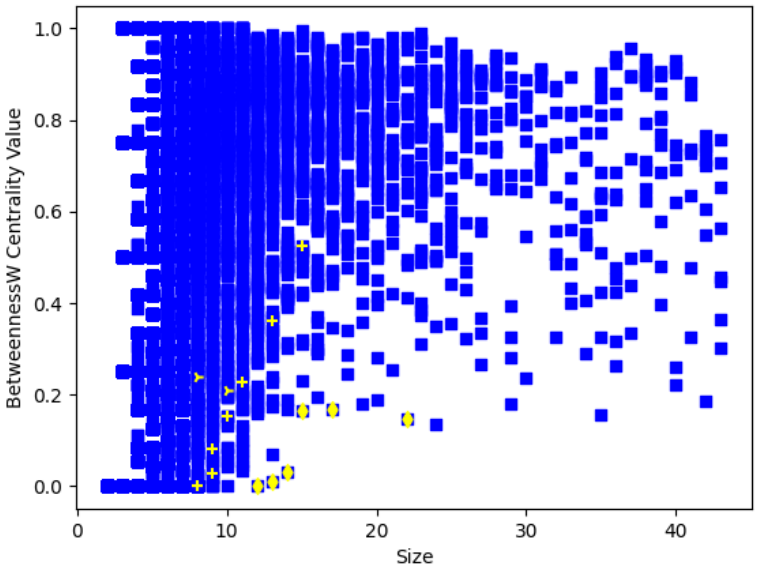


Figure 5.2: Scatter plot of the weighted betweenness centrality. Yellow symbols represent the outliers within the *large_x_x* subsets

one for each subset which recorded outliers, in order to see how outliers of the same subset behave. For the generation of the subgraphs we randomly picked 7 nodes within each category/subsets where it was possible, otherwise all the nodes. In the graphs, the outliers have different colors in order to be highlighted. Working with **Gephi** it was possible to visualize the graphs with different layouts. For our purpose the most interesting layout in order to obtain a pleasurable view to the eye was *Fruchterman Reingold* combined within *Noverlap*. Fruchterman Reingold layout is very similar to Openlord (used to draw the entire graph), it is useful to distinguish components in a network and it diverges by Openlord given the fact that it doesn't discard any edges. The layout can be combined with Noverlap layout in order to further reduce the overlapping of nodes and edges. All the graphs can be seen in the appendix A, here we report the most interesting case: the subset *Large_Normal_Unbalanced*, since it contains the different categories of centrality (Figure 5.5).

In the graph we can understand the roles of the outliers from the different categories of centrality. The thickness of an edge indicates the number of messages on it.

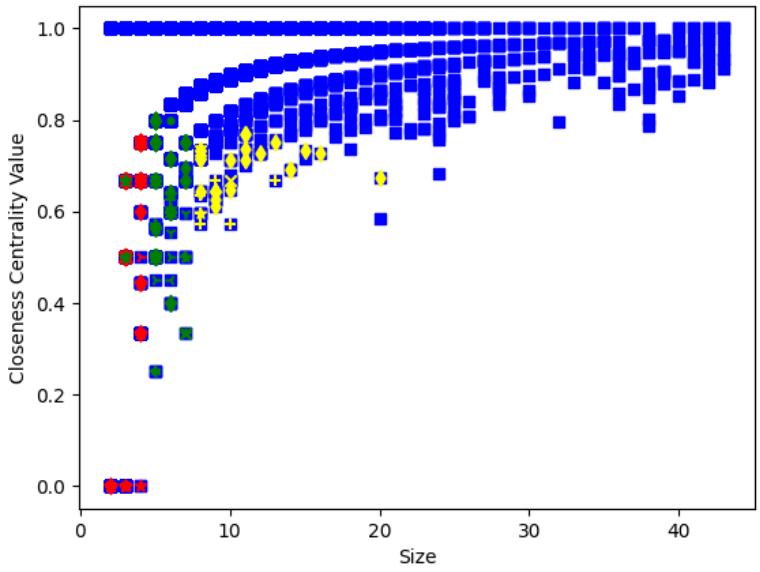


Figure 5.3: Scatter plot of the closeness centrality. Blue squares represent all the values, red symbols represent the outliers within the *small_x_x* subsets, the green symsbols the outliers within the *medium_x_x* subsets.

Outliers with respect betweenness, weighted betweenness and closeness centrality have a low value of centrality. Indeed in the component in which they are present, they don't play a central role. For instance, we can have a look at the bottom right of the figure: the red node (outlier with respect to betweenness centrality) has a lot of connections but it is not an important nodes within its neighbors. We can see that the edges around that node have higher weights.
While the outliers with respect to the weighted closeness centrality (characterized by a high value) are well connected and link different components of the graph.

5.3 Cliques Approach

5.3.1 Experiments

On the other hand, to get a different view of the graph we took advantage of the cliques, assuming a clique the most realistic way to represent a room within the online game. Through the *find cliques* function implemented in **NetworkX** applied to the undirected graph, we were able to extract 325'338 cliques. The size of the cliques goes from 2 to 23 and the number of cliques of reduced size are the majority. It is a

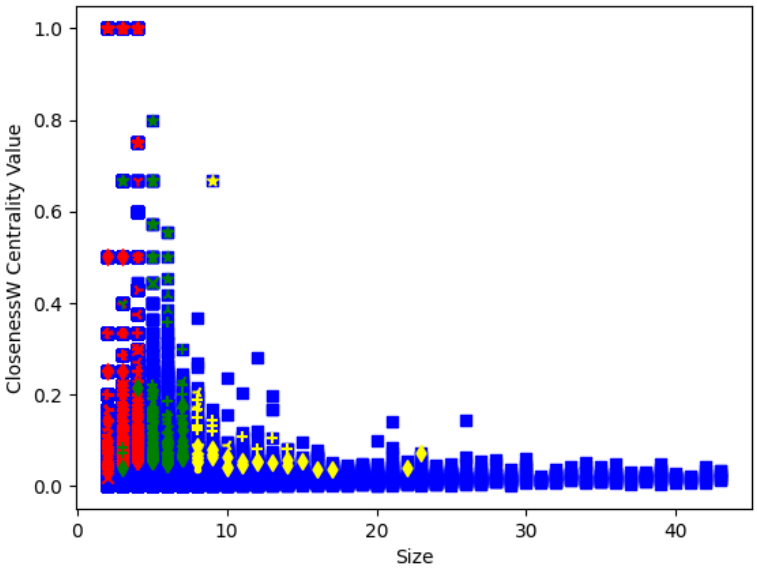


Figure 5.4: Scatter plot of the weighted closeness centrality. Blue squares represent all the values, red symbols represent the outliers within the *small_x_x* subsets, the green symsbols the outliers within the *medium_x_x* subsets.

sign that normal users tend to play with a restricted circle of friends. The smallest cliques (size equal to two) are not so interesting but still taken into consideration. A list of the size of the clique with its respective frequency is reported in table 5.5.

We considered all the cliques in order to find the outliers. We performed the betweenness, the closeness centrality and also the weighted versions with respect to the single clique. Once we got the results we organized them on the base of the size of the cliques. We computed the z-score with respect to the size of the cliques and set the threshold equal to three times the z-score.

Therefore to proceed, we subdivided the outliers on the base of the centrality's category and the ratio c, the same used in the neighbor approach.
Before showing the categories, we want to stress the percentage of outliers within the "pure" centrality measures. By analizing all the nodes within the cliques, we studied almost 1'380'000 nodes. The outliers with respect the closeness centrality were 35'007 (2.5%); the outliers with respect to the weighted closeness centrality were 23'484

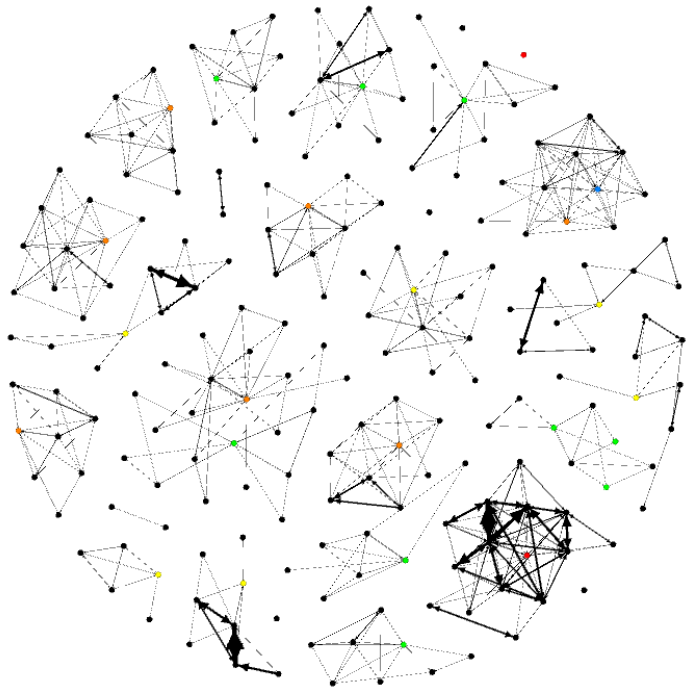


Figure 5.5: Graph representing the subset *Large_Normal_Unbalanced*. Red nodes are the outliers with respect the betweenness centrality, orange nodes the ones with respect the weigheted betweenness centrality, the yellow nodes the ones with respect to the closeness centrality, the green nodes the ones with respect to the weighted closeness centrality and the blue nodes with respect the betweenness and the weighted betweenness centrality.

(1.7%); the outliers with respect to the betweenness centrality were 37'267 (2.7%); the outliers with respect to the weighted betweenness centrality were 5'613(0.4%).

The table 5.6 shows the outliers with respect all possible combinations of centrality and the ratio c. Combinations not present in the table are empty.

From the table 5.6, we can see that the outliers from the "only" centrality (only betweennes centrality, only weighted betweenness and only weighted closeness

Clique list					
Size	Number	Size	Number	Size	Number
2	90'301	3	65'681	4	57'475
5	33'848	6	26'895	7	20'646
8	12'877	9	7'023	10	3'444
11	1'648	12	929	13	798
14	718	15	602	16	667
17	598	18	451	19	410
20	194	21	91	22	35
23	7	-	-	-	-
Total number of cliques : 325'338					

Table 5.5: Number of cliques with respect to the size of them

Centrality's category	Number of outliers	Normal	Less requested	More requested	Only requested	Little spammer	Huge spammer	Pure Spammer
onlyb	13'664	29.7%	56.99%	0.01%	0%	13.80%	0.02%	0%
onlyc	17'124	1.16%	1.05%	0%	0%	24.99%	6.35 %	66.45%
onlycW	7'680	41.65%	39.28%	0.07%	9.78%	9.18%	0.04%	0%
onlybW	451	31.93%	53.88%	0%	0%	20.84%	0%	0%
b_c	3'091	7.60%	47.20%	0%	%	44.52%	0.68%	0%
b_cW	4'093	10.09%	80.58%	0%	0%	9.33%	0%	0%
b_bW	1'023	4.20%	90.13%	0%	0%	5.67%	0%	0%
cW_c	725	4.69%	8.41%	0%	0%	79.45%	7.45%	0%
cW_bW	648	9.88%	24.85%	0%	0%	43.83%	21.45%	0%
b_bW_c	882	4.19%	72.11%	0%	0%	3.00%	0%	0%
b_bW_cW	3'133	0.77%	96.23%	0%	0%	3.00%	0%	0%
b_cW_c	3'089	3.92%	61.57%	0%	0%	34.51%	0%	0%
c_bW_cW	156	4.49%	17.95%	0%	0%	77.56%	0%	0%
b_c_bW_cW	7'055	3.94%	68.49%	0%	0%	27.57%	0%	0%

Table 5.6: Number of outliers with respect to the category of centrality and the ratio c

centrality) tend to have more or less the same number of incoming and outgoing edges; except for the outliers with respect only the closeness centrality, which tend to have more outgoing edges than incoming ones, up to the borderline case of the *pure_spammer*. Instead, if we look at the outliers for more than one centrality, the number of normal nodes within them decreases and the majority of the nodes goes toward a *less_requested* or *little_spammer* situation.

