Master's thesis

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

Anna Fridtun Aarekol

A graph theoretical approach to online predator detection

Master's thesis in Communication Technology
Supervisor: Patrick Bours
Co-supervisor: Natasa Gajic
June 2022

NTNU
Norwegian University of Science and Technology

Anna Fridtun Aarekol

A graph theoretical approach to online predator detection

Master's thesis in Communication Technology
Supervisor: Patrick Bours
Co-supervisor: Natasa Gajic
June 2022

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

Title: A graph theoretical approach to online predator detection

Student: Anna Fridtun Aarekol

Problem description:

In our modern society, a big part of the communication takes place online rather than meeting physically. Online communication poses new threats that did not exist to the same degree for interactions in the physical world. Children are especially exposed when they spend time online. Online platforms create easy access to conversations with children for people with both good and bad intentions. Predators may seize the opportunity to get in contact with children to abuse them sexually. The main goal of this thesis is to find out if the use of graph theory and network analysis can detect predators in chat platforms created for children. A children's game platform where the users can chat with each other in private and group chats will be studied in the research. The platform will be represented as a graph, where the users of the system will be represented as nodes and the conversations between them as edges.

The thesis studies which graph features may contribute to detecting predators. Features that are common for either predators or normal users will be used when classifying the test data. Both supervised and unsupervised machine learning methods will be tested for classification. Supervised learning methods will use a training data set aiming to learn how a typical predator behaves, which then can be used to find similar users. Unsupervised learning algorithms, or clustering algorithms, may detect different kinds of predators, and should hence also be tested in the thesis. The data set provided for the thesis does not have ground data for all users, so the research will also study how to best measure the performance of the methods used.

Date approved: 2022-02-14

Responsible professor: Patrick Bours, NTNU, IIK

Supervisor(s): Natasa Gajic, NTNU, IIK

Abstract

Many children spend much time online, watching videos, playing games, or talking with friends or strangers on social media. Many different online platforms are created targeted at children. The Internet has enabled kids to meet new friends and stay in touch with each other without physically meeting. Although these platforms may contribute significantly to children's social life, they may also pose threats to the children. The online platforms give easy access to conversations with children, even for people with bad intentions. On these platforms, predators can come in contact with children with a low risk of getting disclosed. This master thesis aims to find a method for recognizing predators online using a graph-theoretical approach.

There are research projects that have already studied online predator detection. Most of the research in this area uses textual analysis for the task, many with promising results. The methods involve recognizing specific words or phrases that a predator would use that are unusual for children. There are multiple challenges with this approach. First, when making a predator detection system that analyses text, it can only function if used in the language it was developed for. It will be impossible to create such a system independent of the language. Secondly, the text messages on a chat platform are often informal and contain many slang words. This makes it challenging for machines to interpret what the messages mean.

To avoid the challenges posed by the textual analysis, we use a graph-theoretical approach to detect predators online. Using a real-world data set collected from a social network for children, graph representations of the network will be used to detect predators. The users will be represented as nodes, and the messages between the users as edges. The main goal of the thesis is to study if it is possible to recognize a predator by studying the properties of the nodes in the graph.

We have, throughout the study, designed and implemented a set of features that has been used in various clustering algorithms. From the results of the clustering algorithms, we have discovered multiple users that we considered likely to be predators. To assess some specific users in more detail, we studied anonymized text messages from relevant users and concluded whether the users were predators or not.

We concluded that a graph theoretical approach can be used for online predator detection. However, in the future, both unsupervised and supervised learning in static and dynamic graphs should be studied further for predator detection to find more precise methods to find users with abnormal behavior.

Sammendrag

Mange barn bruker mye tid på nettet, enten de ser på videoer, spiller spill eller snakker med venner eller fremmede over sosiale medier. Mange ulike plattformer er skapt med barn som målgruppe. Dette har muliggjort at barn kan treffe nye venner og holde kontakt med de uten fysisk å møtes. Selv om disse plattformene kan bidra til å øke barns sosiale krets, kan de òg ha mørkere sider ved seg. Plattformer på nett gir tilgang til samtaler med barn for folk med dårlige hensikter. På disse plattformene kan overgripere komme i kontakt med barn med lav risiko for å bli avslørt. Denne masteravhandlingen forsøker å finne en metode for å gjenkjenne overgripere på nett ved hjelp av grafteori.

Det eksisterer prosjekter som studerer deteksjon av overgripere på nett fra før av. Størsteparten av disse benytter tekstlig analyse, og flere kan vise til gode resultater. Analysene går ut på å finne ord eller fraseringer som er mer typisk for en overgriper å bruke enn et barn. Det er flere utfordringer med å analysere tekst i overgrepsdeteksjon. Når man lager et system for analyse av tekstmeldinger vil analysen kun fungere på det språket som systemet ble laget i. Det vil ikke være mulig å lage et deteksjonsprogram som kan fungere uavhengig av språket. En annen utfordring er mengden uformelt språk i tekstmeldinger. Tekstmeldinger inneholder gjerne mer slang, skrivefeil og generelt uformelt språk, som er utfordrende for maskiner å tolke.

For å unngå utfordringene som er beskrevet i prosjekter med tekstanalyse for overgrepsdeteksjon, vil vi bruke grafteori for deteksjon av overgripere på nett. Et datasett fra en chatteplattform med barn som målgruppe, vil presenteres som en graf. Brukerne blir representert av noder, og meldingene som sendes mellom brukerene representeres av kanter. Hovedmålet med arbeidet er å finne ut om det mulig å gjenkjenne en overgriper ved å studere egenskapene til de ulike nodene i grafen.

Vi har gjennom studiet designet og implementert et sett av funksjoner som har blitt brukt i flere forskjellige klyngealgoritmer. Fra resultatene av klyngealgoritmene har vi funnet flere brukere som vi anser som sannsynlige overgripere. Vi har til slutt studert anonymiserte tekstmeldinger som er sendt fra relevante brukere for å konkludere om de er overgriper eller ikke.

Gjennom denne studien, har vi funnet ut at grafteori kan brukes som metode for overgrepsdeteksjon på nett. Videre bør både ikke-overvåket

og overvåket maskinlæring i statiske og dynamiske grafer bli studert videre for å finne en mer presis metode for å finne brukere med unormal oppførsel.

Preface

This thesis is written as the completion of the 5-year MSc in Communication Technology with a specialization in information security at Norwegian University of Science and Technology (NTNU). The supervisor for the thesis has been Professor Patrick Bours at the Department of Information Security and Communication Technology at NTNU. Ph.D. Candidate Natasa Gajic at the Department of Information Security and Communication Technology at NTNU has been co-supervising the thesis work. A preliminary study for this master thesis was completed in the fall of 2021. This thesis was carried out from January to June 2022.

Acknowledgments

I want to thank my supervisor Patrick Bours who provided advice and support throughout the thesis. His knowledge and insights have been central to the outcome of this thesis. I would also like to thank Natasa Gajic, who co-supervised me and contributed with many helpful thoughts and perspectives. I also want to thank the incredible girls at my office for good discussions and laughs. Lastly, thanks to friends and family for all the support.

Contents

List of Figures

List of Tables

List of Acronyms

BIRCH Balanced Iterative Reducing and Clustering using Hierarchies.

BoW Bag of Words.

CC Clustering Coefficient.

CF Cluster Feature.

CNN Convolution Neural Networks.

DBSCAN Density Based Spatial Clustering of Applications with Noise.

GMM Gaussian Mixture Model.

ML Machine Learning.

NN Neural Networks.

NTNU Norwegian University of Science and Technology.

PCI Predatory Conversations Identification.

PII Personal Identifiable Information.

SCI Suspicious Conversations Identification.

SVM Support Vector Machine.

TF-IDF Term Frequency-Inverse Document Frequency.

VFP Victim from Predator disclosure.

VPD Victim from Predator Distinction.

Chapter 1
Introduction

The purpose of this master thesis is to determine if it is possible to use a graph-theoretical approach to detect predators in an online chat network created for children. The thesis aims to detect anomalies in the chat network, which later need to be revisited by a human to determine further if a user is a predator or not. In this chapter, the motivation of the study will be presented. Before the research questions are provided, an explanation of deviations from the problem description will be given. Lastly, the outline for the thesis will be represented, followed by a short disclaimer.

1.1 Motivation

The first known written text was from over 5000 years ago, and since then, the way humans communicate has developed and is still in continuous development. Less than 50 years ago, the Internet was invented. The Internet has provided accessible and fast communication between anyone anywhere. Great benefits have been made possible, such as easy sharing of information, meeting new people, staying in touch with people, and more. Endless social media platforms have been created to enable friends and strangers to communicate with each other by text or multimedia. Facebook and YouTube are examples of widely known and used social networks. In addition, small networks such as dating websites and online games can also be considered social media.

A large number of social media platforms are targeting children. For instance, game platforms that provide chats between the players are popular amongst children. They can meet peers from far away, talk with each other and play games together. These games allow children to find friends, and many kids meet most of their friends online. These social media platforms are thus essential for many children's social life.

Despite the clear advantages of social media created for children, they also pose threats to the users. The platforms give easy access to conversations with children for anyone, also people with bad intentions. Predators can create fake profiles,

masquerading as a child, and quickly start conversations with children with the goal of abusing them sexually. Children who use these kinds of platforms are exposed and vulnerable to predators. According to a report from the National Society for the Prevention of Cruelty to Children from 2014, 12% of children aged 11 to 16 in the UK have received unwanted messages online [CFA14]. An online assault on a child may negatively affect its life, both psychologically, physically, emotionally, behaviorally, and psycho-socially [ZLA+18]. Preventing sexual abuse online and detecting predators is hence crucial.

There already exists some research on predator detection. Most of the studies are based on analysis of the written text between two users, aiming to define language or behavior typical for predators [Mor13; BK19; ZLA+18]. The research projects show that it is possible to achieve good performance with lexical analysis for predator detection. However, there are some disadvantages to analyzing the text between social media users. Firstly, the textual detection systems are built up with a given language. Translating a detection system to a new language is demanding, as there is no one-to-one relation between words of different languages. Hence, the detection system will not work globally. Another challenge is the informal language in online chats. The wide use of incorrect spelling, slang, and informal language in the chats, makes it challenging for computers to interpret what is communicated between two users. With a graph theoretical approach, we will only analyze the chat behavior between users on the platform. By recognizing patterns of how the users communicate with each other, this thesis aims to detect users displaying behavior that deviates from normal behavior. Predators will most likely display this abnormal behavior, but the behavior may also, for instance, display spammers or other types of users. No other studies are known to use the same methods for predator detection, so this study will provide new insights into the field of online predator detection.

The data set that we will use for this master thesis is from a children's game called MovieStarPlanet[1]. The game is targeted at children aged 8 to 15 years and allows the users to chat and play games with each other. The data set is real-life chat data collected over five months from the game. All data used for the study is pre-processed and anonymized before being analyzed. The data used for the analysis is users represented by a randomized ID number and the number of messages between the different users. A graph representation of the chat network will be created using this information. Nodes represent the users, and the chats with other users are represented with edges labeled with the number of messages sent. No actual text messages were used to do the graph analysis, solely the chat patterns of the users. Text messages were used later in the process to analyze whether there was predatory behavior from relevant nodes. The study's main hypothesis is that a predator will behave differently than a child on social media. For example, a predator might want

[1] https://moviestarplanet.com/

to contact many different users aiming to get a satisfying response from some of the children, resulting in a lot of short conversations. On the other hand, a child might contact fewer other children but have more extended conversations with them. These differences may be visible in graph patterns of the network.

1.2 Deviations from the problem description

The problem description stated that both supervised and unsupervised learning algorithms would be used for predator detection. We initially planned to use a set of known predators as training data for the supervised learning algorithms. However, we did not get access to the set of known predators, so the supervised learning algorithms were never tested. The problem description also states that time will be spent on finding a way of measuring the performance of the algorithms. It did not make sense to measure the performance of the unsupervised learning algorithms. The unsupervised algorithms do not give all users a class the same way as supervised learning. Hence, this part of the thesis was also dropped.

1.3 Research questions

With the MovieStarPlanet data set, we will create a graph consisting of nodes representing users and weighted, directed edges representing the number of messages sent between users. The main data set used consists of private chats on the platform over five months. All data is unclassified and may contain both children and predators. After pre-processing and structuring the data, different graph features will be extracted, and all users will be represented with a feature set. Unsupervised clustering algorithms will be used to find abnormal user behavior in the network. By analyzing text messages, users with abnormal behavior will be classified as predators or not. The thesis consists of one main research question with three sub-research questions. Research questions were created and carried out in the pre-project proceeding of this master thesis [Aar21]. The research questions were later revisited and modified after more insights about the project's limits were known.

> ***RQ:*** *Can we detect predators in online chats for children by using a graph-theoretical approach?*

The main goal of this thesis is to investigate if the graph-theoretical approach can be used to detect predators online. The study focuses on finding users that behave differently than regular users, and these users should be investigated further by a human evaluator. The study uses real-life data with graph-theoretical concepts and machine learning algorithms for predator detection. To answer this research question, the following three sub-research questions are formulated.

RQ a: *Which graph features can be used to detect predators in an online chat network for children?*

The nodes of a graph can be represented with a set of features. There are endless different features that can be used to describe a node in a graph, or this case, a user on a chat platform. There is thus a need to investigate which features describe the user in a manner that separates children from predators. The goal is to find a set of features that works well at detecting predators. The feature set will be based on related research and be developed through a manual investigation of visualizations of the chat network.