5.3.2 Result

As done for the neighbor approach, we looked at the scatter plots and then at subgraphs. Figure 5.6 shows the scatter plot of the betweenness centrality against the scatter plot of the weighted version. As a difference from the neighbor approach, We can observe a well defined behaviour also for the betweenness centrality . In the

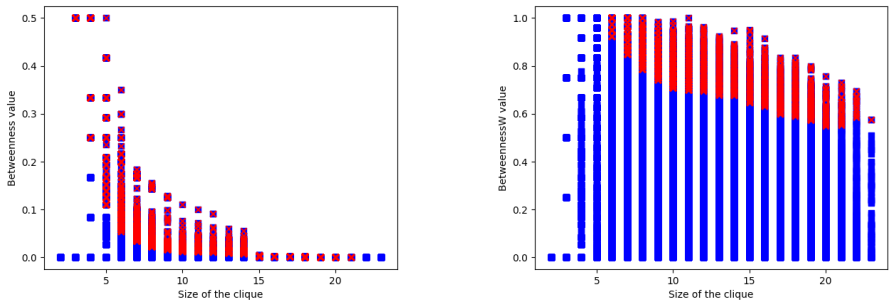


Figure 5.6: Scatter plot of the betweenness centralities. At the left the betweenness centrality. At the right the weighted betweenness centrality. Blue squares represent all the values, red squares represent the outliers.

cliques approach the outliers with respect to the two kind of betweenness are clearly defined as the node with an high value. So they can be interpreted as the bridges between rooms.
Figure 5.7 shows the closeness centralies. Since the formula of closeness centrality, we can make the same considerations as in the neighbor approach.

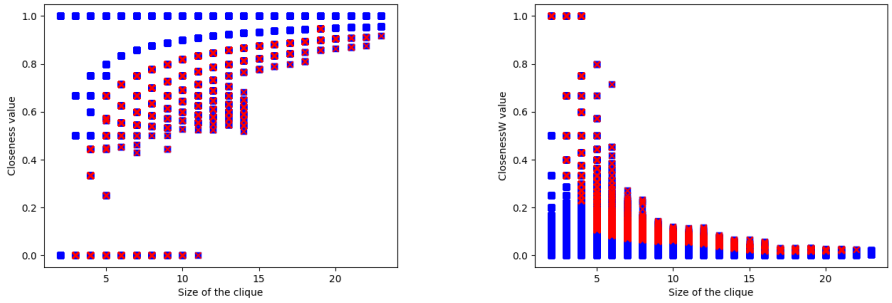


Figure 5.7: Scatter plot of the closeness centralities. At the left the closeness centrality. At the right the weighted closeness centrality. Blue squares represent all the values, red squares represent the outliers.

With reference to table 5.6, we generated subgraphs in order to obtain frames of the network. For the creation of the subgraphs, we randomly picked 5 nodes within each category/subsets where it was possible, otherwise all the nodes. In contrast to the neighbor approach, here we obtained the subgraphs through the composition of

the cliques in which a node were present.
All the result are provided in the appendix B.

Figure 5.8 shows the most interesting case. It is the graph with the outliers with respect to betweenness and weighted betweenness centrality. These kind of outliers are characterized by having a high value of centrality.
If we focus on the left part of the figure, it is present a dense component of the graph in which we can interprete the role of the outliers. Red nodes from the *normal* subset play a central role in the clique/room in which they are present. The green nodes (from the *less_requested* subset) are in the middle of the more dense cliques, they can be interpreted as users that play in different rooms but without writing so much. While the purple node (*little_spammer*) are in the proximity of the more dense cliques, therefore they can be interpreted as node trying to enter to “big rooms”.

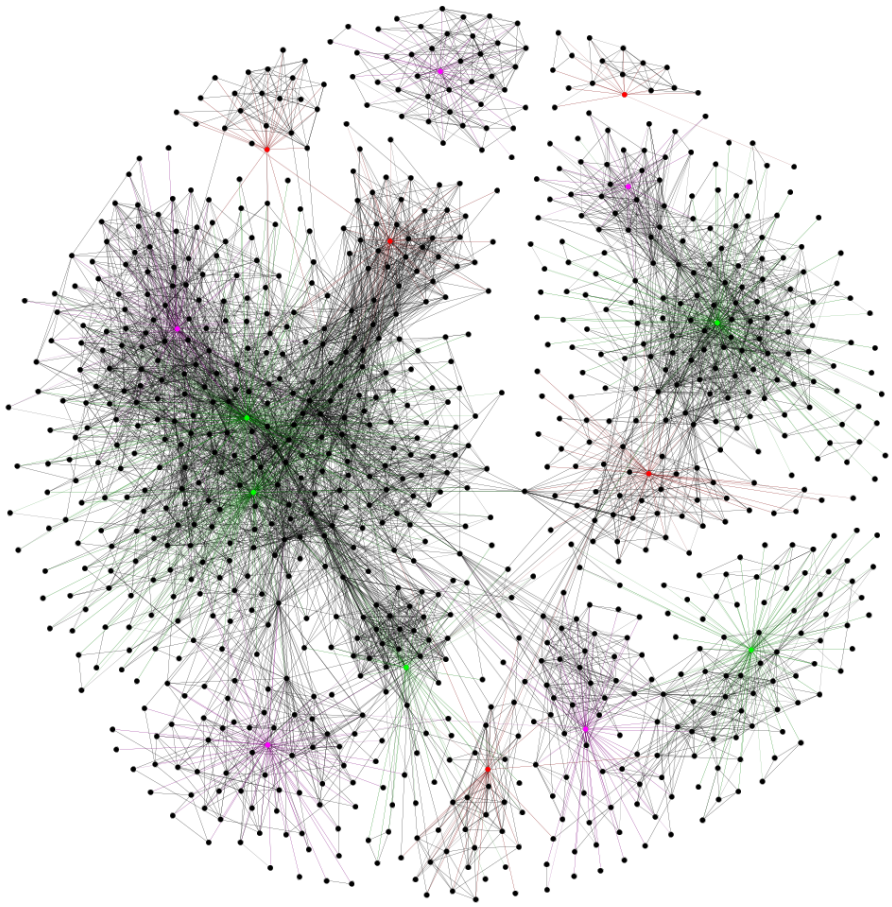


Figure 5.8: Graph representing the outliers with respect to betweenness and weighted betweenness centrality. Red nodes are outliers from the **normal** subsets, green nodes the ones from the *less_requested* and the purple nodes the ones from the *litlle_spammer*.

Chapter 6
Discussion

Through the previous chapters we have presented background knowledgement and related work to our research. In addition, the methodologies and the experiments are shown as well as the results. This chapter provides discussions about the experiments and the results and then limitations will be reported. The initial main object of the thesis was to test features of the social network analysis, in particular taking advantage of the graph theory, in order to find possible indicators for detecting cyber sexual predators. Due to the lack of a dataset containing chatlogs by cyber sexual predators, the object moved to the detection of abnormal behaviours.

6.1 General Discussion

In order to find out abnormal behaviours in our dataset, we followed two different paths. With the neighbors approach we wanted to have a local view of the network by looking at the adjacent nodes and creating subsets of nodes with similar characteristics. On the other hand, through the cliques approach we aimed to recreate the room within the games and see how users behave.
To see the role/position a node occupied in the network, we took advantage of the centrality measures: the betweenness centrality and the closeness centrality. For both two, there were possible to implement the weight of the edges as key values. Betweenness centrality is used to find nodes that connect two distinct part of the graph, nodes that act as bridges. Instead of, closeness centrality is a way to detect nodes which can efficiently spread information within a graph.
While trough the z-score we were able to detect outliers (nodes that behave very different by the mean). It is important to stress that the total number of nodes taken into consideration in the two approaches were completely different (90'000 nodes against 1'400'000). But in both cases the outliers were in the order of 2-3% of the total nodes.

6.1.1 Neighbors Approach

We performed a two levels classification: the first one based on the normalized difference between incoming and outgoing edges; and the second one based on three attributes (the total number of edges, the messages per link and the variation within the number of incoming messages and the number of outgoing messages). In such a way, due to the first classification we divided the nodes in seven sets. We considered the most populated set: the normal one. Therefore, we further splitted the nodes in twenty-seven subsets.
We have seen that usually users tend to communicate with a close number of friend. Indeed the majority of the nodes were in the subsets *small_x_x*. By increasing the number of connections and actively participation of a user, nodes moved toward an unbalanced situation.
By considering the neighbors of a node, we created subgraphs of the entire graph and performed the centrality measures. By organizing the subgraphs on the base of the number of nodes within them, we were able to detect the outliers for every subsets.

Through this approach (by splitting nodes into subsets, and so by creating subsets containing nodes with similar attributes), we have seen that outliers were such that they differed from common nodes just by one category of centrality.
The outliers with respect to the betweenness and the weighted betweenness centrality were characterized by a low value of the centrality measures. It means nodes that don't play for sure the role of bridges. The values were very sparse.
While as regard the other kind of centrality, the scatter plots of the closeness and the weighted closeness centrality had totally different (and well-defined) shape. The outliers with respect to the closeness centrality were the nodes with a a low value of centrality, so no well-connected nodes. Instead of, the outliers with respect to the weighted closeness centrality were the nodes characterized by a high value of centrality, so nodes with a high reachability. This difference is explained due to the formula of the closeness centrality: it is not a ratio as the one for the betweenness centrality but it is the inverse of the sum of the shortest paths that flow through a node. A low value of this centrality can be interpreted in different ways: as a user who exchanges a huge amount of messages but with a restricted group of friends, otherwise as a quiet stand alone user. Despite that, a high value represents a node which plays a center and dominant role within the network.
We have checked our hypothesis by looking at the visualization of the subgraphs thanks to Gephi.

Therefore, by looking at the proximity of a node we have been able to check whether the node in question behaved different from its neighbors. We have seen

that through the centrality measures it is possible to distinguish different kind of outliers: the ones that occupy a more central position and the ones which play the role of extras. For a deeper view of the behaviours of the nodes, a more complete knowledgement of the network is required. It is important to stress that betweenness centrality is not so useful in small simple graphs. Indeed the outliers with respect this kind of centrality mainly from the *large_x_x* subsets, so graphs with more than twelve nodes.

6.1.2 Cliques Approach

In this approach we assumed cliques as the most realistic way to represent the chat rooms of the game. We classified the cliques due to the their size and performed the centrality measures (the same as in the neighbors approach). At this point, we computed the z-score in order to find the outliers. Then we subdivided the outliers with respect to the category of centrality for which the outliers behave different as well as with respect to the normalized difference between incoming and outgoing edges. By the distribution of the cliques (so the number of the cliques with respect to their size), we have seen a confirmation on the fact that users tend to communicate with a close number of other users.
From the scatter plots, as regard the closeness and the weighted closeness centrality we have taken the same conclusion as for the neighbors approach: outliers from the closeness centrality were nodes not easily reachable, while outliers from the weighted closeness centrality represented the dual situation. While for the betweenness and the weighted betweenness centrality, well-defined shapes can be observed. The outliers were the nodes with a high value, so the bridges between different cliques/chat rooms. In addition, as difference from the neighbors approach, outliers were such they diverged from the mean with respect to more than one category of centrality. Indeed, by putting the outliers in relation to the degrees of the nodes (so by splitting the outliers within sets due to the ratio c) we have been able to see that: the outliers with respect to the closeness centrality, characterized by a low value of centrality, tend to have more outgoing edges than incoming ones; the outliers with respect to the betweenness, weighted betweenness and weighted closeness, characterized by a high value of centrality, tend to have the similar number of incoming and outgoing edges or more incoming edges than outgoing ones; while if we consider outliers with respect to combinations of centralities, the number of incoming and outgoing edges is less balanced.

We created subgraphs by randomly picking nodes within this classification and by composing the cliques in which that mentioned nodes were present. Despite assuming cliques as rooms and by taking advantage of the visual representations of

the subgraphs, we have been able to provide an interpretation of the outliers with respect to the degree of them.
First of all, we have noticed that the betweenness and the weighted betweenness centrality are the most suitable as measures to detect nodes which put in connection different cliques. While the closeness centrality may be interesting for detecting abnormal users like bots which spam messages.
We have seen that the outliers classified within the *less_requested* set are in the middle of the more dense cliques and they can be interpreted as users who play in different rooms without writing so much. While the ones classified in *little_spammer* set tend to be around the more dense cliques. A possible interpretation is that they represent users who try to participate to the "big room". Instead of, outliers with more or less the same number of incoming and outgoing edges have a more central position in cliques characterized by a limited number of nodes.

6.2 Limitation

Starting by the state of art related to cyber sexual predators, limitations came up. First of all, there is lack of real chatlogs between paedophiles and children to public domain. The work that PJ was doing contributes a lot to the research as well as arresting cyber predators, but the chatlogs provided in this way are conversations between adults pretending to be children and predators. They are still a good starting point to recognize typical chat behaviours of cyberpredators and so to detect these kind of behaviours in other chats.

In addition, getting hold of online public chats (as the ones in public rooms of an online game) is quite hard for privacy reasons. The dataset we worked with was from an online platform of games. All sensitive information were encrypted and even more, we didn't know much about the platform. We weren't interested of the content of the messages, but we performed assumptions and hypothesis in order to be able to proceed.
More information regarding how the public rooms work can easily improve the results as well as the explanation of the results. Moreover, even know more information regarding the game can be very useful. For instance the type of the game and so the different roles within the game or the level of the players can be useful to understand the position of a node in the network: in a Role Playing Game a leader of a faction might occupy a more central role than an lonely knight, while a trader might have a lot of connections but with small weights. On the other hand, by including in the dataset information regarding the login history can detect passive users otherwise it is not possible to unmask them.
Therefore, the knowledge of these kind of information reduces the false positives.

Other limitations are strictly related to the computation. We worked on a dataset gathered over a period of time of a month, such that a relative short amount of time. Despite that, the entire graph enclosed a huge quantity of information (almost 200'000 nodes and 750'000 edges). Due to the facts that the time complexity of the betweenness and closeness centrality is on the order of $\mathcal{O}(N * E + N^2)$ (so performing these kind of measures is very demanding) and that the graph was not connected, we had to work with subsets and restricted our study in the neighbors approach to the normal set.

Chapter 7
Conclusion

In this chapter conclusion will be provided. Then, based on our experiments and results, we will propose further features interesting to be investigated.

7.1 Conclusion

More and more children have access to the Internet. Therefore they are increasingly in danger. Through the Internet a cyber sexual predator can easily camouflage its identity or, more simply, mislead a child in order to satisfy its sexual dream.
In the background chapter we have analyzed the phenomenon of cyber grooming, stressing the attention on typical behaviours of cyber sexual predators and the importance of building a trust relationship in the process. Then we have moved to present the features of the graph theory.

Due to the lack of a dataset containing cyber predators conversations and so their pattern, the main object of the thesis moved to test key features for abnormal behaviours detection investigating the number of edges and the number of the messages.
In the thesis we provided two approaches: a neighbors approach based on the creation of subsets and so subgraphs, that can be seen as zoomed in frames of the entire network; and a cliques based approach, in which we used cliques to represent the public chat rooms of the game. In both, we have seen that a simple mathematical tool as the z-score can be very useful for detecting outliers. We have stressed our study in outliers three standard deviation above the mean. But by knowing better the network/possible patterns of the users, the threshold can be fit to highlight different features.
In the neighbors approach we have been able to get local views of the entire network. We have checked that closeness centrality is more meaningful than the betweenness

one whether we take into consideration small and simple graph. By switching into the different kind of centralities, we have been able to detect central nodes or more "suburban" nodes.
On the other hand, the cliques approach best suite as a global approach. The betweenness centrality is able to detect nodes that are present in more than one cliques, while the closeness centrality can find out users characterized by more outgoing edges than incoming ones. In addition, trough graphical representations of the cliques, we have been able to give an interpretation of the outliers with respect to their classifications.

By focusing in public chat rooms, our work shows that detection of users that have abnormal behaviours is possible even without looking at the subjects and the words of a conversation. But at the same time a good knowledgement of the network is required in order to investigate in a local point of view (neighbors approach). While looking at a more global view (so looking at more than one public chat rooms) offers a more complete sight with the possibility of drawing realistic conclusions about the patterns of the users.

7.2 Further Work

We tested features of graph theory with respect to a dataset without evil presence. As future work (in absence of a real dataset containing chatlogs of cyber predators) it can be interesting to manually add in such a network artificial nodes that behave as a cyber sexual predator in order to test for which features they can be defined as outliers.
Cyber sexual predator can be considered as central node or bridge between public rooms since, before building a relationship with a child. They might "cast the hooks" to several potential victims, so more outgoing edges than incoming and probably with an higher weight.
Moreover, the synthesis of these kind of nodes can be improved by manually analyzing real chatlogs between children and cyber predators or the ones provided by PJ.
We expect that cyber sexual predators can be classified as passive users at the beginning of their hunt. They might use to just read the conversations in a chat room in order to check the situation (how many adults, how many minors). Therefore adding information in the dataset as the login history in the various rooms can be an important features to be taken into consideration. This kind of information can be used to create dynamic networks otherwise it can be implemented as label of the nodes or the edges.

Furhter investigating this topic will allow online platform to implement automatic algorithms which are able to detect abnormal behaviours of users. Algorithms such that are based on basic information as the number of edges and the number of the messages. Therefore, they don't need to go against any privacy issues. Once suspicious user is identified, a human moderator can be warned. Human moderator could then also take into account the actual content of the conversations the user is involved in. Whether he recognize that the behaviour of the user is not a false positive, that user can be reported to the law enforcement.
The algorithm would be helpful in reducing the number of nodes the moderator has to investigate. Moreover, it would be set not only for cyber sexual predators detection, but rather against all evil presence that could be found in the Internet.

References

[AA14] Tarique Anwar and Muhammad Abulaish. A social graph based text mining framework for chat log investigation. *Digital Investigation*, 11(4):349–362, December 2014.

[AEMH] Cano Amparo E., Fernandez Miriam, and Alani Harith. Detecting child grooming behaviour patterns on social media. *Knowledge Media Institute, Open University, UK.*

[BK19] Patrick Bours and Halvor Kulsrud. Detection of Cyber Grooming in Online Conversation. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, Delft, Netherlands, December 2019. IEEE.

[BM] J. A. Bondy and U. S. R. Murty. Graph theory with applications. *Department of Combinatorics and Opimization, University of Waterloo, Ontario, Canada.*

[CFA] Amparo Elizabeth Cano, Miriam Fernandez, and Harith Alani. Detecting Child Grooming Behaviour Patterns on Social Media. page 16.

[Chia] Cyber grooming. https://www.childsafenet.org/new-page-15.

[Chib] Statistics about grooming and online predators. https://childsafety.losangelescriminallawyer.pro/children-and-grooming-online-predators.html.

[EC10] Micha Elsner and Eugene Charniak. Disentangling Chat. *Computational Linguistics*, 36(3):389–409, September 2010.

[Esh] Reddy Eshwitha. Proof that subgraph isomorphism problem is np-complete.

[Fre] L. C. Freeman. Centrality in social networks conceptual clarification.

[gep] Gephi documentation. https://gephi.org/.

[GKS12] Aditi Gupta, Ponnurangam Kumaraguru, and Ashish Sureka. Characterizing Pedophile Conversations on the Internet using Online Grooming. *arXiv:1208.4324 [cs]*, August 2012. arXiv: 1208.4324.

[GR] Brown George R. Paedophilic disorder. *MSD Manual*.

[Har] F. Harary. The graph theory.

[IFD12] Farkhund Iqbal, Benjamin C.M. Fung, and Mourad Debbabi. Mining Criminal Networks from Chat Log. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pages 332–337, Macau, China, December 2012. IEEE.

[Ind] Indika. Difference between graph and digraph.

[Joh] John William Johnsen. Algorithms and methods of organised cybercrime analysis.

[KCAC08] Tayfun Kucukyilmaz, B. Barla Cambazoglu, Cevdet Aykanat, and Fazli Can. Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management*, 44(4):1448–1466, July 2008.

[Kon] D. Konig. Graph theory.

[LNJLBLTKK] Olson Loreen N, Daggs Joy L., Ellevold Barbara L., and Rogers Teddy K. K. Entrapping the innocent: Toward a theory of child sexual predators' luring communication. *Communication Theory*, 17(1):231–251.

[LPLC10] Yong Lu, Michael Polgar, Xin Luo, and Yuanyuan Cao. Social Network Analysis of a Criminal Hacker Community. *Journal of Computer Information Systems*, page 13, 2010.

[MB] V. Mastronardi and A. M. Bonura. Criminological profile of pedophile. *Revista de estudios de crimonolia y ciencias penales*.

[Mer] Definition of pedophilia. https://www.merriam-webster.com/dictionary/pedophilia.

[Net] Networkxx documentation. https://networkx.org/documentation/networkx-2.5/.

[O'C] Rachel O'Connell. A typology of child cybersexploitation and online grooming practices. *Cyberspace Research Unit, University of Central Lancashire, UK*.

[Pan] Pandas documentation. https://pandas.pydata.org/pandas-docs/stable/index.html#module-pandas.

[POC09] Pietro Panzarasa, Tore Opsahl, and Kathleen M. Carley. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology*, 60(5):911–932, May 2009.

[Rud21] Joseph Rudman. The State of Authorship Attribution Studies: Some Problems and Solutions. page 16, 2021.

[SS] Alan P Schmidt and Trevor K M Stone. Detection of Topic Change in IRC Chat Logs. page 11.

[Syl] J. J. Sylvester. Nature.

[SYSC06] Dou Shen, Qiang Yang, Jian-Tao Sun, and Zheng Chen. Thread detection in dynamic text message streams. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, page 35, Seattle, Washington, USA, 2006. ACM Press.

[Teo] Paoletti Teo. Leonard euler's solution to the konigsber bridge problem. *College of New Jersey*.

[TL] Mario Gerla Taun Le, Haik Kalantarian. A novel social contact graph-based routing strategy for workload and throughput fairness in dealy tolerant network.

[UPdV14] Editorial Universitat Politècnica de València. Universitat Politècnica de València. *Ingeniería del agua*, 18(1):ix, September 2014.

[Wil] Robin J. Wilson. History of graph theory from: Handbook of graph theory.

Appendix A

Neighbors Approach's Graphs

In this section of the appendix, the graph we obtained in the neighbors approach will be provided.
The red node indicates outliers with respect to the betweenness centrality, orange nodes the ones with respect the weigheted betweenness centrality, the yellow nodes the ones with respect to the closeness centrality, the green nodes the ones with respect to the weighted closeness centrality, the blue nodes with respect the betweenness and the weighted betweenness centrality, dark blue indicates outliers with respect weighted closeness and weighted betweenness centrality and light blue the ones for closeness and weighted closeness centrality.