RQ b: *Can unsupervised clustering algorithms be used to detect predators in an online chat network for children?*

Predators may behave very differently from each other. Some may send out many messages to find a suitable victim, while others may carefully choose victims in other ways. Using clustering, we will obtain groups of users with similar behavior. Both predator users and regular users may have contrasting behavior and can be clustered into several different clusters. Hence, the clustering results may disclose different types of predatory and normal behavior. Small clusters point to abnormal behavior and might be related to predatory or other unwanted behavior in the game. Hence, the thesis should investigate if unsupervised clustering algorithms can be used to detect predators.

RQ c: *How does the length of the time frame for the data set influence predator detection in an online chat network for children?*

The data set used in the thesis consists of chat data collected over relatively small, continuous periods. The smaller data sets are also concatenated to make a more extensive, non-continuous set. The thesis should investigate how data analysis with different lengths of time frames influences how well the clustering algorithms find anomalies.

1.4 Outline

Chapter 2 resents some background theory on predator detection and graph theoretical concepts, and chapter 3 describes a selection of state-of-the-art research on predator detection and related research with graph theoretical approaches. The methodology used in the thesis will be presented in Chapter 4, and Chapter 5 describes the results.

Lastly a general discussion is provided in Chapter 6, followed by conclusion and future work in Chapter 7.

1.5 Disclaimer

This thesis aims to examine new methods to detect predators online. As no classification data was made available to us, we can only look for predatory behavior. Whenever in this thesis we claim that a user is a predator, we do not refer to a convicted person, but we mean a person that shows predatory behavior based on the user's chat patterns and human interpretation of the user's text messages. The classifications of predators in the thesis were confirmed by experts who have read many predatory conversations in the past.

Chapter 2
Background

This chapter will first present the background for the master thesis. Firstly, some background theories about sexual predators and their psychology and behavior will be provided. Then, a brief introduction to basic graph theoretical concepts will be given together with a description of some central graph features. Lastly, an introduction to Machine Learning (ML) is presented, and some relevant ML algorithms will be described briefly.

2.1 Sexual predators

Morris [Mor13] defines *sexual predatory* by identifying two characteristics. The first characteristic is named "age disparity" and revolves around the age of the people involved. A predator is an adult that has conversations with one or more underaged children or teenagers. The second characteristic is named "inappropriate intimacy". The interaction between the adult predator and the underaged child must contain conversation on intimate topics introduced or encouraged by the predator. To summarize, a sexual predator is an adult that encourages intimate conversations with a child. The term *predator* is often related to the term *pedophile*; however, it has a slightly different meaning. A pedophile is physically attracted to children and may want to start long-term relationships with children. Predators do not necessarily have the same physical attraction toward children, but they may abuse a victim when they see an opportunity that they can utilize. Predator is a broader term, and it covers more types of abuse cases and will hence be used as a collective name for adults that sexually abuses children in this paper.

Many predators will go through a specific process when abusing a child online, referred to as *online grooming*. Craven et al. [CBG06] define sexual grooming of children as: "*A process by which a person prepares a child, significant adults, and the environment for the abuse of this child. Specific goals include gaining access to the child, gaining the child's compliance, and maintaining the child's secrecy to avoid*

disclosure. This process serves to strengthen the offender's abusive pattern, as it may be used as a means of justifying or denying their actions." The goal of the grooming process is to be able to abuse the victim without being caught. Without the victim's trust, there is a more significant risk of being disclosed, and predators will hence make an effort to gain trust. Online grooming refers to sexual grooming of a child over the Internet. Craven et al. [CBG06] also define different types of grooming. The first type of grooming is self-grooming. Self-grooming is the process where the predator will try to justify their actions for themselves. The second grooming type is the grooming of the environment. This type refers to ensuring that the victim's environment won't disclose the predator. The environment includes parents and other people close to the victim child. The last grooming type is the grooming of the child, which is the most recognized form of grooming. This type involves gaining the trust of children to later be able to exploit them without being disclosed. Without the child's trust, the risk of the child ending the conversation or talking to its parents gets bigger.

Several research projects have studied the grooming process and have described different stages that the predator will use in a conversation with a victim child [OCo03; Kon09; GACS16; PGS15]. Some of the papers describe three stages, while others list up to six stages. While there are differences in details in the different articles, they all describe a process of firstly creating friendship and trust with the victim before gathering information about the child's environment and assessing the risk. Lastly, they introduce sexual topics in the conversation. Table 2.1 summarizes the grooming process in six stages.

Several research projects have aimed to define typical characteristics common to predators. Babchishin et al. [BKH11] studied demographical variables of sexual of online offenders. The study shows that most predators are male, and most of them are in their late 30s. Around 30% of the offenders in the study were not married, and approximately 15% were unemployed.

Malesky [Mal07] studied the modus operandi of convicted sex offenders. The goal was to be able to identify potential victims of online abuse. Malesky gives insights into the typical characteristics of victims of online abuse and insights into how predators typically choose victims. The study suggests that three main themes characterize a victim. The first theme is when a child mentions sex. The child may use their profile biography or profile name to say something related to sex, which some predators will find as motivation for the abuse. The second theme described in the study is when the child is behaving submissive or needy. The third theme is when the child has a young-sounding name. The child may have a profile name that sounds young to the predator, for instance, "sophie11", which could motivate the predator to reach out to that specific child. The study claims that child's willingness

Table 2.1: Stages of online grooming

Stages	Description
1: Friendship forming stage	The predator will try getting to know the child, and both the child and predator may exchange personal information such as location, age, and family situation.
2: Relationship forming stage	The predator will ask questions to the child about their home- and school situation. By giving compliments and conversations about topics that interest the child, the predator tries to gain trust.
3: Risk assessment stage	The predator will try to gather information to determine the likelihood of getting detected by the child's surroundings, such as parents or siblings.
4: Exclusivity forming stage	The goal of this stage is often to make the child believe that the predator is to be trusted and to establish a feeling of love and exclusiveness in the relationship.
5: Sexual stage	The predator will ask questions and introduce topics related to sex, body, and intimacy. The predator may ask the child to participate in sexual activities online, such as sending sexual pictures.
6: Conclusion stage	The predator may try to organize further contact and physical meetings

to talk about sex was the most common characteristic of the victims. In addition, the paper states that some predators will send messages to many potential victims and choose the victim based on the response the predator receives. O'Connell [OCo03] describes a similar behavior, where predators will typically send out a short message to many children and wait for the children to respond before deciding which children to start long conversations with. The decision will then be based on the answer from the children.

Olson et al. [ODER07] studied how a predator chooses a victim and which properties victims typically possess. The study shows that children with low self-esteem or a lack of confidence, naive and without knowledge about abuse, and children with dysfunctional family dynamics are typical traits of abuse victims.

When predators wish to exploit a child, they may either present themselves as who they are, or they can masquerade as a child or as another adult. Malesky's paper on the modus operandi of sex offenders states that 29% of the predators in the study presented themselves as children when communicating with children [Mal07]. On the other hand, Wolak et al. report that only 5% of offenders present themselves as

children [WFMY10]. The number of predators presenting themselves as children will vary depending on the online chat platform. Some platforms will not allow adults to play, which will force predators to masquerade as children.

2.2 Graph theory

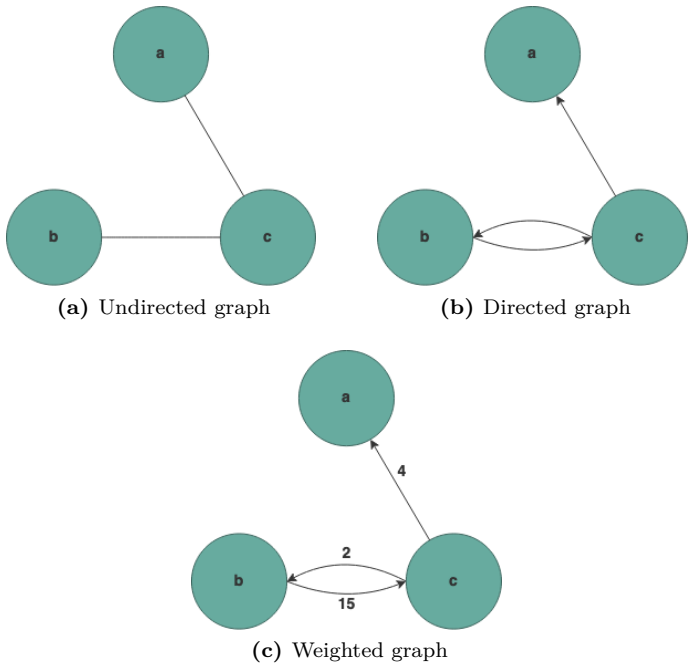
This section will present an overview of graph theory and an introduction to some basic graph theoretical concepts. Graph theoretical concepts were studied, and identification of relevant background material was carried out in the project preceding this thesis [Aar21]. This is amended with a discussion of a few additional concepts that have shown to be relevant after the project.

2.2.1 Introduction to graph theory

According to [Tru13] a graph is defined as "*an object consisting of two sets called its node set and its edge set. The node set is a finite nonempty set. The edge set may be empty, but otherwise its elements are two-element subsets of the node set.*" Graphs are typically presented as $G = (V, E)$ where V is the set of nodes, and E is the set of edges. Each edge is connecting two nodes, called the *endpoints* of the edge. Two nodes joined by an edge are *neighbor nodes*, and two edges that share the same node as an endpoint are *neighbor edges*. A node, v, that is an endpoint of an edge, e, is *incident* on e and e is similarly incident on v.

Graphs are divided into *undirected* graphs and *directed* graphs. In undirected graphs, the edges will have no specific direction, but the edges have a specified direction for directed graphs, presented with arrows. For example, in a chat network, the nodes can represent users of the chat platform, and the edges can represent the conversation between two users. A directed edge would indicate messages sent from one user to another. If the other user is sending messages back, there will be another edge in the other direction between these two nodes. So then, the two nodes would be connected with two edges. The graph can also have edges with *weights* linked to it. The weights give some extra information about the data being represented in the graph. For example, the weight may represent the number of messages sent from one person to another in a chat network. Figure 2.1 shows examples of the different types of graphs. In Figure 2.1a, there is communication between user a and user c and between user b and user c. In Figure 2.1b, user c sends messages to user a, and users b and c send messages to each other. Finally, in Figure 2.1c, user c has sent four messages to user a and two to user b, while user b sends fifteen messages back to user c.

A *path* in a graph is a sequence of distinct edges that connects a series of distinct nodes [BW10]. The *shortest path* between two nodes in the graph is the path with



(a) Undirected graph **(b)** Directed graph

(c) Weighted graph

Figure 2.1: Comparison of different types of graphs

the smallest set of edges in unweighted graphs or the path with the lowest sum of weights in weighted graphs. In directed graphs, a path needs to follow the direction of the path. For example, in the graph in Figure 2.1a, there is a path from a to b and a path from b to a, but in Figure 2.1b, there is a path from b to a, but no path from a to b.

2.2.2 Selection of graph theory concepts

In graph theory, different concepts describe the characteristics of the data presented. Some of the relevant concepts to this research will be discussed further.

A *subgraph* $G' = (V', E')$ of the graph $G = (V, E)$ is a graph where the nodes, V' are a subset of V and the edges E' are a subset of E. A *clique* is a subgraph of a graph G where all the nodes are connected.

Node degree is the number of edges incident on a specific node. The degree of a node can also be interpreted as the number of neighbors of a node. It may be interesting to also look at the *in-degree* and *out-degree* values in directed graphs. In-degree is the sum of all edges incident on a node where the direction of the edges is leading into the node. Out-degree is the sum of all edges incident on a node

where the direction of the edges is outgoing from the node. In weighted graphs, the weight of the edge can be summarized to calculate the weighted degree, in-degree, or out-degree value.

Different types of *centrality* rank the importance of nodes in the graph by various indicators. Degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality are examples of different ways of calculating centrality.

A *Clustering Coefficient (CC)* is a measure of the influence a node has in a network. It measures if the node is a part of a highly connected group of nodes in the graph, also known as a cluster. The CC of a node v is based on the degree $deg(v)$ of the node and the number of edges connecting the neighbors of v, $N(v)$. The following formula can be used to calculate the CC of a node v:

$$CC(v) = \frac{2N(v)}{deg(v) \cdot (deg(v) - 1)}$$

Figure 2.2 shows the CC calculated on the red node in three different cases.

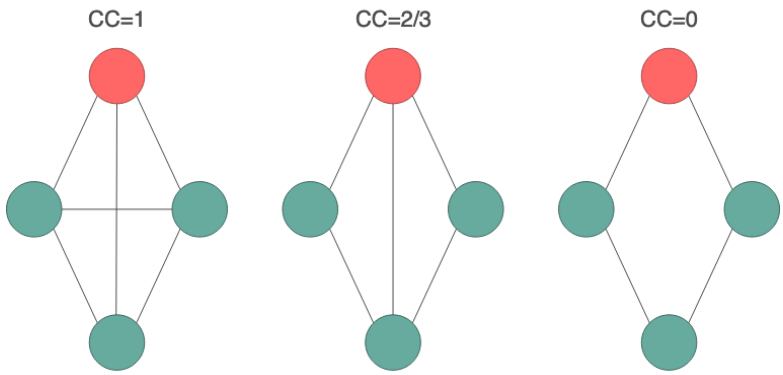


Figure 2.2: Clustering coefficient for the red node in three different graphs

An *outlier* is a node that lies outside of the typical pattern of the nodes in a graph. Outliers will typically be interesting nodes to investigate in anomaly detection.

2.3 Machine learning

Machine Learning (ML) is the science of giving computers the capability to make predictions based on past information without being explicitly programmed [MRT18, p. 1]. ML can be used to do various tasks, from document and text classification to spam filtering, detection of criminals, risk analysis, or clustering of user groups.