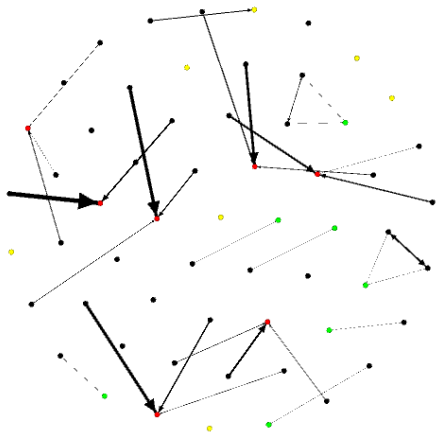


Figure A.1: Graph representing the subset *Small_Casual_Balanced*

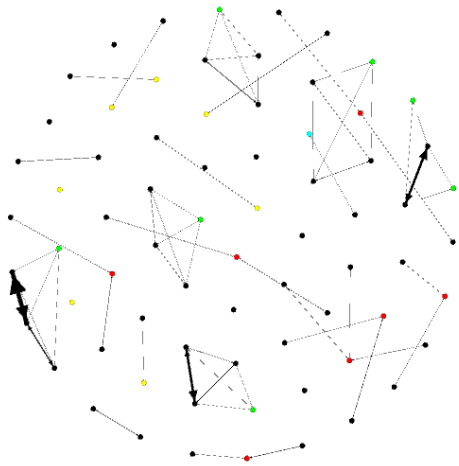


Figure A.2: Graph representing the subset *Small_Casual_notsoBalanced*

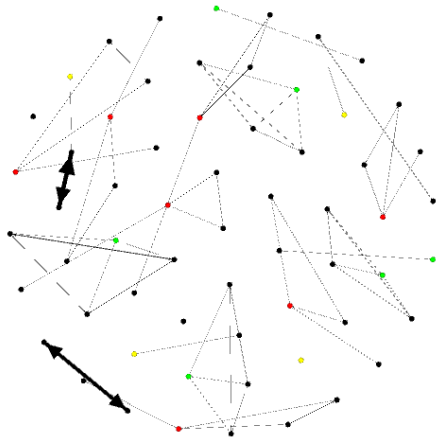


Figure A.3: Graph representing the subset *Small_Casual_Unbalanced*

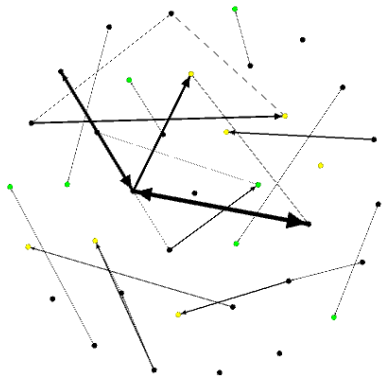


Figure A.4: Graph representing the subset *Small_Normal_Balanced*

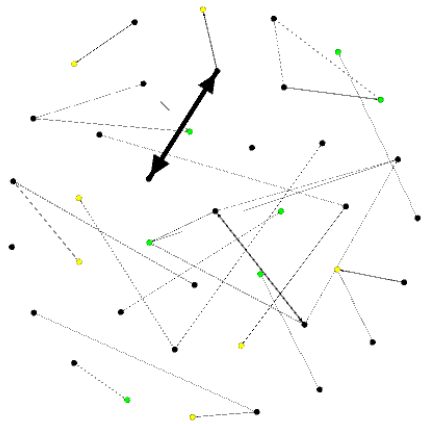


Figure A.5: Graph representing the subset *Small_Normal_notsoBalanced*

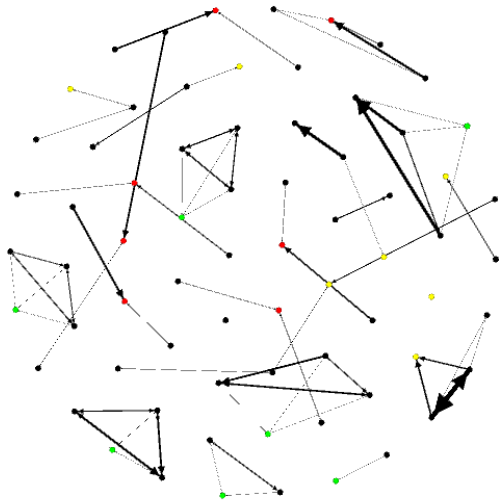


Figure A.6: Graph representing the subset *Small_Normal_Unbalanced*

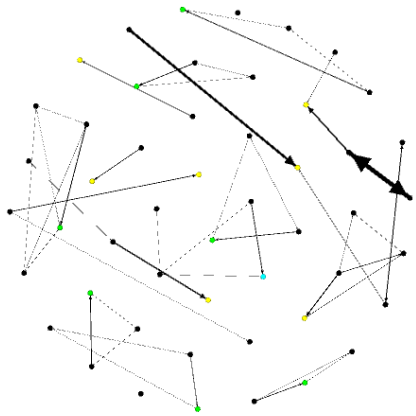


Figure A.7: Graph representing the subset *Small_Active_Balanced*

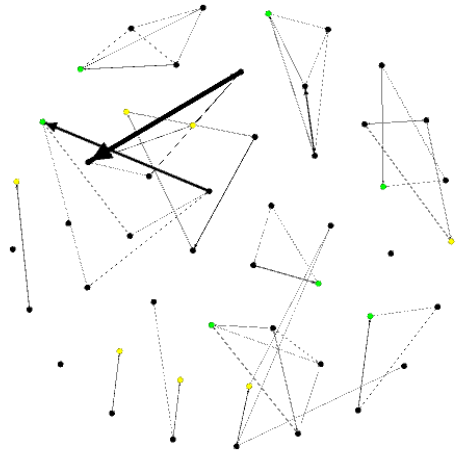


Figure A.8: Graph representing the subset *Small_Active_notsoBalanced*

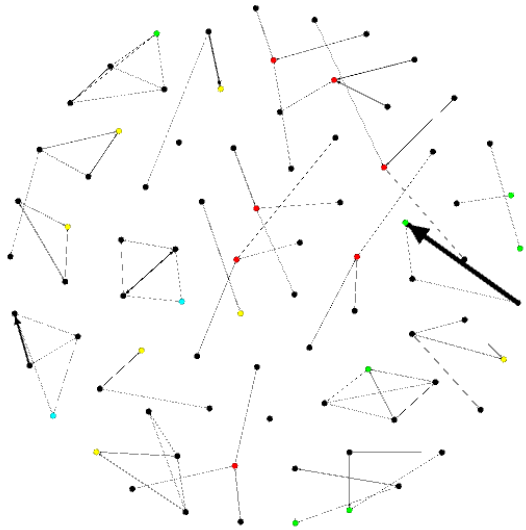


Figure A.9: Graph representing the subset *Small_Active_Unbalanced*

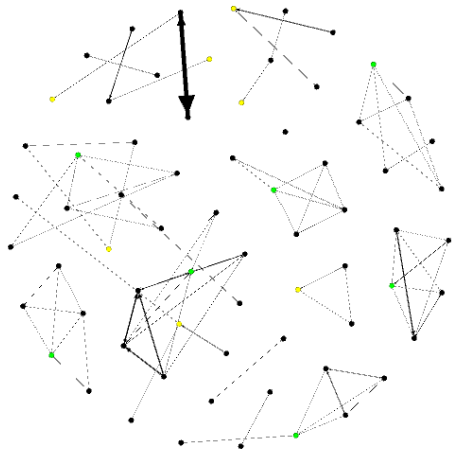


Figure A.10: Graph representing the subset *Medium_Casual_Balanced*

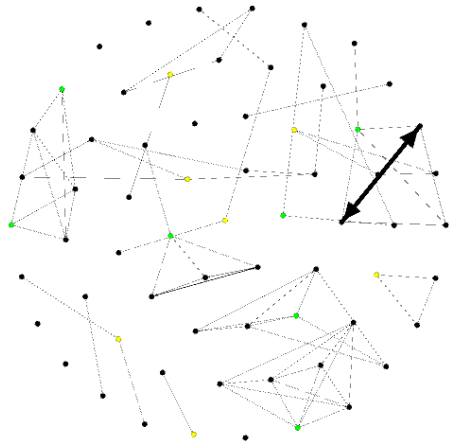


Figure A.11: Graph representing the subset *Medium_Casual_notsoBalanced*

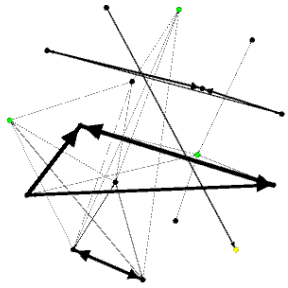


Figure A.12: Graph representing the subset *Medium_Casual_Unbalanced*