In ML, all data set items that shall be classified are represented by the same set of features. The features may be binary, categorical, or continuous, depending on the data set. ML is divided into supervised, reinforcement, and unsupervised learning. The latter ML type will be used to classify the data set in this project and will be explained in more detail in this section, together with some examples of relevant algorithms. Lastly, supervised learning will be explained briefly.

2.3.1 Unsupervised learning

Unsupervised learning aims to find patterns in the data set to classify the different data points [BvLR11]. Unsupervised ML uses only unclassified data without any training data as input for classification. Clustering is a technique of unsupervised learning that is widely studied and used. In general, clustering and unsupervised learning are suitable methods to find unknown properties of the data of the given data set. This subsection describes a selection of clustering algorithms relevant to this thesis.

k-means

k-means clustering is a widely used clustering algorithm that aims to cluster the data set into k number of clusters [HW79]. The algorithm takes the number of clusters, k, as a parameter. The algorithm starts by selecting k random data points as starting points and calculates the distance between each data point in the data set to the k starting points. The data points in the data set will then be clustered with the starting point it is closest to. The mean data point for the new clusters, the centroids, is calculated. New clusters are then calculated by using the computed centroids as starting points. And again, new centroids are calculated, and new clusters are formed based on the centroids. This iterative process will be repeated until the mean values do not change from one iteration to another or until the number of iterations reaches a maximum limit. The whole process will be repeated with new random starting points a given number of times. The algorithm will choose the clustering that gives the least variation.

The advantages of the k-means algorithm are that it is simple, robust, efficient, and can work with various data points. However, the algorithm works poorly for clusters with non-spherical shapes and is sensitive to outliers [Wu12].

Gaussian Mixture Model

Gaussian Mixture Model (GMM) [Rey09] uses a mixture of Gaussian distributions to cluster data, where each of the distributions represents a cluster. The algorithm works similarly to the k-means algorithm, but in GMM, both the mean value and covariance are used to calculate the centroids, where k-means only uses the mean

value. The algorithm has a probabilistic approach, i.e., it calculates the probability that a given data point is in the different classes. The probability calculation is done through the probability density function. The data points are clustered with the class with the highest probability. For a given data set, the algorithm will produce k Gaussian distributions (clusters), where each of them has a mean vector and a covariance matrix. The mean and covariance are calculated through the Expectation-Maximization, a technique for finding the proper parameters for a model.

GMM is a relatively easy algorithm to implement as it requires few parameters. Furthermore, it does not need the data to have a specific geometry, unlike k-means, which assumes the clusters to have a circular form.

Agglomerative clustering

Agglomerative clustering performs hierarchical clustering with a bottom-up approach [SB13]. The algorithm uses a distance measure, for instance, the Euclidean distance, to cluster the data. First, the distance between all data points of the data will be calculated, and the two data points closest will be clustered. Then the same procedure will repeat, but with the new cluster calculated as one instance. Next, the distances are recalculated with the new cluster and then new the smallest distance will form a cluster. This will be repeated until all data points are in one cluster. Figure 2.3 shows how the algorithm works. Node a and b are the closest, thus will be clustered together first. Then, c and d will be clustered, and lastly, the two clusters will be clustered into one large cluster. The algorithm takes the number of clusters as a parameter, and it ends when the chosen number of clusters is achieved.

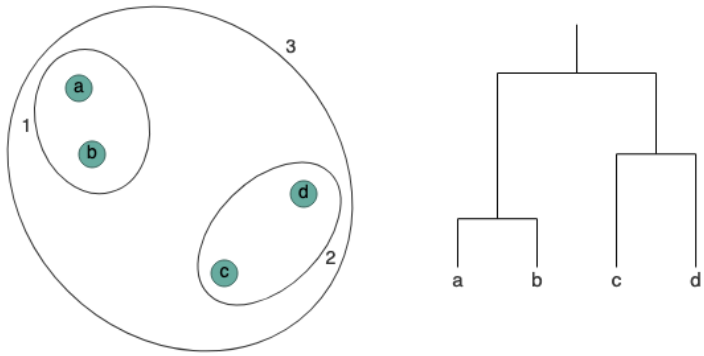


Figure 2.3: Example of the agglomerative clustering algorithm

The agglomerative clustering algorithm is simple to implement and can be a nice way to structure data. However, it is slow and not suitable for larger data sets [WBKP08].

BIRCH

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a hierarchical clustering algorithm created to handle large amounts of data points [ZRL97]. The algorithm aims to create many small radius groups and later cluster these small groups into larger clusters by using other clustering algorithms such as agglomerative clustering. The algorithm takes in three parameters, a threshold T, and a branching factor, B and the number of clusters, k. The BIRCH algorithm will first scan the data set and build a Cluster Feature (CF) tree. CF is a set of three values describing a cluster: $CF = (N, \vec{LS}, SS)$, where N is the number of data points in the cluster, $\vec{LS}$ is the linear sum of the N data points, and SS is the squared sum of the N data points. Using the CF values makes it possible to calculate the distance between two clusters. The nodes in the CF tree are called CF nodes. When a new data point is clustered, it will start at the root CF node and traverse the tree to find the leaf CF node it is closest to. If the radius to the nearest cluster in the CF node is larger than the square of the threshold, T, the data point will form a new cluster at the same CF node. If the number of clusters in the CF node exceeds the branching factor, the node will split and form two new children nodes. After the CF tree is created, the clusters will be clustered further into k clusters with another clustering algorithm.

BIRCH is an algorithm created to handle large data sets efficiently [ZRL97]. However, the algorithm is a bit complex to use as it takes three parameters which can make it challenging to optimize.

Mean shift

In similarity to the k-means algorithm, mean shift is a centroid-based algorithm [Car15]. The algorithm consists of iterations until the centroids converge. During one iteration, all data points will shift towards the mean of the neighborhood surrounding them. The neighborhood is a circle shape with the relevant data point as the center and the parameter *bandwidth* as the radius. For each iteration, the mean is found by calculating the maximum of a density function which includes a kernel function. The kernel function, K, is used to calculate the weight between the center and the data points in the neighborhood. The mean is calculated with the following formula.

$$m(x_i) = \frac{\sum_{x_j \in N(x_i)} K(x_j - x_i) x_j}{\sum_{x_j \in N(x_i)} K(x_j - x_i)}, \tag{2.1}$$

where x_i is the centroid in the i'th iteration, and $N(x_i)$ is the neighborhood of data points within a the radius from the centroid.

An advantage of the algorithm is that the number of clusters does not need to be predefined; it can calculate the number of clusters that best fits the given data set. However, the algorithm does not scale with a high number of dimensions [CM02].

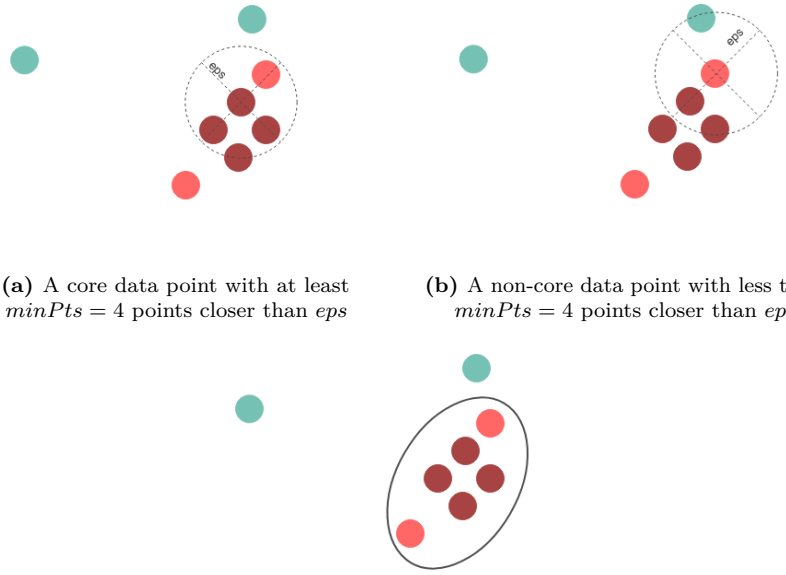
DBSCAN

Density Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering algorithm is a density-based algorithm that separates different clusters from each other where there are continuous regions of objects with low density [EKSX96]. The data points within these regions will be classified as outliers. The DBSCAN algorithm takes two values as input. *minPts* is a threshold for the minimum points that need to be directly connected to a data point for it to be considered as a core point. *eps* is a distance threshold that decides the distance between two points for them to be considered directly connected. The algorithm will find core data points based on the input values, *minPts*, and *eps*. Core points are points that has more than *minPts* of points directly connected to themselves. After all core points are calculated, all connected core points will be assigned to the same cluster. After the core points are given clusters, all non-core points are assessed. The non-core points directly connected to a core point will also be assigned the same cluster as the relevant core point; however, the non-core point will not be further used to expand the cluster. The non-core points that are too far away to be assigned a cluster will be classified as outliers. Figure 2.4 shows an example of DBSCAN clustering. The dark red points are core data points, while the lighter red and the green are non-core data points. *minPts* in the example is 4, and *eps* is marked in Figure 2.4a and 2.4b. The light red points are close to core data points and will join the cluster but do not have enough directly connected data points to be core data points. The green points are classified as outliers. Figure 2.4c marks the cluster created from DBSCAN.

Advantages of the DBSCAN algorithm are that it can cluster data of arbitrary shapes, and it is robust towards outliers. However, it is sensitive to the two parameters *minPts* and *eps* and does not work with data sets with altering densities [AD15].

2.3.2 Supervised learning

Supervised machine learning algorithms classify data based on prior information [KZP+07]. The algorithms use a set of instances with known labels to further predict the classes of the instances with unknown labels. A label is the correct output for an instance of the data set. The set of instances with known labels is called the training set, while the data to be classified is called the testing set.



(a) A core data point with at least $minPts = 4$ points closer than eps

(b) A non-core data point with less than $minPts = 4$ points closer than eps

(c) The cluster created from the algorithm

Figure 2.4: DBSCAN example

The classification rules are built from one or more classes depending on the training data set. For example, if there is information about only one class in the data set, the algorithm should be a one-class algorithm, but the algorithm can also be based on two or more classes. No supervised learning algorithms were used in this study due to the lack of ground truth needed for training data. Some supervised ML algorithms that could have been used are Naive Bayes, k-NN, decision trees, random forest, SVM, and neural networks.

Chapter 3
Related Work

This chapter describes the state-of-the-art and related work for the thesis. Firstly, papers concerning predator detection will be presented. Next, a collection of studies that have utilized graph theory for detection problems related to this thesis will be discussed before presenting a research project that studied a graph-theoretical approach to predator detection. The state-of-the-art was reviewed, and a selection of relevant background material was described in the project preceding this thesis [Aar21]. This is amended with a review of a few papers that have been discovered after the project.

3.1 Predator detection

Former research on detecting predators in platforms for children has used different approaches and methods. Many studies have analyzed the linguistic features in the text written between a child and a predator. Various methods have been developed for detecting predator conversations and identifying predators based on the written text. Other papers focus on the different users' behavior and author analysis for the same overall goal. Some of the most promising research for predator detection will be presented in this section.

The research done by Villatoro et al. [VJE+12] placed 1. at the PAN-2012 competition [IC12] with an $F_{0.5}$ score of 0.93. The study uses a two-stage approach for detecting predators; the Suspicious Conversations Identification (SCI) stage and the Victim from Predator disclosure (VFP) stage. For the SCI stage, Bag of Words (BoW) was used as a feature extraction method, employed with both binary and Term Frequency-Inverse Document Frequency (TF-IDF) weighting schemes. Similarly, for the VFP stage, the study tested BoW with binary and TF-IDF employment. Two different classification algorithms were tested: Support Vector Machine (SVM) and Neural Networks (NN). Villatoro et al. achieved the best result using NN classification and BoW with binary weighting scheme on both the SCI stage and

VFP stage, which gave an $F_{0.5}$ score at 0.93, placing the research at the top of the competition.

Cardi and Rebedea [CR17] is a more recent study using the two-stage classification approach. The study uses the SVM classifier to detect the predator conversations in the first stage, and a Random Forest classifier is used for the identification in the second stage. For feature extraction, behavioral features and interactional attributes are studied alongside BoW features with binary weighting. Behavioral features are not directly related to the words used by the different participants but rather the behavior of the users. Examples of behavior features are question ratio (the number of questions a user asks), slang ratio, and sexual word ratio. The paper results show that the inclusion of behavioral features is beneficial compared with the user of only lexical features.

Fauzi and Bours also studied a two-step approach for Predator identification; however, the research included ensemble strategies aiming to improve detection accuracy [FB20]. For both the Predatory Conversations Identification (PCI) and Victim from Predator Distinction (VPD) phases, different feature sets and classification methods were tested. In addition, ensemble methods were introduced, and comparisons between classification with and without ensemble methods were studied. Ensemble methods take multiple classifiers, which "vote" to classify the text. In the research, the ensemble classifiers vote for predator or non-predatory, both with hard and soft voting, and the classification is then based on the votes. The ensemble method gave promising results in the PCI stage with an $F_{0.5}$ score of 0.99, which outperformed the other classifiers working alone. For the VPD stage, the Naive Bayes classification performed better than the ensemble methods.

Another approach for detecting predators through lexical analysis is author attribution, which is the "task of assigning an author to an unknown text" [MDRR19]. Author attribution can be done in several ways, for instance, with keystroke dynamics or linguistic analysis. Much of the author's analysis research for predator detection is based on lexical analysis. One challenge of author attribution in chat networks is the short length of the text messages, which leads to precise classification being more complex. Several papers on author attribution for short messages have used the merging of all messages from one person into one larger text as a method to circumvent this challenge [Bou11; MP13].

Bours and Kulsrud [BK19] compare the author-based approach with the message- and conversation-based approach for detection. The method includes merging the messages from one author into one larger text. The author-based method obtained an $F_{0.5} = 0.891$ by using Neural Networks classification and TF-IDF features. The research paper concludes that the Author-Based Detection method gives the most

promising results alongside the Conversation-Based Detection method.

Misra et al. [MDRR19] used author attribution with the use of Convolution Neural Networks (CNN). The research proposed two models for the author analysis. Both models did the author analysis with CNN, but one model focuses on Authorship Attribution (AA-CNN) and the other on Predator Classification (AA-CNN-PC). The proposed method showed to be comparable to state-of-the-art research and a simpler method compared with previous work.

One of the main challenges of analyzing textual messages is many misspellings, slang, and informal language. Cheong et al. describe some of these challenges in a paper on detecting predatory behavior in game chats [CJG+15]. Cheong et al. used a data set from the game MovieStarPlanet, which is the only paper to our knowledge, that studies predator detection and uses real-life data in the research. The writing style is highlighted as one of the main challenges in the study, as there seems to be a high level of misspellings, slang, errors, and meaningless symbols in game chats. Another challenge discussed in the paper is that the nature of the game causes normal users to have a language that may be similar to the predator's language. Ordinary users are likely to form virtual relationships in the game, and typical chats revolve around being boyfriends, girlfriends, introducing role play, etc. Thus, it gets more challenging to recognize the predator in the game chats.

3.2 Graph-based network analysis

Graph theory has been used to analyze and detect anomalies in many different research projects and areas. In this study, we use graph theory in a game chat, which can be defined as a social network. Several papers study graph features, characteristics, and how to detect anomalies and outliers in social networks. For example, Panzarasa et al. [POC09] analyze the patterns of user behavior in a social network. One of the main takeaways from the study is that the studied social network has a "small-world" characteristic, which means that there are relatively small shortest paths between the majority of the nodes in the network. The paper also emphasizes the uneven characteristics among regular users. For example, some users are more popular or gregarious than the average user, which causes the graph to exhibit a fat-tailed degree distribution.

DeBarr and Wechsler [DW10] used social network analysis in their research on improving spam detection. The paper proposes to use the degree centrality of the sender's message transfer agent and the path length between the sender and receiver as graph features. The research classifies the test data with the machine learning algorithm, LogitBoost. The results show that including social network analysis improves spam detection significantly.

Similar to DeBarr and Wechsler, Fire et al. [FKE12] studied how to detect spammers and fake profiles in social networks with a graph-theoretical approach. Four features were extracted and used for the detection: the user's degree, the user's connected communities number, the number of connections between a user's friends, and the average number of friends inside connected communities. Decision trees and Naive Bayes were used as classifiers. To evaluate the performance of the proposed method, a random control group of most likely fake profiles was collected and assessed manually by an expert group. The proposed algorithm gave good results with F-scores up to 0.999. The paper concludes with the method being sufficient for small to medium-sized social networks.

Almaatouq et al. [ASN+16] also studied how to detect spammers. The research tests different kinds of features, content features, profile features, and social interaction features, which correspond to graph features. Degree-, density- and centrality values are examples of features included in the social interaction category. The performance of the classification of users is compared based on the different feature types. When using combinations of the different categories, the classification is the most precise. Social interaction features outperform the other categories when looking at the performance individually. Different supervised learning algorithms are also tested for classification in the research. The decision tree performed the overall best out of the classification methods tested.

Johnsen [Joh16] analyzed cybercrime networks seeking to identify interesting users within a social network. He studied which features could be used to identify central individuals in the network and how graph theory could be used in identification. The method included using several features and neighborhood approaches: k-NN and e-neighborhood. Some of the different features analyzed in the thesis were degree, in-degree, out-degree, betweenness, and closeness. The results suggested that betweenness, closeness, and in-degree were the features that gave the most accurate indications of the interesting users in the social network.

3.3 Graph-based predator detection

Matteini Palmerini [Mat21] is, to the best of our knowledge, the only research on detecting sexual predators online through graph analysis. He studied if a graph theoretical approach could contribute to detecting sexual predators. The data set used in the study was gathered from an online game where users communicate in private chats. Two classification approaches were used; the neighbors' approach and the cliques approach. Both classification methods used betweenness- and closeness centrality features.

The neighbors' approach uses classification in two steps, firstly creating subsets of

the graph and then subgraphs. The first step uses the difference between unweighted in-degree and out-degree values to create subsets. The step divides the data into users that behave more like spammers, regular users, and users that mainly receive messages. The next step divides the normal users based on three attributes: the total number of edges, the messages per link, and the variation between in-degree and out-degree values. Outliers were detected by using both betweenness- and closeness centrality features. The two different types of centrality measures showed different kinds of outliers. With betweenness centrality, outliers had high centrality values but a low centrality value when closeness centrality was used. The paper concludes that the neighbors' approach is appropriate for getting a network's local views.

The cliques' approach is based on the assumption that a clique is the most realistic represent a group chat in the game. Betweenness- and closeness centrality was calculated concerning the cliques and used to find outliers. The same result regarding betweenness and closeness centrality was found in the cliques' approach: betweenness centrality showed outliers with high centrality values, and closeness centrality showed low centrality values for the outliers. The cliques' approach was concluded to suit best as a global approach.

One limitation of the thesis was the data set that was used. All sensitive information in the game was encrypted, and there was little information about the game itself. Thus, the object of the project was moved to "test key features for abnormal behaviors detection investigating the number of edges and the number of the messages" [Mat21]. There was no conclusion on how well the methods described would detect predators.

Chapter 4
Methodology

This chapter describes the methodology used for the master thesis. The preprocessing phase will be described, followed by a section on how the data set was manually studied before choosing the feature set. Section 4.3 goes more into detail about how the feature set was formed. Next, section 4.4 describes how the Machine Learning (ML) algorithms were implemented and tested. Lastly, the limitations of the thesis will be presented.

4.1 Preprocessing

The data set used in this master thesis is from the online children's game MovieStarPlanet. The game is aimed at children between the ages of 8 and 15. In the game, the users play a character in a virtual world, where they can play games and meet other users. The game allows for private chats and chats in groups and forums. A part of the game revolves around forming relationships and friendships with other users. Thus, many of the conversations in the chats revolve around topics such as relations and romance. To protect the users of the game, MovieStarPlanet filters out words that may appear inappropriate in chats[1]. Many of the text messages are characterized by circumventions of the filters, for instance, intended misspelling or separating of words.

MovieStarPlanet has provided a data set collected from the USA in over five months, between January and May 2021. The data set consists of chat data from private chats, such as text messages and usernames, from users of the game. No data used for the thesis is from group chats or chats in forums. The data set is not continuous over five months but split into five different continuous sets. In addition to the five-month data set, we got access to a data set from one separate day. This data set was accessible to us before we got the five-months data set. It was used to

[1] https://moviestarplanet.zendesk.com/hc/en-us/articles/115000385689

prepare the software used for visualization and clustering, and it was used for feature extraction. The data were collected was from 31/12/21.

The preprocessing of the data revolved around making the data anonymous. This was done outside of this master thesis by the thesis supervisor. All users were given a randomized ID, and the only information used for the predator detection was the IDs and the number of messages sent between different user IDs. The anonymous user IDs can only be linked to the original usernames by the supervisor. So if a node has abnormal behavior, it can be investigated further by reading the messages sent to and from the related user. None of the data used to examine if graph theory could be used for predator detection could identify any user of the game. However, the anonymized data provides enough information to create a graph of users and to study the properties of the different anonymous users. Figure 4.1 shows a part of the data file used. Column 1 and 2 shows user IDs, and the third column shows the number of messages sent from the user in column 1 to the user in column 2.

```
Source,Target,Weight
2,241265,20
3,22927,5
3,46217,196
3,112556,3
3,149033,6
3,162553,2
3,216174,4
5,18903,1
5,69610,5
5,157028,8
5,215491,1
5,227284,4
5,227680,1
5,242764,9
6,91667,1
6,123564,1
9,10975,15
9,19513,5
9,28261,2
9,46234,6
9,63112,3
9,87316,3
9,88262,1
9,96417,1
9,105473,3
9,118897,2
9,142865,4
9,145368,1
9,157459,3
9,161166,9
9,161167,1
9,205133,1
9,211460,6
9,217557,3
```

Figure 4.1: Snippet from the data set file

4.2 Study of the data set

At the beginning of the project, we had only access to a data set collected from one day of the game. The data set consists of 18 364 users and 1 130 468 messages. This data set was used to learn about the properties of the nodes of the game network. Different charts of properties were created to identify important aspects of the data further.

When using clustering algorithms, a set of features is needed. The feature set was based on looking at the graph, but also on assumed behavior of normal and anomaly users. To learn about the properties of a user in the game, we visualized parts of the graph. We used a data set from one day to visualize a network of chats between users. The graph was built with nodes representing the users, and with the number of messages sent between users as the weight of the directed edges. This network was too large to extract any helpful information. Therefore, we created subgraphs based on ego graphs for particular nodes in the graph from one day in the game. These ego graphs were created by choosing a node and a depth. Then the subgraph would consist of the node and its neighbors if the depth is 1, the same set of nodes plus all the user's neighbors' neighbors if the depth is two, and so on. The blue nodes around node n in Figure 4.2 shows n's ego graph with depth 1. The rest of the nodes except node 10 and 13 belongs to n's ego graph with a depth of 2. These subgraphs were able to give some more information about the nodes. It was, for instance, easy to see how many other users the node was talking to, how long the conversations were, how many of the nodes were communicating, and more.

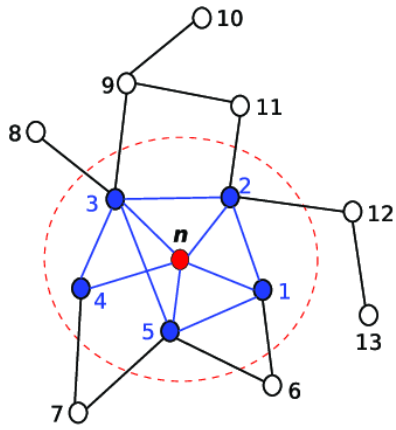


Figure 4.2: Example demonstrating ego graph around node n [AIP+18]

We used the python library Pyvis[2] to visualize the graph representation of the networks. The graphs were visualized with nodes labeled with their ID, and the

[2]https://pyvis.readthedocs.io/en/latest/index.html

edges between the nodes were labeled with the number of messages between the users. A color system was also created to separate central and less central nodes. Nodes with a higher out-degree value than 12 were colored red, and nodes with an out-degree value between 5 and 12 were colored purple. Nodes with out-degree between 1 and 5 were colored yellow, while nodes with only 1 or 0 outgoing edges were colored green. Figures 4.3 and 4.4 show one example of a visualized subgraph created from a random node, node 13458, with a depth of 2. The visualizations show more prominent edges where the conversation consists of relatively many messages compared to other conversations in the graph. Figure 4.4 demonstrates this property. When the larger data sets were studied, the boundaries for the colors in the visualizations were adjusted. Nodes with an out-degree value higher than 32 were colored in red. Nodes with an out-degree value between 7 and 32 were colored purple. Nodes with out-degree between 1 and 7 were colored yellow, while nodes with only 1 or 0 outgoing edges were still colored green.

Figure 4.3: Example of visualization. The graph is the ego graph for node 13458 with depth 2

.

When studying the different subgraphs, nodes with many neighbors, i.e., with a high degree value, stand out. Also, nodes with many incoming and few outgoing messages and vice versa are interesting. We created plots that could visualize the density of the degree of the nodes. Figure 4.5 shows a plot of the density of the degree, in-degree, and out-degree values from the one-day data set in the same plot. From the plot, we can see that most users had less than ten conversations that day. We made similar observations on the number of messages, where most conversations

Figure 4.4: Zoomed-in example of visualization. The graph is the ego graph for node 13458 with depth 2.

consisted of less than 50 messages that day. The weighted degree distribution, shown in Figure 4.6, is similar to the plot without weights. This observation can indicate that users with either many neighbors or with a lot of sent and received messages might be unusual.

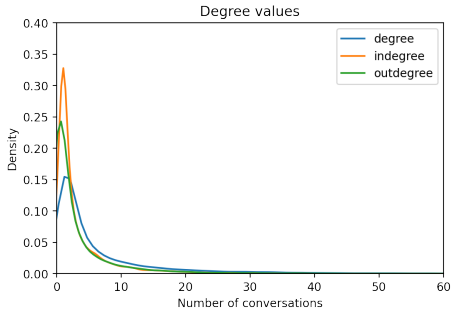


Figure 4.5: The density of degree values from one day in MovieStarPlanet

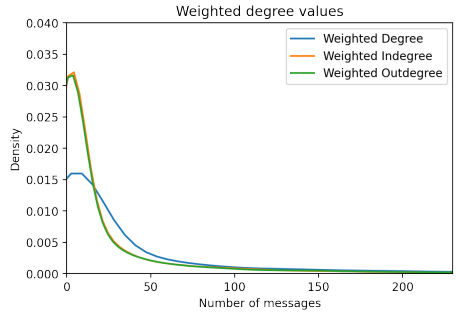


Figure 4.6: The density of weighted degree values from one day in MovieStarPlanet

We also looked at the number of incoming and outgoing messages related to the total number of messages for nodes individually. Figure 4.7 shows the distribution of incoming messages divided by the total number of messages from a node to all its neighbors and outgoing messages divided by the total number of messages. We

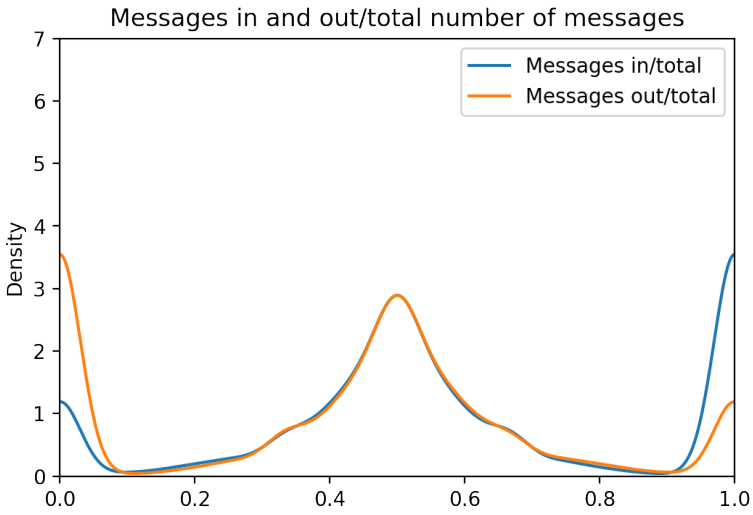


Figure 4.7: Number of incoming and outgoing messages divided by the total number of messages

observed that many users send approximately the same number of messages as they receive. In addition, some users only send messages, and many users only receive messages. The two different graphs in the plot are opposite from one another:

$$\frac{\#outgoingMessages}{\#totalMessages} = 1 - \frac{\#incomingMessages}{\#totalMessages}$$

Thus, there is no additional information gained by using both as individual features for the clustering algorithms.