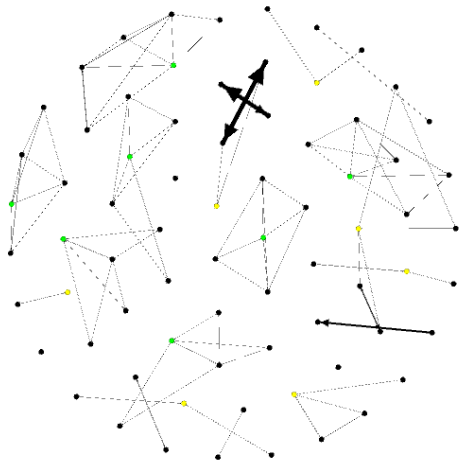


Figure A.13: Graph representing the subset *Medium_Normal_Balanced*

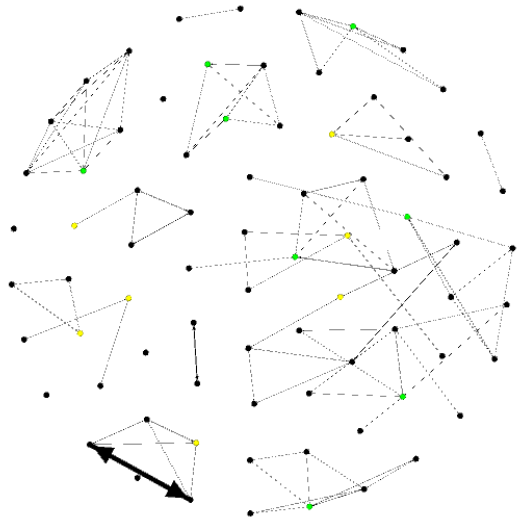


Figure A.14: Graph representing the subset *Medium_Normal_notsoBalanced*

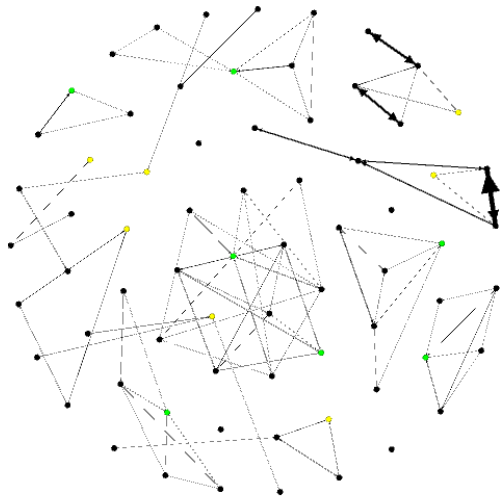


Figure A.15: Graph representing the subset *Medium_Normal_Unbalanced*

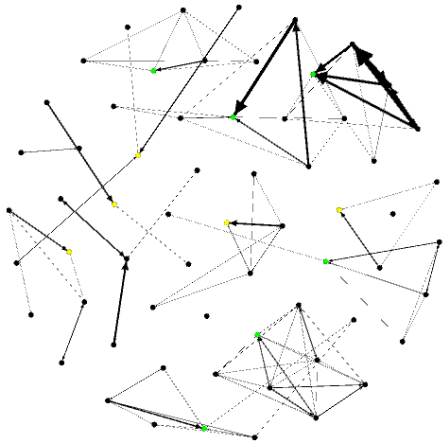


Figure A.16: Graph representing the subset *Medium_Active_Balanced*

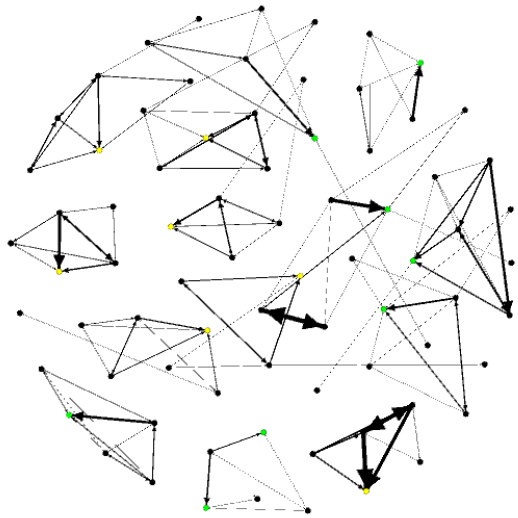


Figure A.17: Graph representing the subset *Medium_Active_notsoBalanced*

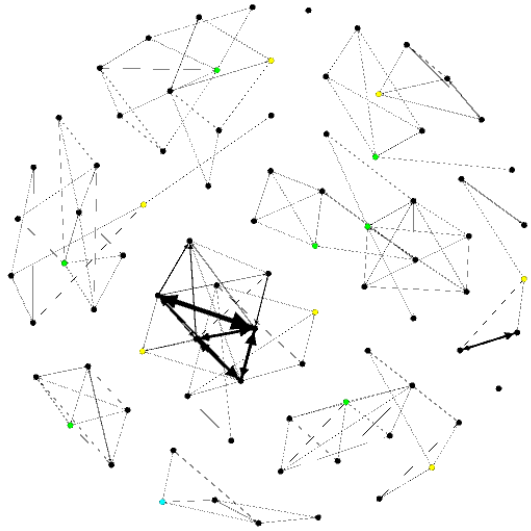


Figure A.18: Graph representing the subset *Medium_Active_Unbalanced*

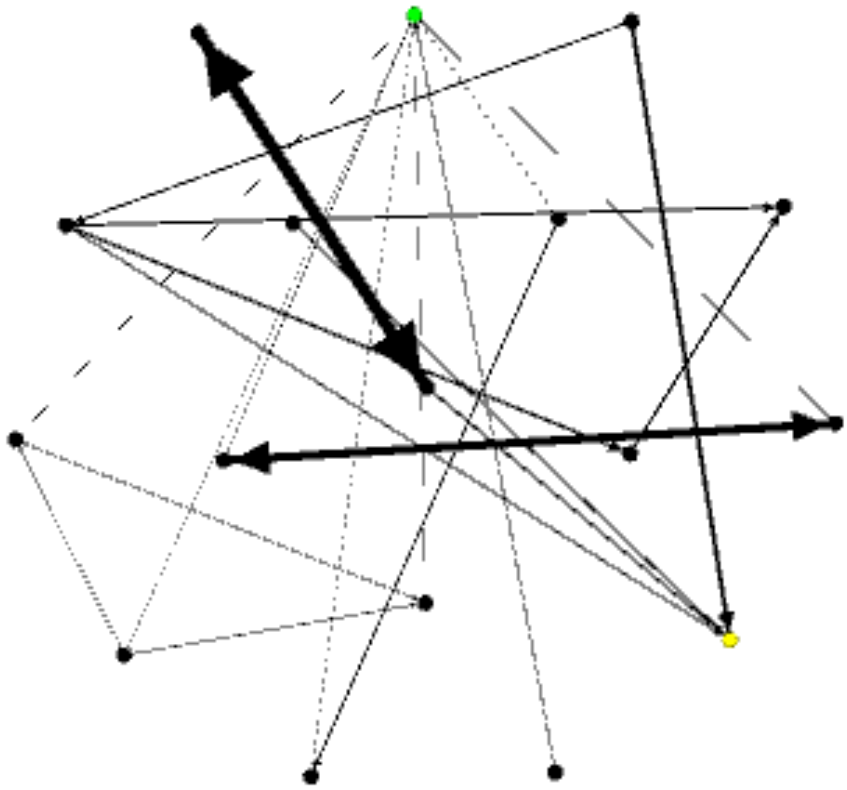


Figure A.19: Graph representing the subset *Large_Casual_Balanced*

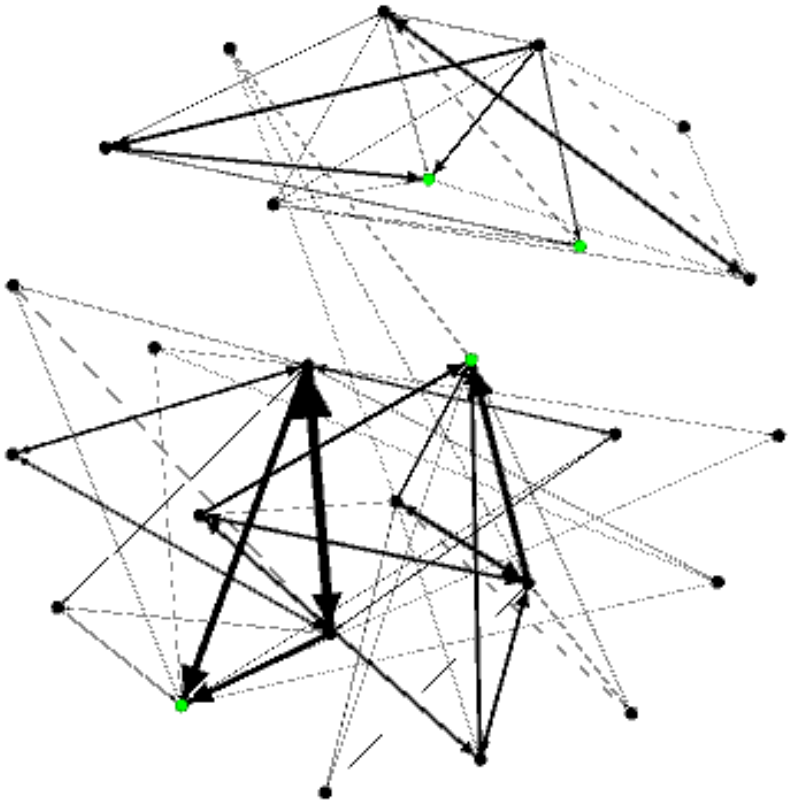


Figure A.20: Graph representing the subset *Large_Normal_Balanced*

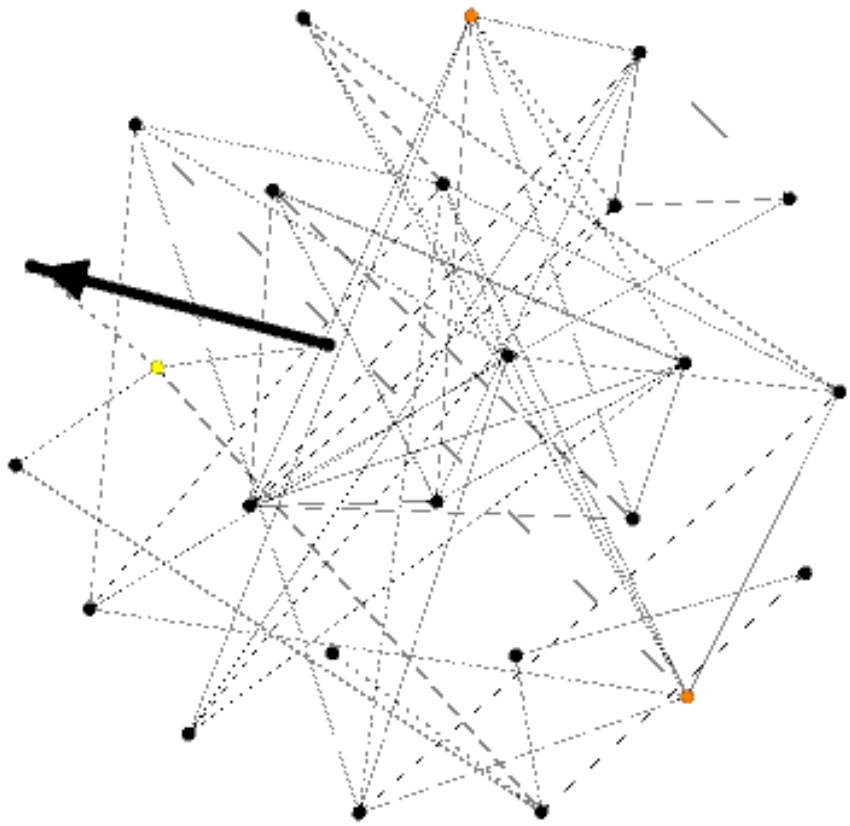


Figure A.21: Graph representing the subset *Large_Normal_notsoBalanced*

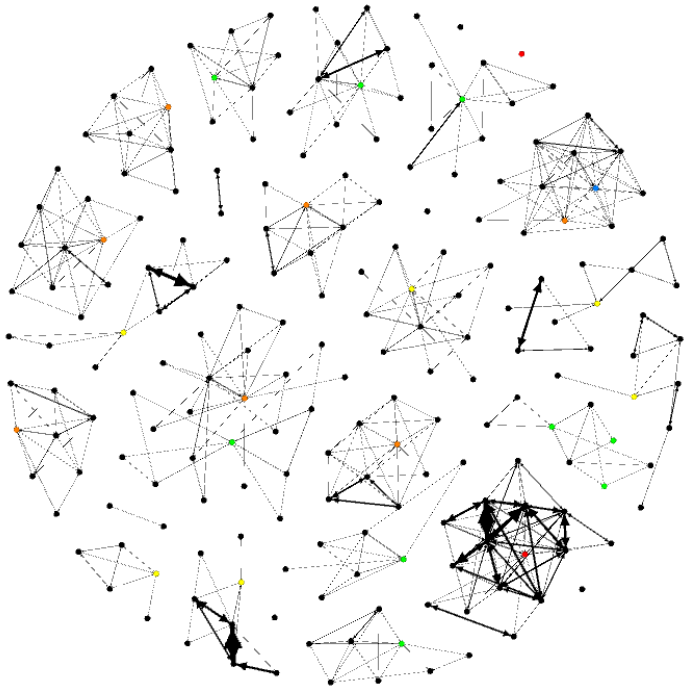


Figure A.22: Graph representing the subset *Large_Normal_Unbalanced*

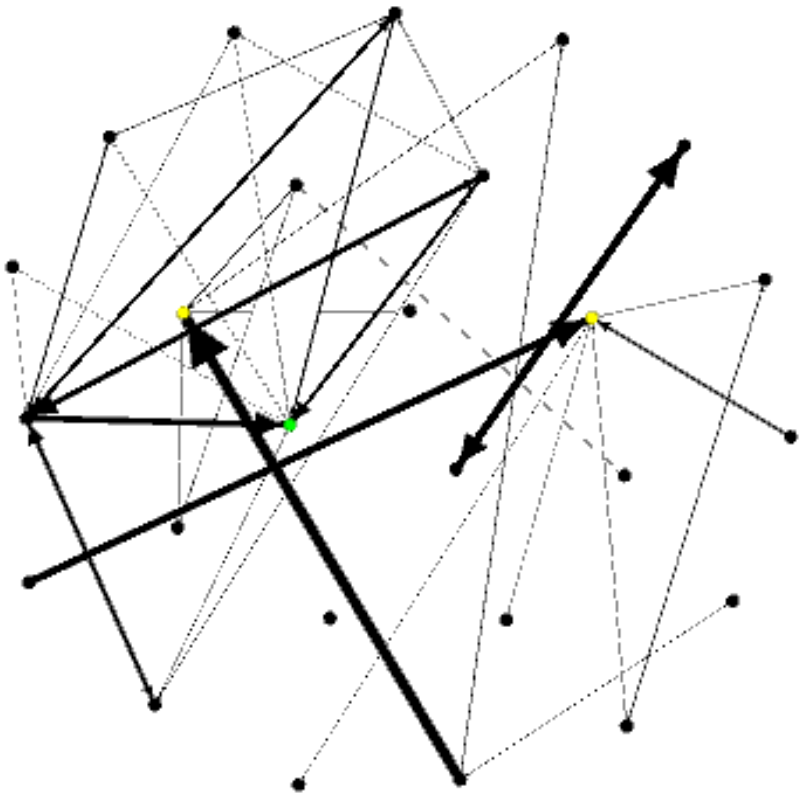


Figure A.23: Graph representing the subset *Large_Active_Balanced*

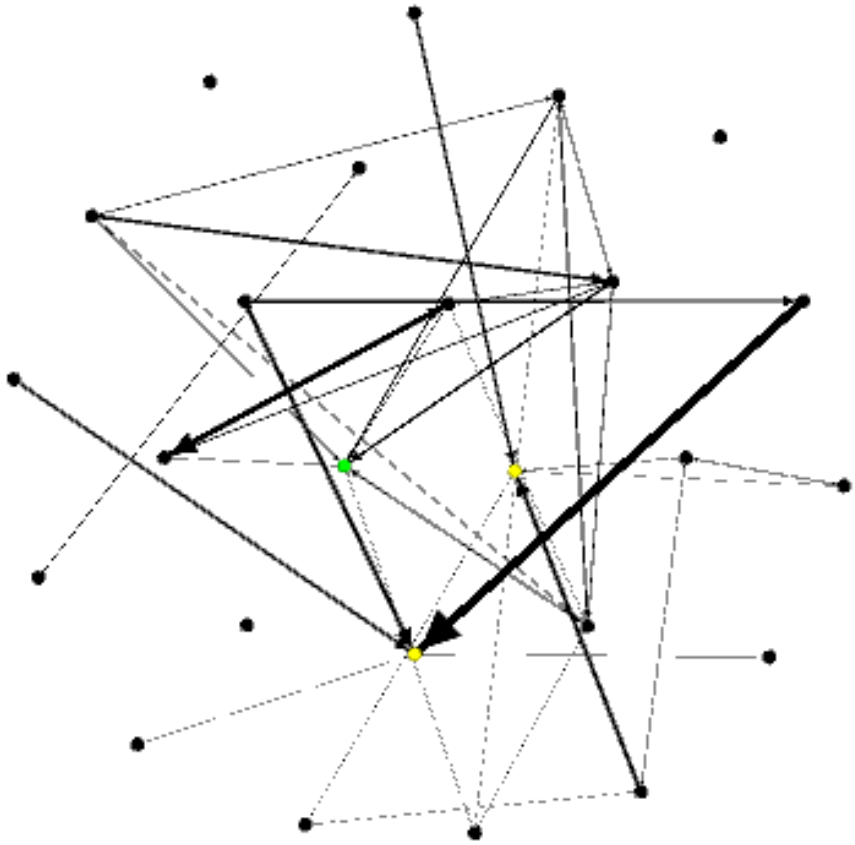


Figure A.24: Graph representing the subset *Large_Active_notsoBalanced*