Another node characteristic that was interesting to investigate was whether the node's neighbors were chatting with each other. If a user of the game is part of a group of friends, it is natural to think that that user's friends are also friends. Therefore, a node in the network with many connected neighbors might indicate that the node is a user playing with its group of friends. This property was visualized by using the Clustering Coefficient (CC). The CC compares the number of possible edges between neighbors of a node with the number of edges between the neighbors. Figures 4.8 and 4.9 show the density of the CC calculated with and without the weights of the edges. The plots show that most users did not have many neighbors talking with each other during the one day of the game. This property might change when the data set is collected over a more extended period.

Lastly, we studied the distribution of the number of messages in the different conversations belonging to a user. We divided the conversations into four categories based on the number of messages. The four categories were: conversations with

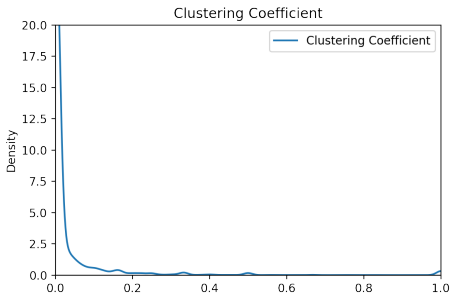


Figure 4.8: Clustering coefficient from one day in MovieStarPlanet

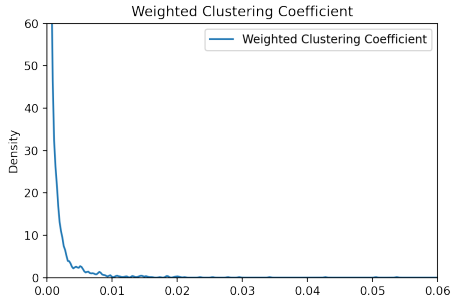


Figure 4.9: Weighted clustering coefficient from one day in MovieStarPlanet

one message, conversations with two to five messages, conversations with six to 20 messages, and conversations with more than 20 messages. For each node, a vector of four values was calculated. For instance, a node with the vector $(4, 6, 2, 1)$ has four conversations containing one message, six conversations with less than or equal to five messages, two conversations with less than or equal to 20 messages, and one conversation with more than 20 messages. A similar vector was created but only containing the number of outgoing messages, and one additional vector containing the same data but in percent was calculated. The corresponding vector in percent would then be $(0.30, 0.47, 0.15, 0.08)$. Different plots were created to visualize the normal behavior of a node. Figures 4.10 and 4.11 show the message distribution plots. The plots are a bit cluttered, but it is possible to see that it is more usual with short conversations than longer ones. The box plot also indicates that most users have only a few conversations at all lengths. These properties may change when the data set is collected over more extended periods.

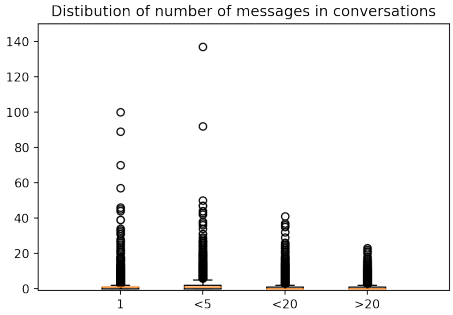


Figure 4.10: Box plot of the distribution of messages

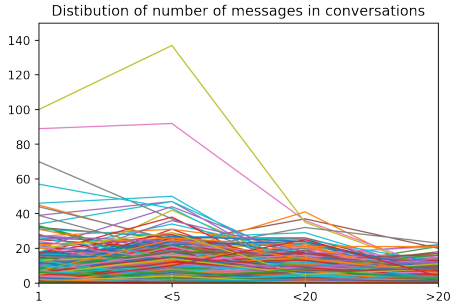


Figure 4.11: Plot of distributions of messages from all nodes

Figures 4.12 and 4.13 show the data but with percentage. It shows that conversations consisting of one message are a large part of the conversations for many

users. The more extensive conversations are usually a smaller part of the users' conversations. The plot shown in Figure 4.13 is more challenging to analyze, and gives little useful information.

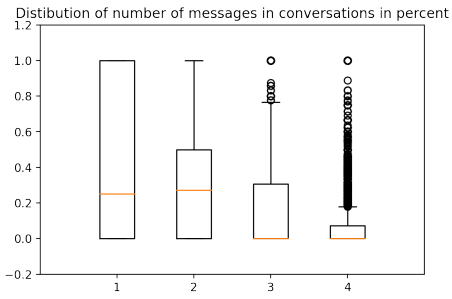


Figure 4.12: Box plot of percentage of message distribution

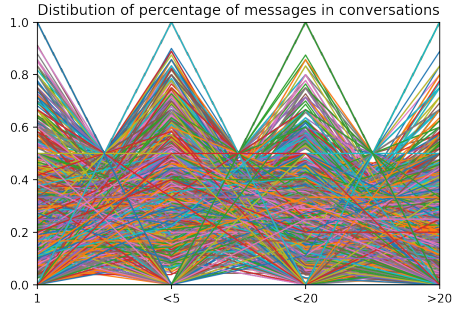


Figure 4.13: Plot of the percentages of message distribution

4.3 Feature extraction

In the feature extraction phase, 22 features were selected to be tested in the clustering algorithms. The feature extraction were based on the analysis of features from the previous section.

The first six features were related to the node's degree. The *degree* of the node is the number of conversations a user has or the number of neighbors of the node representing the user. *In-degree* and *out-degree* were also included as features. The in-degree value corresponds to the number of conversations where messages are sent to the user, and the out-degree value corresponds to the number of conversations where messages exist sent from the user. In addition to these values, the corresponding values with weights were included, namely *weighted degree*, *weighted in-degree*, and *weighted out-degree*. The weighted degree is the number of messages sent to and from the user. The weighted in-degree value is the number of messages the user received. The weighted out-degree value is the number of messages the user sent to other users.

The next feature is the *proportion of outgoing messages*. It is also based on the degree values, the weighted out-degree value related to the weighted degree. The value is calculated with Equation 4.1.

$$\frac{degW_{out}(n)}{degW(n)} \tag{4.1}$$

$degW_{out}(n)$ is the number of messages sent from node n and $degW(n)$ is the total number of messages sent between node n and its neighbors.

The two next features is *Clustering Coefficient (CC)* and *weighted CC*. The python library, NetworkX, was used to calculate values[3]. The CC and weighted CC are calculated with Equations 4.2 and 4.3, respectively.

$$cc_n u = \frac{2T(u)}{deg(u)(deg(u)-1) - 2deg^{\leftrightarrow}(u)} \tag{4.2}$$

$$wcc_u = \frac{\sum_{vw}(\hat{w}_{uv}\hat{w}_{uw}\hat{w}_{vw})^{1/3}}{deg(u)(deg(u)-1) - 2deg^{\leftrightarrow}(u)} \tag{4.3}$$

$T(u)$ is the total number of edges between node u's neighbors, $deg(u)$ is node u's total degree value. $deg^{\leftrightarrow}(u)$ is the reciprocal degree of u, which is the proportion of edges in both directions related to the total number of edges incident on u. v and w are neighbors of node u, and $\hat{w}_{uv}$ is the normalized weight on the edge, calculated with the maximum weight in the network: $\hat{w}_{nu} = w_{nu}/\max(w)$.

The next feature is called *non-common neighbors*. For all neighbors of a node n, the non-common neighbors calculates the number of neighbors the neighbor has that is not shared with node n. The number is calculated for each node and is then divided by the degree of the node n. To exemplify, if a node is a part of a close group of friends that all are users of the game, it is likely that all the users talk with each other, and might talk to few other users. Then the non-common neighbors value would be small. However, if the node is not a part of a group of friends, the number will possibly be more prominent as its neighbors probably talk with other users than the node's neighbors.

The remaining features capture the distribution of the length of the conversations between a user and its neighbors. The size of a conversation is measured by the number of messages in a conversation. To measure this distribution, the conversations were split into four features; conversations consisting of one message, conversations of two to seven messages, conversations of eight to 32 messages, and conversations of more than 32 messages. Each node will then have four different values representing the distribution of the length of conversations. In addition, four new values only calculating the outgoing messages are included in the same manner. Lastly, we calculated the distribution in percent. So the first value would represent the number of conversations with one message related to the total number of conversations and equivalent for the other three features. The boundary values were adjusted from the one-day data set as the conversations between users are likely to get longer over time. The adjustment was made by looking at the number of messages in the five month

[3]https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.cluster.clustering.html#networkx.algorithms.cluster.clustering

data set. Table 4.1 sums up all feature values calculated for each node. The set of features from the table will be calculated for all nodes and form a feature vector used in the clustering algorithms.

Table 4.1: Initial set of features

Feature	**Equation**	**Explanation**
Degree	$deg(n)$	The number of conversations containing messages to or from user n
In-degree	$deg_{in}(n)$	The number of conversations containing messages to user n
Out-degree	$deg_{out}(n)$	The number of conversations containing messages from user n
Weighted degree	$degW(n)$	The number of messages sent to or from user n
Weighted in-degree	$degW_{in}(n)$	The number of messages sent to user n
Weighted out-degree	$degW_{out}(n)$	The number of messages sent from user n
Proportion of outgoing messages	$\frac{degW_{out}(n)}{degW(n)}$	The number of outgoing message divided by the total number of messages
Clustering Coefficient	$\frac{2T(u)}{deg(u)(deg(u)-1)-2deg^{\leftrightarrow}(u)}$	The degree to which node n's neighbors are connected to each other
Weighted Clustering Coefficient	$\frac{\sum_{vw}(\hat{w}_{uv}\hat{w}_{uw}\hat{w}_{vw})^{1/3}}{deg(u)(deg(u)-1)-2deg^{\leftrightarrow}(u)}$	The degree to which node n's neighbors are connected to each other in a weighted graph
Non-common Neighbors	$\frac{non\_common\_neighbors(n)}{deg(n)}$	The sum of node n's neighbors' neighbors that are not directly connected to the node n, divided by the degree of the node.
#Conversations of 1 message	$deg_1(n)$	The number of conversations with node n consisting of one message

#Conversations of less than 7 messages	$deg_{\leq 7}(n)$	The number of conversations with node n consisting of two to seven messages
#Conversations of less than 32 messages	$deg_{\leq 32}(n)$	The number of conversations with node n consisting of eight to 32 messages
#Conversations of more than 32 messages	$deg_{\geq 32}(n)$	The number of conversations with node n consisting of more than 32 messages
#Conversations of 1 outgoing message	$deg_{out,\leq 1}(n)$	The number of conversations consisting of one or zero outgoing message from node n
#Conversations of less than 7 outgoing messages	$deg_{out,\leq 7}(n)$	The number of conversations consisting of two to seven outgoing message from node n
#Conversations of less than 32 outgoing messages	$deg_{out,\leq 32}(n)$	The number of conversations consisting of eight to 32 outgoing messages from node n
#Conversations of more than 32 outgoing messages	$deg_{out,\geq 32}(n)$	The number of conversations consisting of more than 32 outgoing messages from node n
Proportion of conversations of 1 message	$\frac{deg_1(n)}{deg(n)}$	The proportion of conversations with node n consisting of one message
Proportion of conversations of less than 7 messages	$\frac{deg_{\leq 7}(n)}{deg(n)}$	The proportion of conversations with node n consisting of two to seven messages
Proportion of conversations of less than 32 messages	$\frac{deg_{\leq 32}(n)}{deg(n)}$	The proportion of conversations with node n consisting of eight to 32 messages

Proportion of conversations of more than 32 messages	$\frac{deg_{\geq 32}(n)}{deg(n)}$	The proportion of conversations with node n consisting of more than 32 messages

4.4 Implementation and testing of clustering algorithms

Throughout testing the clustering algorithms, we used different data sets. Firstly, the data set from one day in the game was used to study the properties of the nodes in the network and to get an understanding of which features should be included in the testing. Later, a non-continuous data set from 2 months was used to test the algorithms on a more significant amount of data. The data set from 2 months is a subset of a larger data set collected from a five-month period. While testing clustering algorithms for predator detection, we used only data sets collected from the five months. The data collected over five months was not collected continuously, meaning that there were gaps of several days where no data was collected. Hence, we both run the cluster algorithms on the whole non-continuous five-month set, and five smaller sets consisting of continuous data from smaller periods. Table 4.2 shows an overview of all the different data sets and some of their properties. Although the smaller periods are under one month of data, they will be referred to as month 1, month 2, and so on as they were collected in their respective months. All the data sets except the one from one day are from the same five months and contain many of the same users. The different users have the same user ID across the different data sets from the five months.

Table 4.2: Overview of the different data sets

Data set	**Time period**	**Continuous**	**Number of nodes**	**Number of edges**	**Number of messages**
One day	24 hours	yes	18 364	65 968	1 130 468
Five months	5 months	no	272 699	2 449 064	65 663 048
Two months	2 months	no	142 695	1 047 646	32 538 059
Month 1	6 days	yes	86 639	534 343	12 835 006
Month 2	8 days	yes	84 113	545 633	13 716 358
Month 3	9 days	yes	81 353	558 461	14 299 596
Month 4	7 days	yes	68 482	410 984	10 018 760
Month 5	11 days	yes	84 968	586 012	14 789 416

4.4.1 Implementation

Agglomerative clustering, BIRCH, k-means, mean shift, DBSCAN, and Gaussian Mixture Model (GMM) are the clustering algorithms tested for predator detection. All clustering algorithms were programmed in python using the Scikit-Learn library [PVG+11]. The programs that run the algorithms take in a CSV file containing nodes represented by a feature vector. The clustering scripts normalize the features and do clustering with Scikit-Learn. The normalization scales the different feature values, so all features share the same scale. The scaling was done with the StandardScaler from the Scikit-Learn library [PVG+11]. The normal score for a feature sample, x, is calculated with the formula:

$$z = (x - \mu)/\sigma,$$

where μ is the mean of all samples and σ is the standard deviation of the samples. The scripts then make two output files, one containing the nodes sorted by their cluster, and the other is a new CSV file containing the same rows and columns as the input file but with a new column named "Cluster", which contains the cluster assigned to the node. The next sections describe some details on how we implemented the different algorithms. All algorithms are explained in more detail in Section 2.3.1

k-means

The k-means algorithm takes the number of clusters as a parameter to the algorithm, and 10, 7, and 5 were tested as values for the number of clusters. The algorithm takes in an *init* parameter that decides how the initial centroids are calculated. The parameter was set to "k-means++", an algorithm to choose the initial centroid seeds, to make the algorithm converge faster than with random seeds. *n_init*, the number of times the algorithm is run with different centroid seeds, was set to 10. The maximum number of iterations in the algorithm was set to 300.

Gaussian Mixture Model

GMM takes the number of components as parameters, corresponding to the number of clusters. 10, 7, and 5 were tested for the number of components. When first running the algorithm with 10 clusters on the largest data set with data from all five months, there were no clusters that stood out. Therefore, we tested the algorithm with 15 clusters, aiming to get smaller clusters. Hence, the algorithm was run with 15 clusters instead of 7 and 5 clusters for the 5-months data set.

Agglomerative clustering

The agglomerative clustering algorithm takes the number of clusters as a parameter. The algorithm was tested with different parameters for the one-day data set but was

not used further on the larger sets as it was too slow. The Euclidian distance was used to calculate the distances.

BIRCH

The BIRCH algorithm takes a threshold value, branching factor, and the number of clusters as parameters. 1.5 was used as the threshold and 50 as the branching factor. 10, 7, and 5 were tested as values for the number of clusters. The agglomerative clustering algorithm is used for the last clustering part after the initial grouping of the nodes.

Mean shift

Mean shift does not need the number of clusters as input to the algorithm, as it iterates to the clusters converge. Instead, the parameter that is central for the mean shift is the bandwidth. This was calculated through a bandwidth estimator but later adjusted to avoid creating too many different clusters. The bandwidth used was 12. The maximum number of iterations was set to 300.

DBSCAN

DBSCAN takes *eps* (a distance threshold) and *minpts* (the minimum number of points required to form a dense region) as parameters. The algorithm does not require the number of clusters as input. The *minpts* was set to $\#features + 20$. *eps* was calculated using an elbow method on the nearest neighbors for the data points, which gave the value 3. The elbow method is a method that can be used to select the optimal value for a parameter. The *KneeLocator* from the kneed library [SAIR11] calculated the *eps* value. The distances were calculated with the Euclidean distance metric.

4.4.2 Testing

The two-months months data set was used at the start of the testing period to test how well the different algorithms could handle the larger data sets. Various parameters were tested with this data for different algorithms. We decided to drop the agglomerative algorithm for clustering in this phase as it was too slow.

DBSCAN was run with different values for *eps* and *minpts*. Using $eps = 3$ and $mipts = \#features + 20$, we got one large and one small cluster together with a group of outliers. This was the most useful clustering we achieved with DBSCAN, and hence these parameters were used with the other data sets as well.

Mean shift was also tested with different bandwidth parameters with the two-months months data set. An estimator was used to calculate the appropriate value

for bandwidth, but the algorithm produced over 100 different clusters. Hence, the bandwidth was adjusted up to 12. With this parameter, the algorithm created 22 clusters.

After testing the different algorithms with the two-months months data set, the five-months set was tested on the algorithms. The data set was a good way to get an overview of how the users behave over a more extended period. However, the fact that the data set contained data from 5 different periods with longer periods in between the data set made the large set a bit more challenging to analyze. The challenge of analyzing data collected over a long period is that users will behave more similarly over time. For example, while it may be abnormal behavior to have many different conversations in a short period, it is not unusual to have the same amount of different conversations over a more extended period. It was, therefore, easier to separate abnormal behavior in shorter periods. Furthermore, the fact that the large set is non-continuous may give the nodes different feature values than what is most suitable. Hence, the most effort went into analyzing the continuous data sets collected in the various months.

After the different algorithms on the various data sets were run, key numbers from the clustering were calculated. Since we had no ground truth, much of the analysis was based on assumptions and hypotheses. For example, clusters containing few nodes, large average degree, large average weighted degree, or mismatched values for the degree and weighted degree were considered suspicious. High or low values from the other features were also used to find anomalies in the data sets. The clusters that stood out were visualized to understand the properties of the clusters better. A group of randomly chosen nodes was visualized with their ego graphs from most of the clusters that stood out.

Some of the nodes were also investigated further to determine if there was predatory activity or any other illegal activity happening in the relevant conversations. For these nodes, conversations between the associated user and its neighbors were read through to identify if the user was a predator or engaged in other illegal activities. Before the chats were studied, the project's supervisor did rigorous anonymization of all users involved. As a result, there were no possibility to identify the users participating in the chats that were investigated in the thesis.

4.5 Limitations

The most prominent limitation of the thesis was the lack of ground truth. Initially, we planned to have a set of users that were classified as predators by MovieStarPlanet. However, we were not successful in getting the set and hence had no ground truth to work with. As a result, we were limited to only using clustering algorithms as we did

not have any data to train supervised learning algorithms. In addition to dropping the supervised ML algorithms, the lack of ground truth made assessing the clustering algorithms more complex. Instead of looking at where the classified predators would be clustered, we had to manually investigate all interesting clusters or nodes, reading chats from relevant users. This process is time-consuming and not very effective. Also, it is likely that we have missed interesting clusters as a result of this. However, the method used in the thesis is adequate for answering the main research question.

The feature script that calculates the features of the nodes initially took more than 25 hours to complete on the data set containing chat data from 2 months. The script was changed to only handle 1/8 of the nodes to shorten the period. The script was then run in 8 parallels, and it took about four hours to complete. The feature script created the list of nodes in unique ways from each time it was run, so not all nodes were calculated during the first run of the scripts. A new script to find the set of forgotten nodes was written, and the feature script was then run with that specific set. Overall, running scripts in parallel was essential for testing all the different data sets for predator detection.

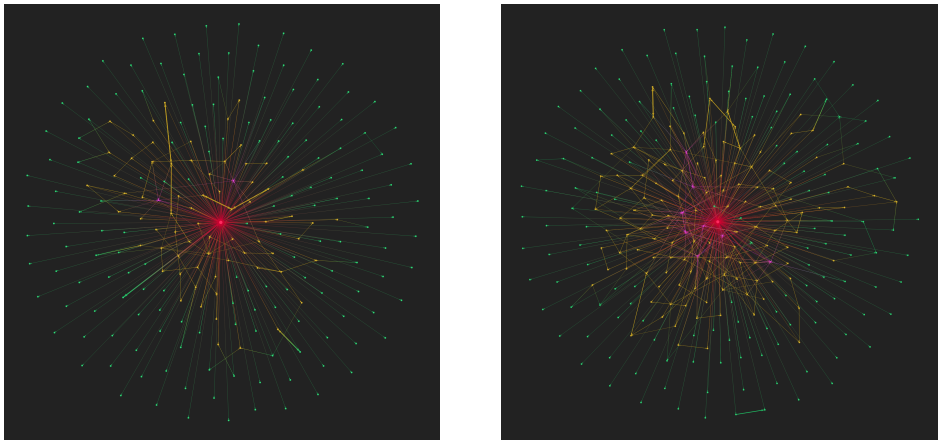
Chapter 5
Results

This chapter describes the results of testing the machine learning algorithms with different data sets and parameters. The results from the different algorithms are summarized in tables, where key data from the different clusters are provided. For each algorithm in each data set, the result clusterings with different parameters are shown and discussed. More extensive tables including more information about the distinct clusterings, are found in the Appendix A, B, and C. The extensive tables consists of key values from all features in all clusterings, together with standard deviation for each feature value.

Throughout the analysis of the results from the clustering, some nodes that are likely to be predators were found. We can not say with certainty whether a user is a predator, but we will refer to the users that are likely to be predators as predators, to simplify the read. In addition to predators, some users send out one or a few messages to a large set of users, often with the exact text sent to all users. These kinds of users will be referred to as spammers, and the activity they do will be referred to as spamming. To discuss the different findings and nodes, we will include visualizations of relevant nodes. This is done by displaying ego graphs based on the nodes, always with the depth of 1.

5.1 One-day data set

At the beginning of the thesis, we had access to the data set collected in one day. This data set was visualized and analyzed. Nodes 15451 and 8902 stood out from this set as they were the most prominent nodes when visualizing the whole data set. Both nodes had many neighbors, meaning the two users had a lot of different conversations in one day. Figures 5.1a and 5.1b shows the ego graphs from the two nodes that stood out. The two nodes are both colored in red as they have sent messages to more than 20 users. The neighbors do not seem to be tightly connected, which may result from the small time frame of the data set. For example, the neighbors of a node



(a) Node 15451 **(b)** Node 8902

Figure 5.1: Ego graph for two central nodes from the one-day data set

might usually chat with each other but might not have done that this specific day.

We tested the k-means algorithm with the feature set described in Table 4.1 on this data set to test the clustering software created. The algorithm was run with 10 clusters as a parameter. Table 5.1 shows the clustering. Nodes 15451 and 8902 were clustered with no other nodes in cluster 9. The nodes could not be confirmed as predators or regular users, as we could not find the related user from the randomized ID for this data set. This is also why we did not investigate the other algorithms. Table A.1 contains some additional values from the clustering.

From Table 5.1, cluster 9 stands out, both for being a small cluster and because of the significant average degree value and average weighted degree value. In addition, cluster 7 is relatively small and would be interesting for further investigation if we had access to the chats from this data set.

5.2 Five-months data set

As described in Chapter 4, the feature set consisted of all features listed in Table 4.1. With this set of features, the algorithms were run multiple times. The parameter giving the number of clusters was adjusted in iterations to optimize the clusterings. However, as we did not have any ground truth, it was challenging to compare the different clusterings effectively. Generally, it made the most sense to look at the clusterings with more significant numbers of clusters as they often gave smaller clusters with properties that were easier to recognize. Therefore, mainly the clusterings containing 10 clusters for k-means, GMM, and BIRCH were studied,

Table 5.1: k-means on data from one day

Cluster	Number of nodes	Average degree	Average weighted degree	Average weighted out-degree	Average CC
10 Clusters					
0	2294	2.68	8.07	4.95	0.01
1	1678	14.46	202.79	102.77	0.04
2	1682	3.24	132.75	65.83	0.03
3	3886	1.33	1.85	0.09	0.0
4	4759	2.3	9.62	4.4	0.01
5	224	2.44	75.18	36.92	0.88
6	349	23.74	891.21	451.87	0.04
7	176	45.82	401.49	207.28	0.02
8	3314	3.33	28.51	13.76	0.03
9	2	252.0	1198.0	676.5	0.0
The entire data set					
	18364	4.38	61.56	30.78	0.03

together with DBSCAN and mean shift which does not require the number of clusters as input. The clusterings with a smaller number of clusters were studied to some degree and compared with the 10-cluster clusterings.

The rest of this section presents the results from clustering the data collected over five months. Firstly, results from testing the clustering algorithms on the complete data set will be given. This data set is not continuous and contains data from 5 different continuous periods. We will then analyze the five smaller continuous periods. When analyzing the complete data set from five months, it became evident that time played an important role. It was more challenging to detect abnormal behavior when the data set was collected over a long and noncontinuous period. The clusters became more extensive and had more similar traits. Therefore, more effort went into analyzing the smaller individual continuous data sets. When working with this set, we had access to the chats belonging to the user represented by a randomized ID.

5.2.1 Data set from all five months

k-means, GMM, and BIRCH were the only algorithms run on this data set. This is because the remaining algorithms were too slow for a data set of that size. The results of the different clustering algorithms are summarized in tables and discussed

in this section. Appendix B contains more extensive tables from the clustering with the complete five-months data set.

k-means

Table 5.2 shows an overview of the clusterings made by the k-means algorithm with 10, 7, and 5 as the number of clusters. An extended table is found in Table B.1.