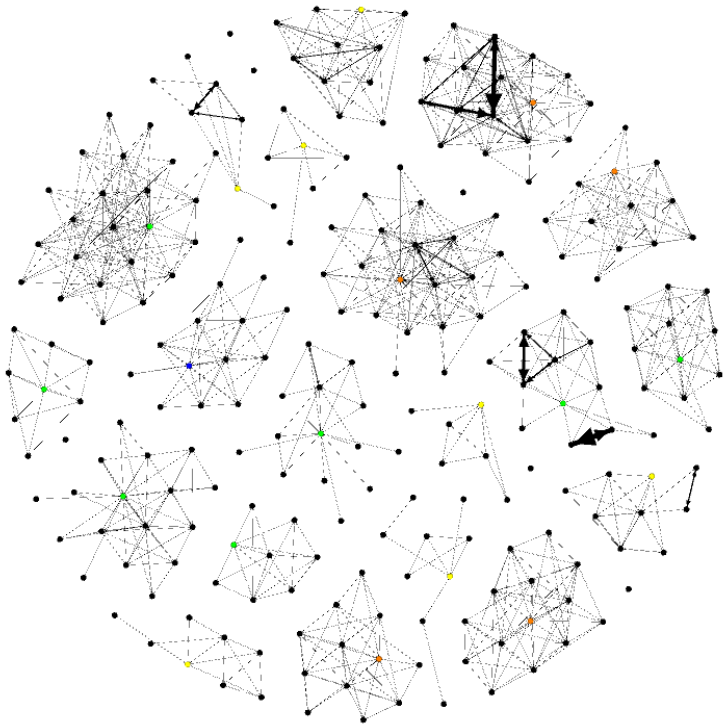


Figure A.25: Graph representing the subset *Large_Active_Unbalanced*

Appendix B
Cliques Approach's Graphs

In this section of the appendix, the graph we obtained in the cliques approach will be provided.
The red node indicates outliers classified in the *normal* set, green node outliers classified in the *less_requested* set, purple node the ones in *little_spammer*. Where there are present, blue nodes are outliers classified as *huge_spammer* and light blue nodes the ones in *pure_spammer*.

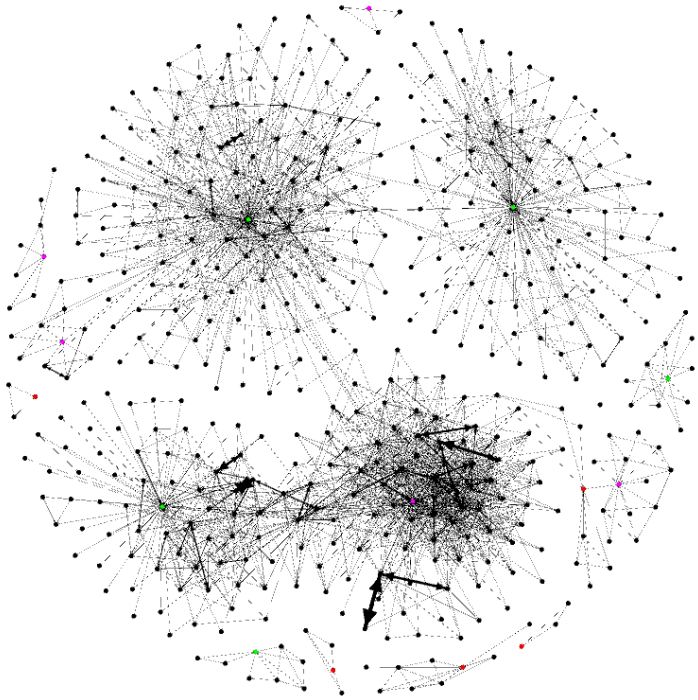


Figure B.1: Graph representing the outliers with respect to betweenness centrality.

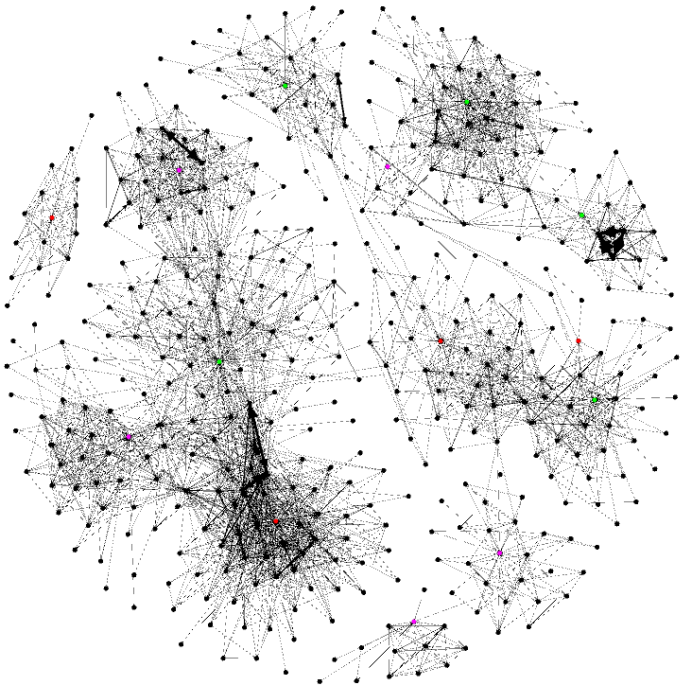


Figure B.2: Graph representing the outliers with respect to weighted betweenness centrality.

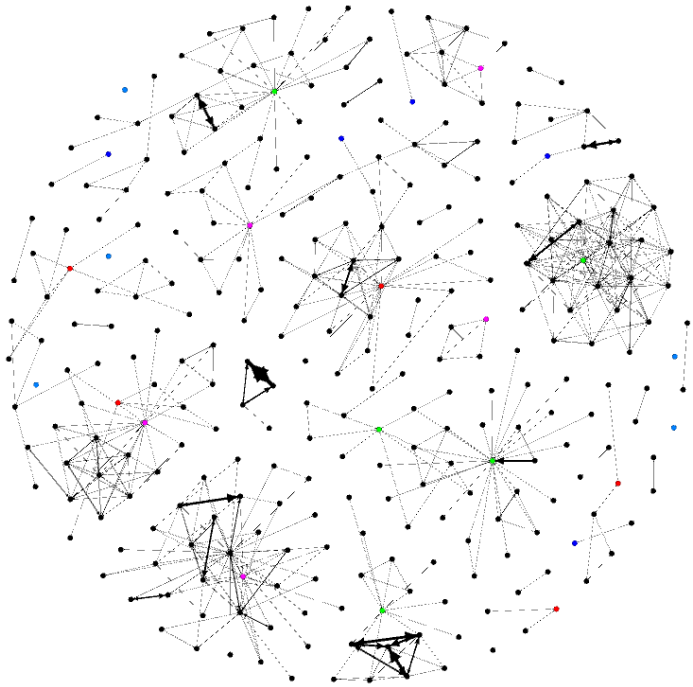


Figure B.3: Graph representing the outliers with respect to closeness centrality.

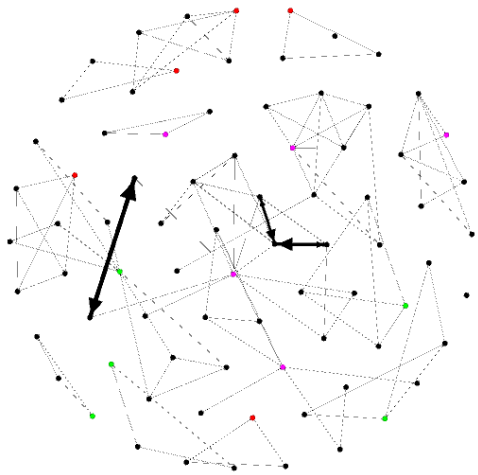


Figure B.4: Graph representing the outliers with respect to weighted closeness centrality.

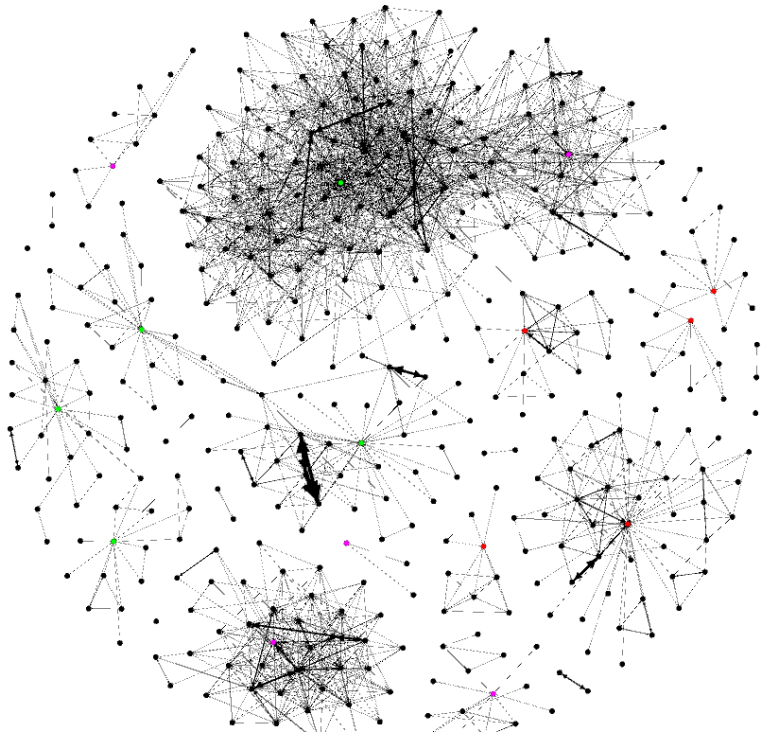


Figure B.5: Graph representing the outliers with respect to closeness and weighted closeness centrality.

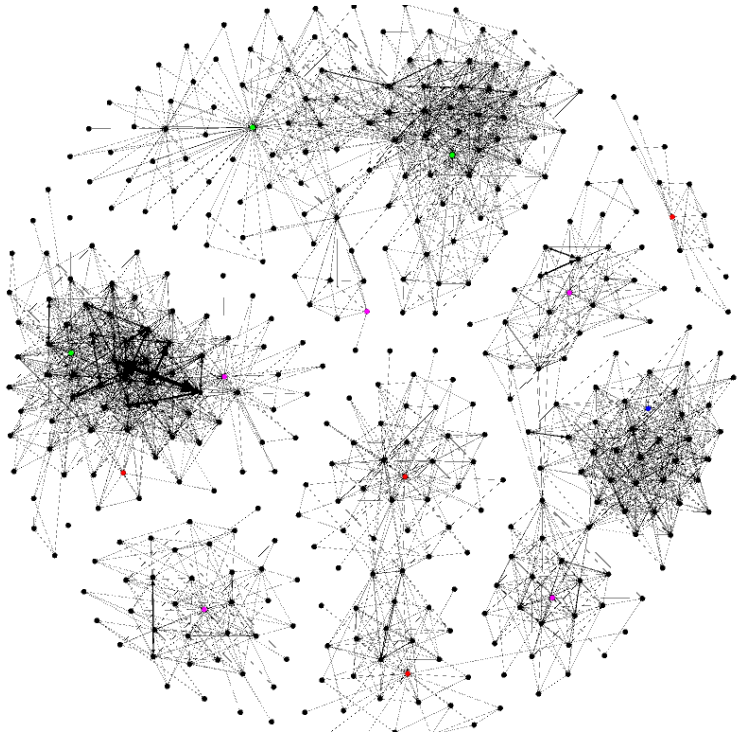


Figure B.6: Graph representing the outliers with respect to weighted betweenness and weighted closeness centrality.