Table 5.2: k-means on data from 5 months

Cluster	Number of nodes	Average degree	Average weighted degree	Average weighted out-degree	Average CC
Clustering 1: 10 Clusters					
0	60477	2.13	10.78	4.94	0.01
1	3750	161.02	3975.94	1991.4	0.04
2	33241	2.84	39.87	20.05	0.02
3	790	212.79	15067.58	7483.89	0.04
4	56312	1.49	2.08	0.12	0.0
5	87363	8.74	111.86	56.7	0.04
6	248	538.91	10645.54	5322.83	0.02
7	9323	2.73	262.7	129.77	0.03
8	14855	56.93	1429.15	718.29	0.04
9	6340	2.77	104.66	51.45	0.86
Clustering 2: 7 Clusters					
0	76328	1.83	4.48	1.6	0.01
1	3708	164.89	5392.81	2700.76	0.04
2	90250	3.98	27.32	13.34	0.02
3	14834	60.51	1435.46	720.89	0.04
4	591	387.21	15119.6	7491.94	0.03
5	69957	8.2	133.35	67.63	0.04
6	17031	3.19	193.83	96.21	0.38
Clustering 3: 5 Clusters					
0	9557	101.85	2873.75	1441.55	0.04
1	77453	1.84	4.13	1.4	0.01
2	85774	3.3	17.48	8.48	0.03

3	1313	303.37	11322.85	5623.38	0.03
4	98602	10.82	218.18	109.89	0.09
The entire data set					
	272699	10.5	240.79	120.39	0.04

When running the k-means algorithm on the data set with 10 clusters, clusters 3 and 6 stood out based on their small sizes. Table 5.2 shows that the nodes in both clusters have approximately 20 and 50 times as high degree values on average, compared to the rest of the data set. 5 randomly chosen nodes from cluster 6 were visualized. Node 146547 is one example of the nodes from cluster 6, and the visualization of the node's ego graph is shown in Figure 5.2. As the figure shows, the node has many neighbors, and many neighbors are also connected.

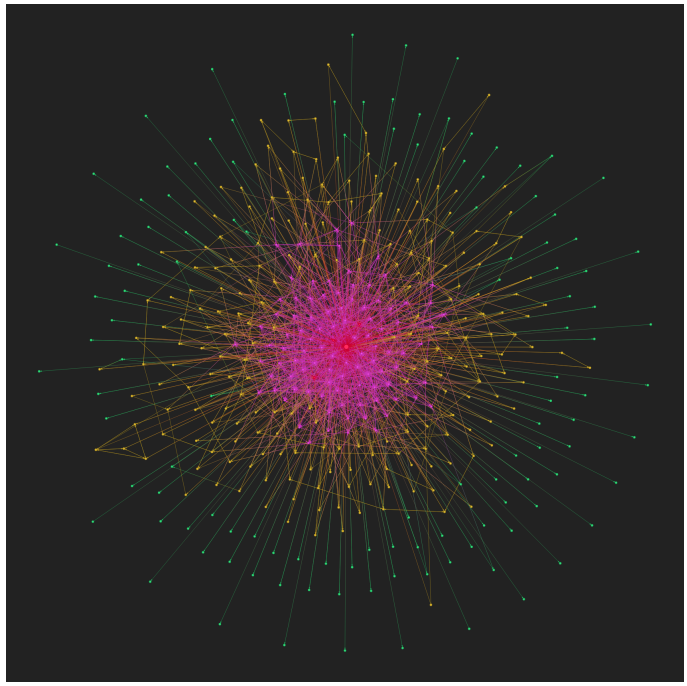


Figure 5.2: Ego graph for node 146547

Cluster 3 is the second smallest cluster, with 790 nodes, only containing around 0,3% of the nodes in the data set. The key values from Table 5.2 show that these nodes have smaller ego graphs than the nodes in cluster 6. The neighbors of the nodes in this cluster also seem to be relatively tightly connected. This shows in

the average CC value, which is 0.04, relatively high for nodes with many neighbors. Figure 5.3 shows one example of a node from cluster 3.

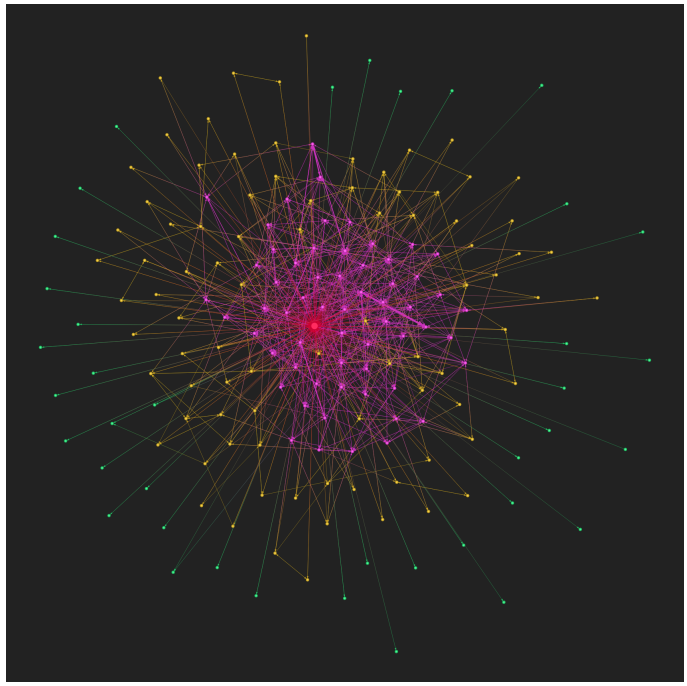


Figure 5.3: Ego graph for node 904994

Gaussian Mixture Model

Table 5.3 shows the overview of the clusters created by running the GMM algorithm on the five-month data set. The algorithm was firstly run with 10 clusters. However, the result from the clustering was ten relatively large clusters. Therefore, we tested the algorithm with 15 clusters instead of 5 and 7. Table B.2 shows some additional values for the clusterings.

Table 5.3: GMM on data from 5 months

Cluster	Number of nodes	Average degree	Average weighted degree	Average weighted out-degree	Average CC
Clustering 1: 15 Clusters					
0	11100	6.1	109.68	56.2	0.07
1	16161	1.3	3.12	0.06	0.0

2	4528	5.9	793.11	391.38	0.22
3	48321	1.13	1.13	0.0	0.0
4	38375	2.48	21.47	10.71	0.0
5	10382	31.08	297.89	149.35	0.03
6	8251	81.43	1438.16	722.51	0.04
7	4469	178.8	6498.82	3231.48	0.04
8	9830	2.37	51.17	26.07	0.45
9	29654	8.23	44.23	21.78	0.14
10	7992	31.12	1324.23	671.1	0.06
11	33116	1.25	4.47	2.47	0.0
12	12126	13.4	219.99	112.44	0.05
13	4092	1.06	134.23	66.62	0.0
14	34302	2.38	4.69	2.52	0.0
Clustering 2: 10 Clusters					
0	30262	8.08	55.56	27.4	0.1
1	8267	6.85	585.63	292.34	0.18
2	81539	1.71	4.3	2.07	0.0
3	14304	45.27	864.7	436.97	0.04
4	50254	1.16	1.22	0.0	0.0
5	19232	17.51	229.18	116.69	0.04
6	18599	2.44	53.04	26.9	0.32
7	5686	115.1	5886.81	2935.03	0.05
8	40681	2.76	22.73	11.33	0.0
9	3875	146.92	1694.97	844.49	0.02
The entire data set					
	272699	10.5	240.79	120.39	0.04

Neither the clustering with 10 nor 15 clusters gave any clusters that stood out from the rest. Therefore, we spent no additional time analyzing the results of this algorithm on the large data set.

BIRCH

Table 5.4 shows key values from clustering with the BIRCH algorithm with the five-month data set. More details about the clusterings are found in Table B.3.

Table 5.4: BIRCH on data from 5 months

Cluster	Number of nodes	Average degree	Average weighted degree	Average weighted out-degree	Average CC
Clustering 1: 10 Clusters					
0	268	382.76	20308.22	10050.93	0.03
1	21189	56.09	1399.7	702.81	0.04
2	502	317.77	6096.02	3016.91	0.03
3	1860	181.44	3931.03	1969.98	0.04
4	59	701.44	7556.44	3794.73	0.02
5	387	158.66	14801.19	7399.09	0.04
6	2	2.5	2399.0	1312.5	0.75
7	1	2701.0	13738.0	6146.0	0.01
8	248344	3.91	56.05	27.93	0.04
9	87	2.55	906.17	415.09	0.85
Clustering 2: 7 Clusters					
0	60	734.77	7659.47	3833.92	0.02
1	23049	66.21	1603.97	805.07	0.04
2	502	317.77	6096.02	3016.91	0.03
3	268	382.76	20308.22	10050.93	0.03
4	248431	3.91	56.35	28.06	0.04
5	387	158.66	14801.19	7399.09	0.04
6	2	2.5	2399.0	1312.5	0.75
Clustering 3: 5 Clusters					
0	889	248.51	9885.56	4924.56	0.03
1	60	734.77	7659.47	3833.92	0.02
2	248433	3.91	56.36	28.07	0.04
3	268	382.76	20308.22	10050.93	0.03
4	23049	66.21	1603.97	805.07	0.04
The entire data set					
	272699	10.5	240.79	120.39	0.04

The BIRCH algorithm gives several smaller clusters. Cluster 7 is the smallest

cluster containing only one node. This node has 2701 neighbors, meaning that the related users have conversations with 2701 different users. The node from this cluster is node 52855. The node is visualized in Figure 5.4 where we see that the ego graph for the node is extensive. By reading the chats from this user, it was evident that it was a spammer. However, no illegal sexual activity was detected.

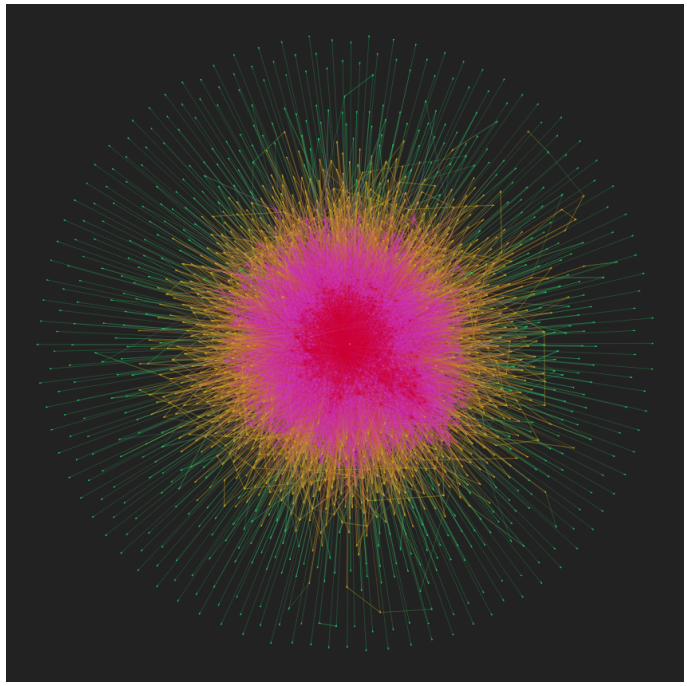
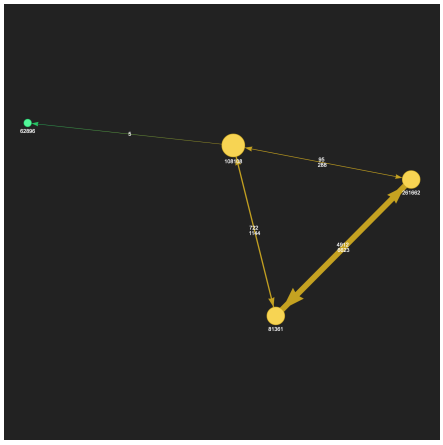


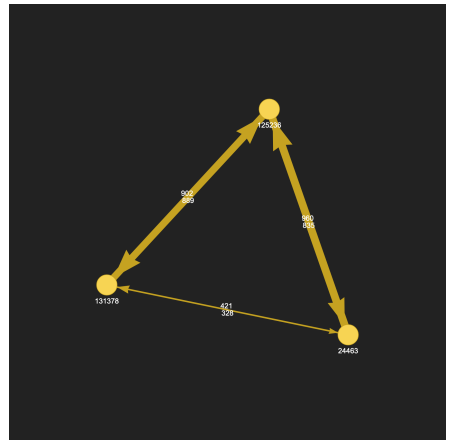
Figure 5.4: Ego graph for node 52855

Cluster 6 in clustering 1 is the second smallest cluster with only two nodes. The nodes in this cluster have, in contrast to cluster 7, only 2.5 neighbors on average. Despite the low number of nodes, the average weighted degree is 2399, which means that the nodes have few but long conversations. The average CC is the value that stands the most out. This is at 0.75, which is significantly higher than the average CC value for the entire set at 0.04. Figure 5.5 shows the two nodes. In node 24463, the average CC value will be 1 as all neighbors to the node are connected. Node 108108 has CC= 0.5, which also is relatively high. It is easy to achieve a higher CC value for nodes with few neighbors, and it does not make the relevant users more suspicious.

Apart from clusters 6 and 7, several of the other clusters could be interesting to investigate further. All clusters except 1, 3, and 8 are small enough to examine. When comparing the clustering with clustering 3, most of the small clusters were



(a) Ego graph for node 108108



(b) Ego graph for node 24463

Figure 5.5: Two nodes from cluster 6 in clustering 1 for using BIRCH

in clusters 0, 1, and 3. These clusters are relatively small, and the nodes in these clusters have many neighbors and might be interesting nodes. However, since the data set is so large, it is expected that many users will have conversations with many different users. Hence, no further investigation was done with this data set.