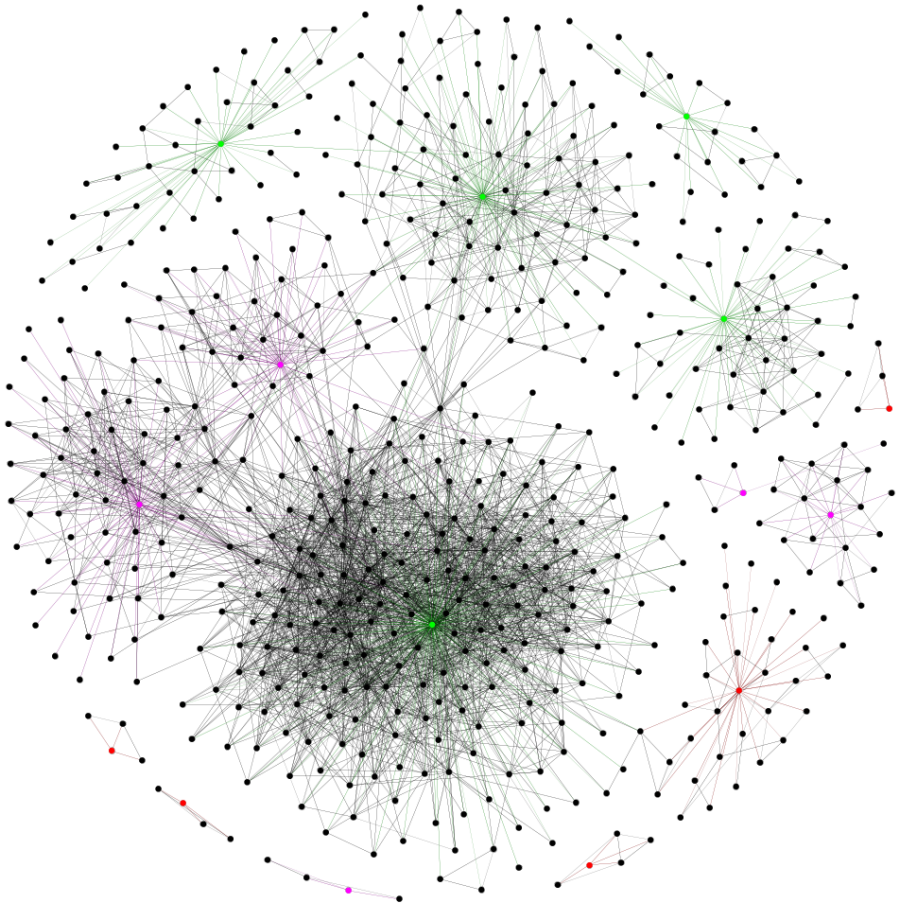


Figure B.7: Graph representing the outliers with respect to betweenness and weighted closeness centrality.

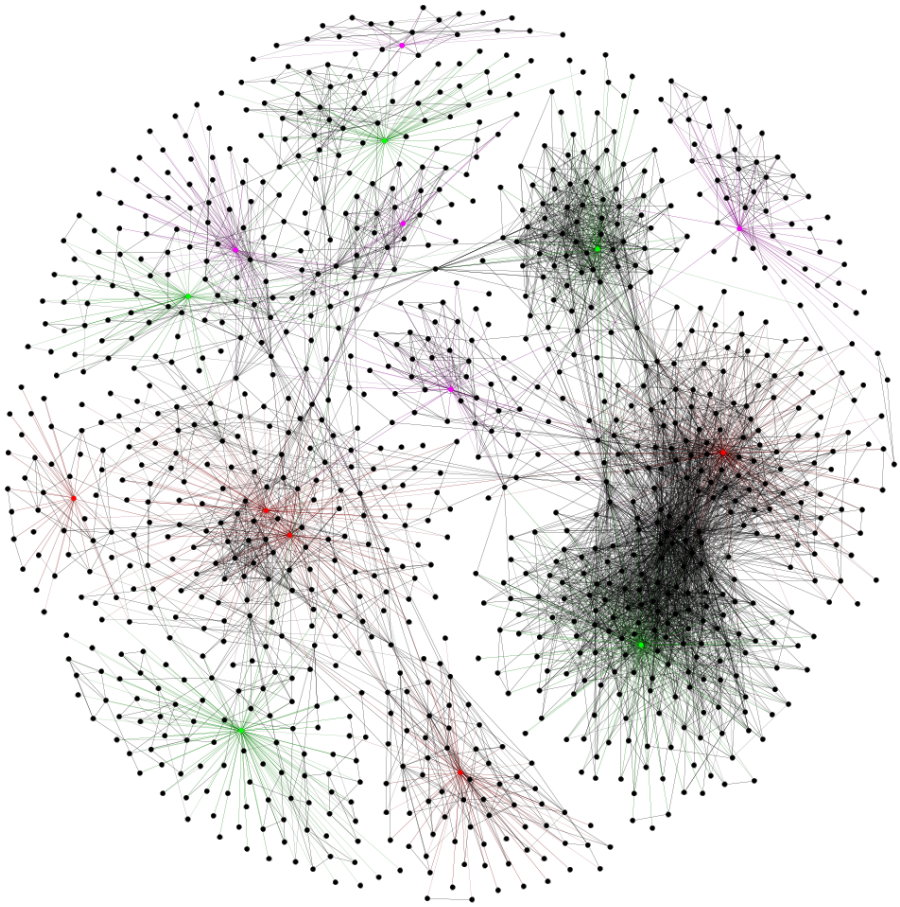


Figure B.8: Graph representing the outliers with respect to betweenness and closeness centrality.

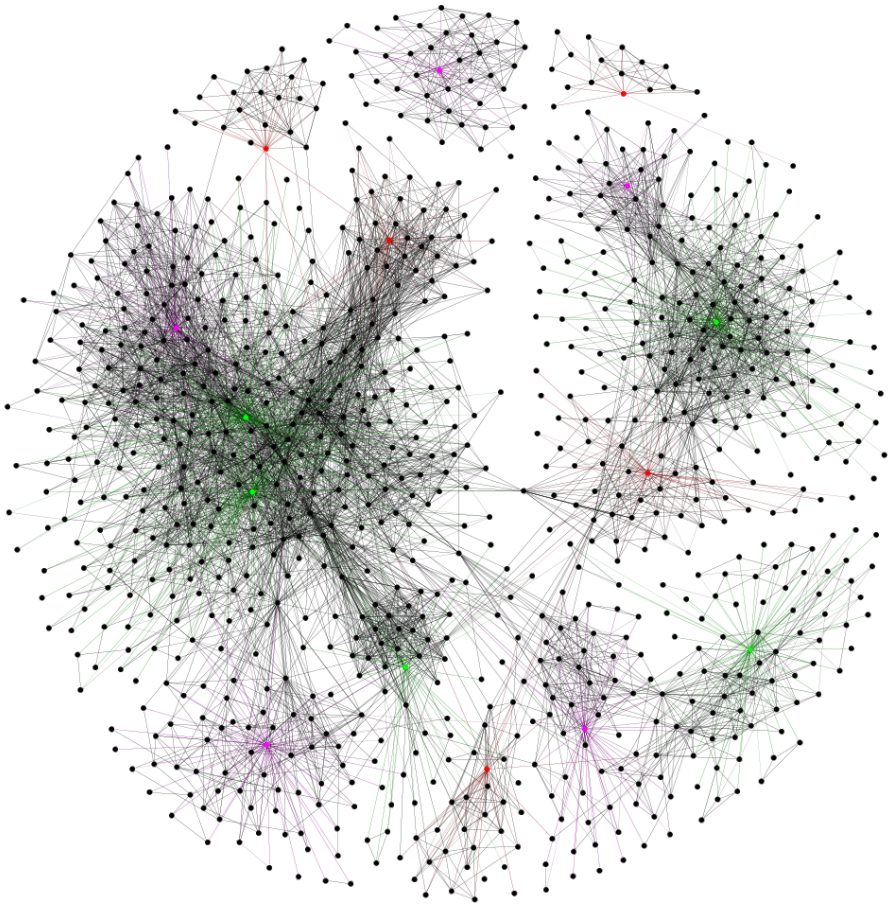


Figure B.9: Graph representing the outliers with respect to betweenness and weighted betweenness centrality.

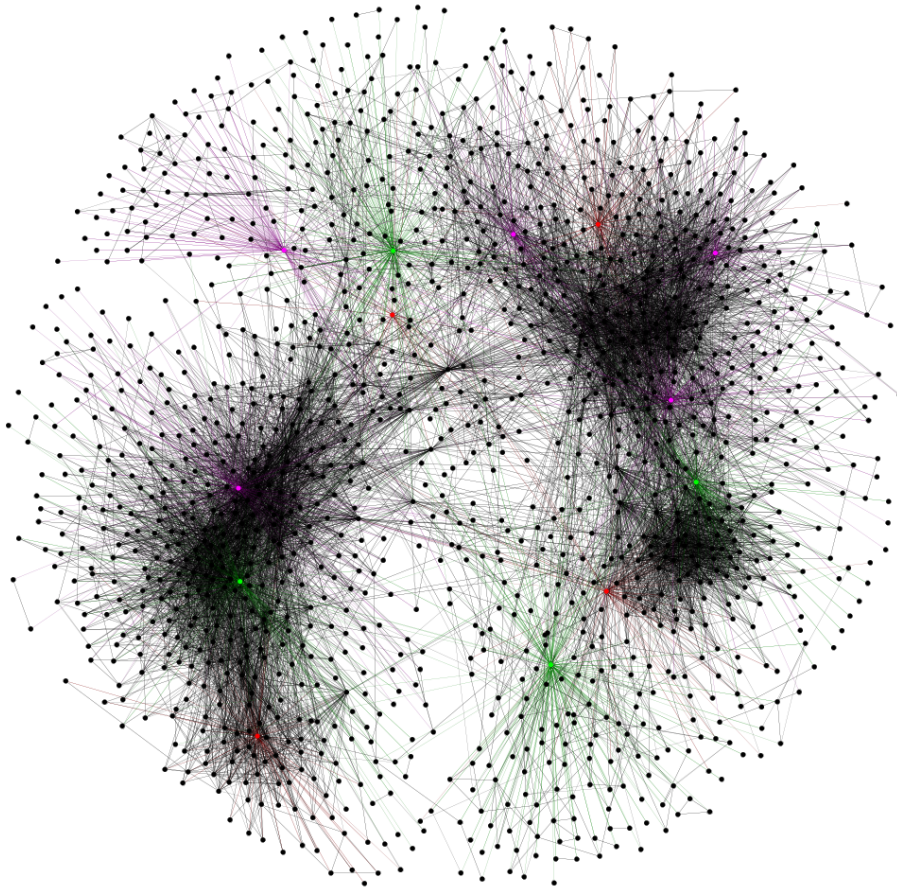


Figure B.10: Graph representing the outliers with respect to betweenness, weighted betweenness and closeness centrality.

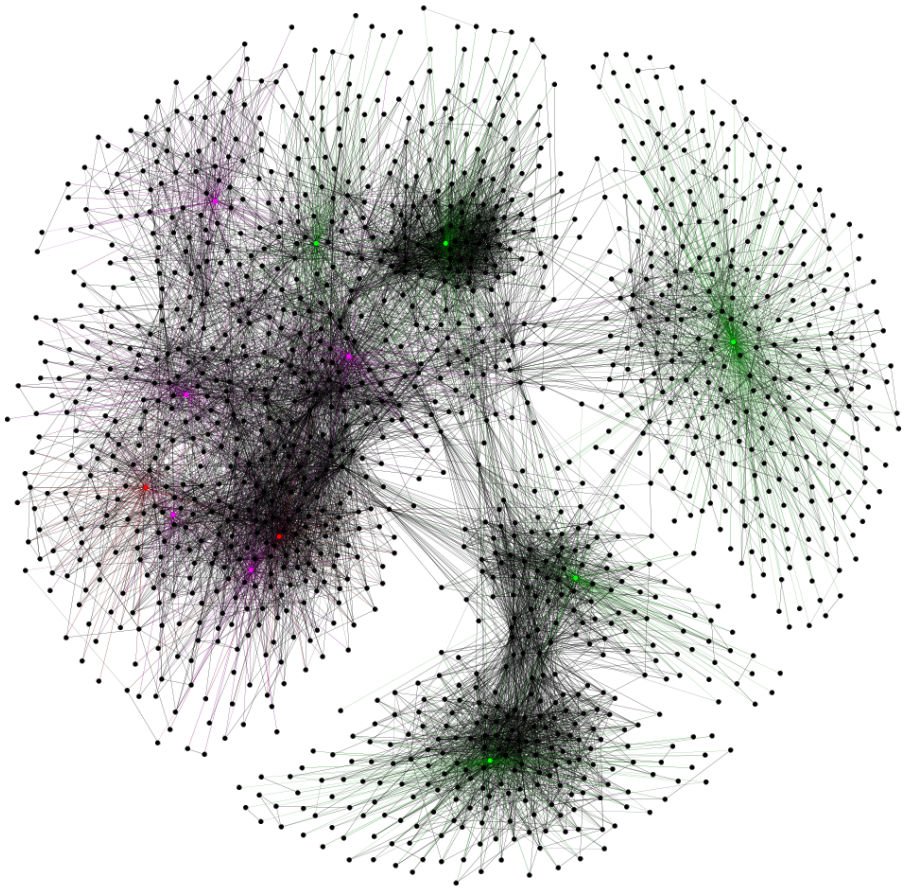


Figure B.11: Graph representing the outliers with respect to betwenness, weighted betweenness and weighted closeness centrality.

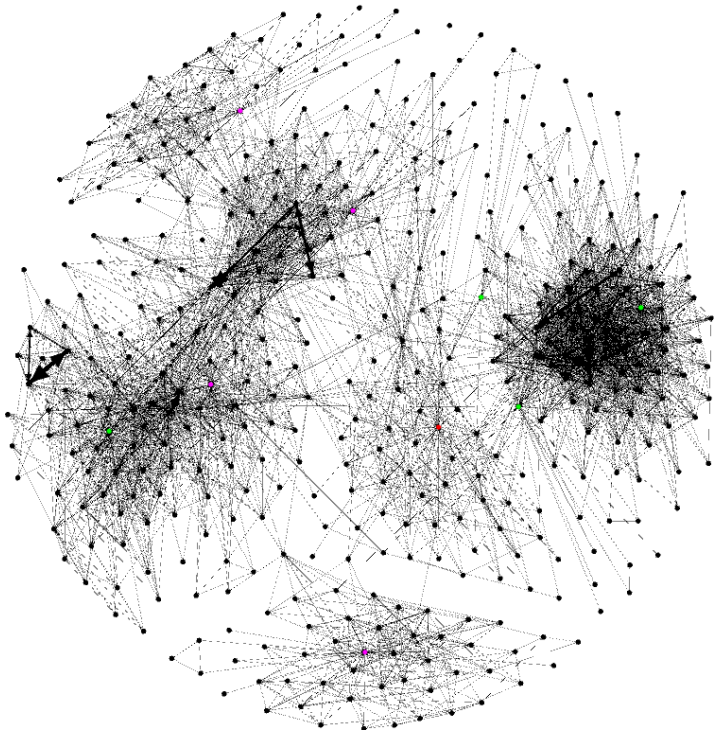


Figure B.12: Graph representing the outliers with respect to weighted betweenness, closeness and weighted closeness centrality.

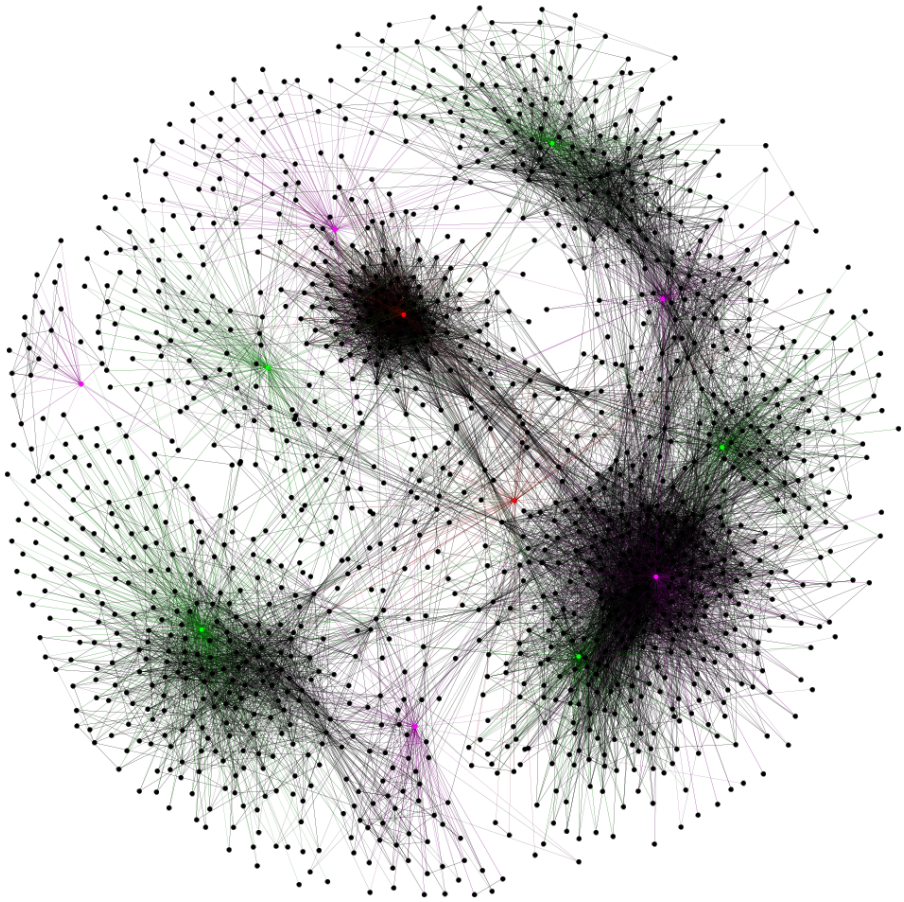


Figure B.13: Graph representing the outliers with respect to betweenness, closeness and weighted closeness centrality.

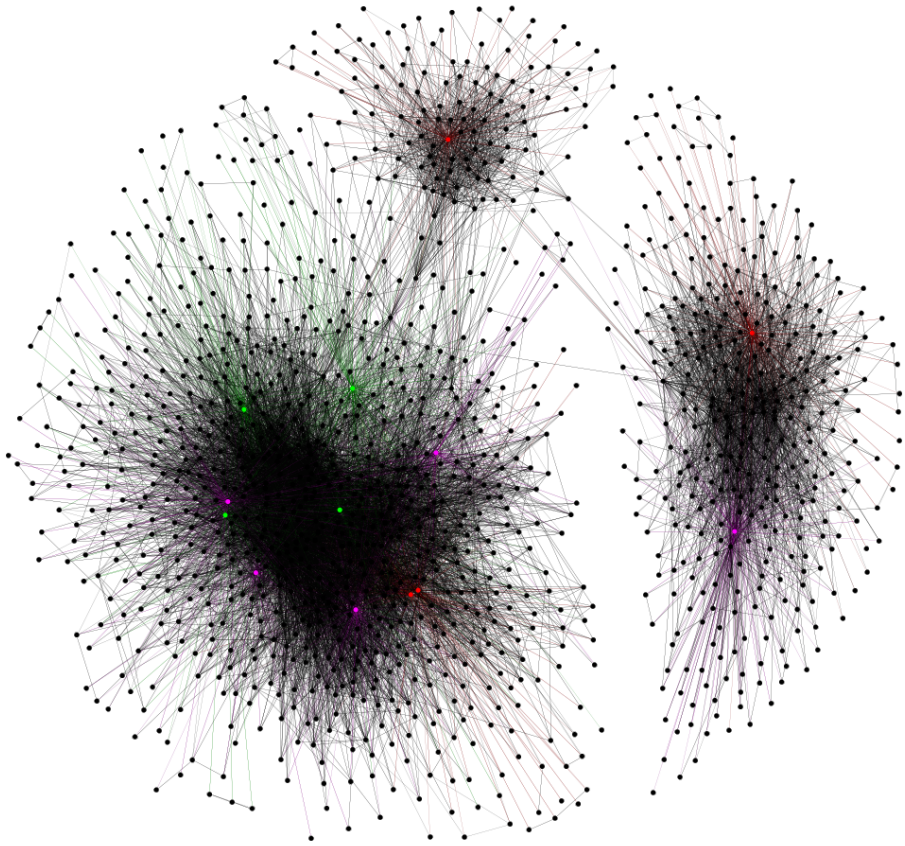
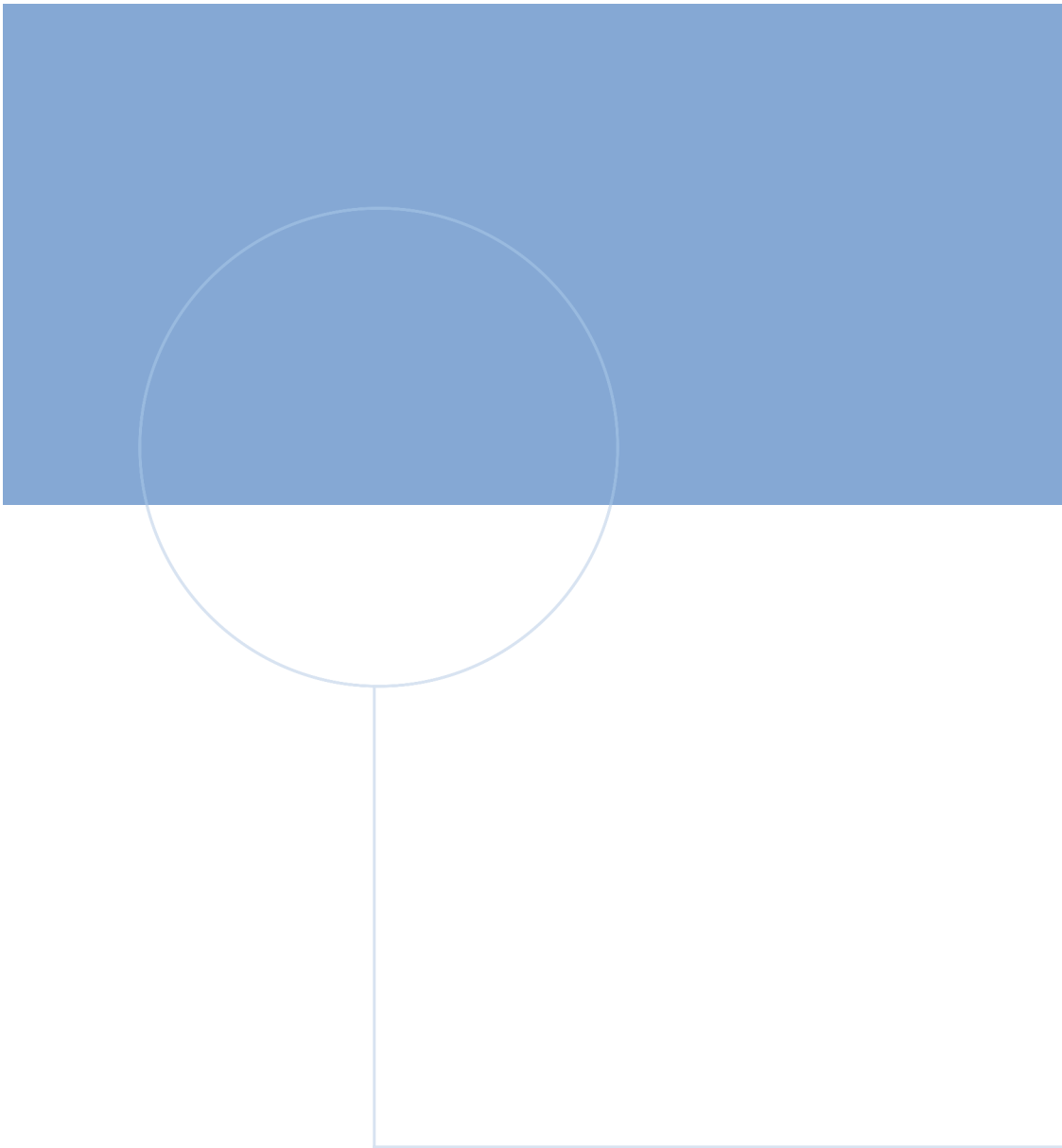


Figure B.14: Graph representing the outliers with respect to betweenness, weighted betweenness, closeness and weighted closeness centrality.



NTNU

Kunnskap for en bedre verden