5.2.2 Data sets from individual months

The data sets from individual months are five different sets from different periods that all are collected continuously. Therefore, for each algorithm, results from month 4 are presented as an example. However, all months were studied, and the resulting clusterings are provided in Appendix C. For these data sets, we studied the results closer and investigated written messages between anonymized users to identify predators.

k-means

Table 5.5 shows a overview of the clusterings made by the k-means algorithm with 10, 7, and 5 as the number of clusters. Tables C.1, C.2, C.3, C.4, and C.5 in the Appendix C shows the more extensive tables for all of the months.

Table 5.5: k-means on data from the 4th month

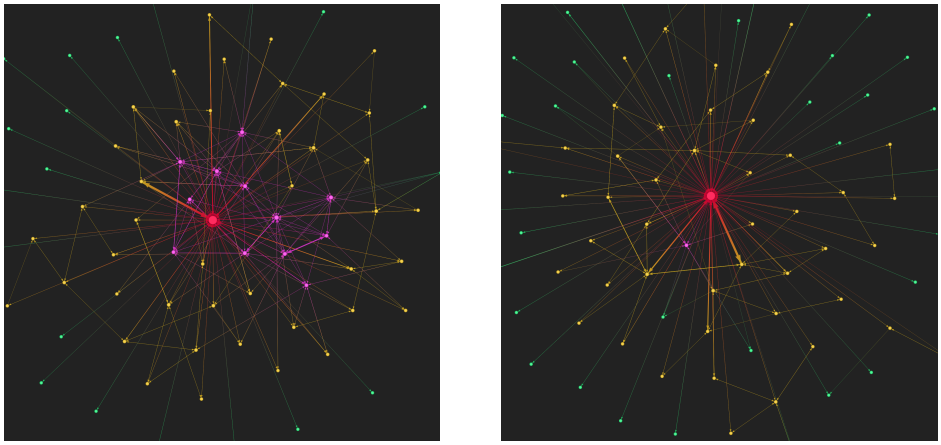
Cluster	Number of nodes	Average degree	Average weighted degree	Average weighted out-degree	Average CC
Clustering 1: 10 Clusters					
0	23649	3.27	22.05	10.84	0.02
1	218	103.68	6484.11	3217.22	0.03
2	3666	3.71	243.59	120.9	0.03
3	5501	27.07	445.86	224.06	0.05
4	12351	4.04	54.32	27.54	0.03
5	1549	47.55	1977.69	994.97	0.04
6	1370	2.81	93.33	46.99	0.83
7	459	122.85	1733.96	870.61	0.03
8	19716	1.68	3.64	1.33	0.01
9	3	880.33	2761.33	1465.33	0.01
Clustering 2: 7 Clusters					
0	21574	2.76	13.64	6.58	0.02
1	3967	40.96	1066.61	536.73	0.04
2	708	110.4	3645.05	1818.57	0.04
3	19745	1.67	3.28	1.11	0.01
4	1431	2.87	105.49	52.61	0.81
5	21054	6.75	127.7	64.08	0.04
6	3	880.33	2761.33	1465.33	0.01
Clustering 3: 5 Clusters					
0	38377	4.56	47.65	23.85	0.02
1	4406	39.19	1000.92	503.54	0.04
2	19592	1.6	3.21	1.08	0.0
3	714	113.8	3620.8	1807.18	0.04
4	5393	4.02	209.92	104.57	0.28
The entire data set					
	68482	7.04	146.3	73.15	0.04

Cluster 9 from clustering 1 and 6 from clustering 2 contains the same three nodes. These nodes are also found in other clustering algorithms and will be reviewed at

the end of this section.

Two other clusters from the first clustering are small and have a significant average degree value, clusters 1 and 7. The nodes from cluster 1 have 104 neighbors on average and over 6400 messages sent and received. The nodes from cluster 7 have 123 neighbors on average, but this cluster has a lower average weighted degree value of 1733.

Five randomly selected nodes from cluster 1 were investigated by visualizing the nodes and looking at the chats containing the nodes. Figure 5.6 shows nodes 201563 and 47777, which both were a part of the five investigated nodes from cluster 1. Both nodes had a couple of extensive conversations and many smaller ones. This pattern was also found for the other nodes investigated from the cluster. None of the nodes was considered a predator based on the text messages sent.

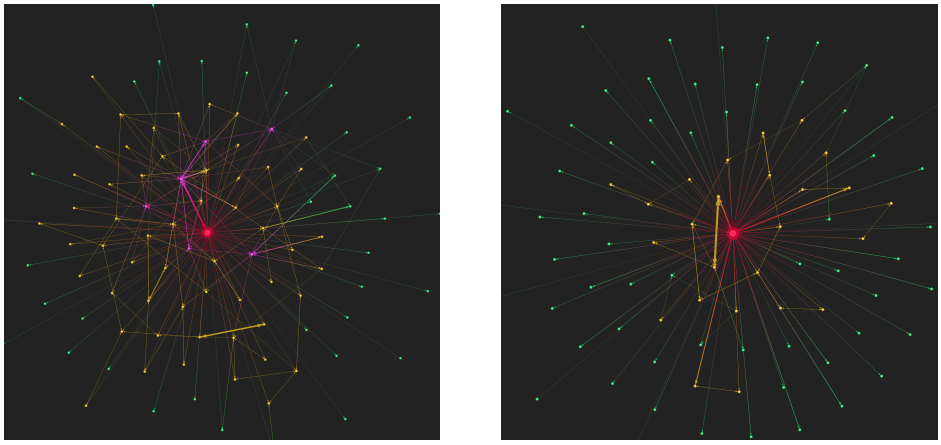


(a) Ego graph for node 201563 **(b)** Ego graph for node 47777

Figure 5.6: Two nodes from cluster 1 in clustering 1 using k-means

Two nodes, 246889 and 106005, from cluster 7 were studied closer, and the visualizations of the ego graphs from the nodes are shown in Figure 5.7. Node 106005 seemed to want pictures from the users it talked with. There is a possibility this user is a predator, but the user's age was challenging to confirm. The user claims to be 16, which is likely to be accurate, making the user less likely to be a predator. Node 246889 did not have chats indicating predatory, but we found spamming from the user.

From month 1, cluster 5 is the smallest, with 273 nodes. From this cluster, node 113379 was investigated by looking into the user's chats. This user was concluded to be a predator. The user claimed to be a 23-year-old male wanting sexualized



(a) Ego graph for node 246889 **(b)** Ego graph for node 106005

Figure 5.7: Two nodes from cluster 7 in clustering 1 using *k*means

relations with 13-year-old girls. Node 113379 is visualized in Figure 5.8. The figure shows that the node has a relatively similar pattern to the nodes that stood out from month 4. The user had 223 conversations that month and 1737 messages sent to or from the user. 28% of the user's conversations consisted of one message, and 40% of the conversations were between 2 and 7 messages. This means that almost 70% of the conversations are relatively short. Figure 5.9 shows a part of the chats between user 113379 and other users. All Personal Identifiable Information (PII) have been covered because of the anonymization in the figure of the chat. From the figures, we see that the user wants to add young girls on Snapchat[1] ("sc") and that he wants some sexualized role play, including a sister/brother relation. In the chats, he sent many messages to many users hoping that some would be interested in participating in the activities he wanted them to. The same node was also found in cluster 7 for the third, fourth, and fifth months, all relatively small clusters. He was highly active in all five months studied.

k-means clustering did also help disclose a suspect user from the fifth month. Node 6255 claimed to be a male, and he wrote messages to other female users indicating that he wanted sexualized pictures. The user's age was claimed to be around 16 to 17 years old, and he asked girls aged 12 years old for pictures. Figure 5.10 shows the visualization of the ego graph created by the user, and Figure 5.11 shows how the user communicates with other users. From the chats, we see that the user claims to be 17 years, and he wishes to "have fun" with girls at the age of 12. Based on the thesis' supervisor's extensive experience in reading predatory

[1] https://www.snapchat.com/

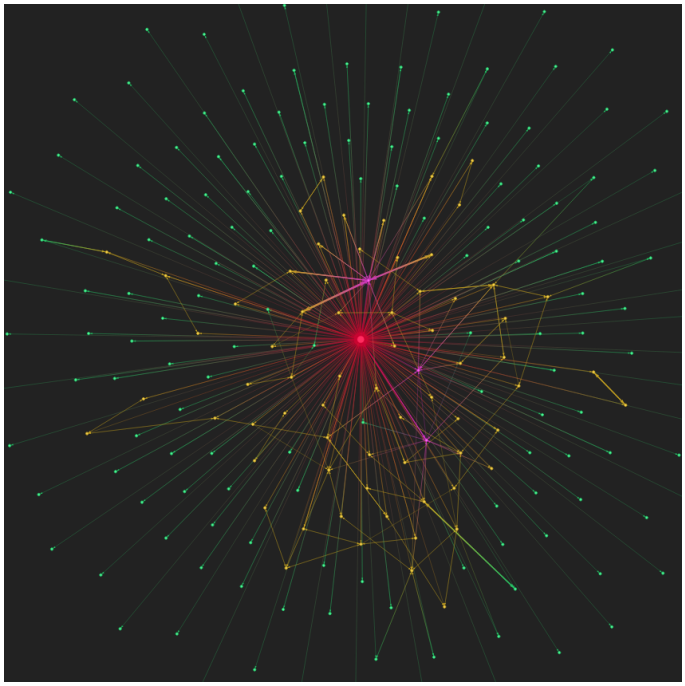


Figure 5.8: Ego graph for node 113379

conversations, "having fun" is a very strong indicator of pointing to sexual activities online. This user was, therefore, considered a predator. The user participated in 47 conversations that month, and he sent and received 248 messages. 35% of the conversations contained only one message, and 36% of the conversations consisted of between 2 and 7 messages.

No other users examined from the k-means clusterings were assumed to be predators. However, some other activities that are against the game's rules were found, such as giving out personal information and sexting.

When comparing clustering 1 with clustering 2 from Table 5.5, we saw that the clusterings had a similar form. 100% of the nodes from cluster 1 were in cluster 2 in the second clustering. Over 70% of the nodes from cluster 7 were also in cluster 2 in the second clustering. We see that the 7 clusters make a bit bigger clusters, but it seems to separate most of the nodes that stand out in the first clustering. The third clustering has a similar form to the second one, where all nodes from cluster 1 in clustering 1 are clustered in cluster 3, while 73% of the nodes from cluster 7 were clustered in cluster 3. So even if the clusterings in k-means have a smaller number of clusters, they seem to group the nodes in valuable manners.

```
01-02 : 02:14    113379 -> 203826: hey wanna b sister with benefits
01-02 : 02:23    203826 -> 113379: yeah
01-02 : 02:24    113379 -> 203826: u do
01-02 : 02:25    203826 -> 113379: yeah I guess
01-02 : 02:26    113379 -> 203826: do u have
01-02 : 02:26    113379 -> 203826: sc
01-02 : 02:26    203826 -> 113379: yeah
01-02 : 02:26    113379 -> 203826: add me
01-02 : 02:27    203826 -> 113379: what is ur thing
01-02 : 02:28    113379 -> 203826: [redacted]
01-02 : 02:28    113379 -> 203826: [redacted]
01-02 : 02:28    113379 -> 203826: [redacted]
01-02 : 02:29    203826 -> 113379: how old are u
01-02 : 02:29    113379 -> 203826: im 23 chu
01-02 : 02:35    113379 -> 203826: hello
01-02 : 02:35    203826 -> 113379: oh Im 13
01-02 : 02:36    113379 -> 203826: i dont mind
01-02 : 02:36    203826 -> 113379: but I do
01-02 : 02:36    113379 -> 203826: what u mean
01-02 : 02:37    203826 -> 113379: ur 23 Im 13...
01-02 : 02:47    113379 -> 203826: but u dont want a older brother
01-02 : 02:48    203826 -> 113379: Im good
01-02 : 02:48    113379 -> 203826: mmmm
01-02 : 02:49    203826 -> 113379: I already got a older brother dont need more lol
01-02 : 03:02    113379 -> 203826: ok
01-02 : 03:05    203826 -> 113379: do u understand
01-02 : 03:08    113379 -> 203826: ig
01-02 : 03:08    113379 -> 203826: i guess
01-02 : 03:08    203826 -> 113379: ok good
```

(a) Chat showing conversations with 13 year old user

```
01-03 : 12:52    113379 -> 232446: hey girl what race ru
01-03 : 13:19    232446 -> 113379: straight
01-03 : 13:20    232446 -> 113379: o

01-03 : 13:03    113379 -> 147425: hey girl wha race  ru
01-03 : 13:08    147425 -> 113379: what race?
01-03 : 13:08    147425 -> 113379: why that matter?
01-03 : 13:09    113379 -> 147425: kinda
01-03 : 13:09    147425 -> 113379: huh why?
01-03 : 13:11    113379 -> 147425: bc

01-03 : 13:05    113379 -> 146988: hey girl what race r u

01-03 : 13:11    113379 -> 88039: hey girl what race r u
01-03 : 13:11    88039 -> 113379: white why?
01-03 : 13:12    113379 -> 88039: r u mix
01-03 : 13:13    88039 -> 113379: no will u go away
01-03 : 13:14    113379 -> 88039: why
01-03 : 13:14    88039 -> 113379: im taken
01-03 : 13:15    113379 -> 88039: well wanna b sister with benefits
01-03 : 13:15    88039 -> 113379: no i dont
01-03 : 13:15    88039 -> 113379: bye

01-03 : 18:08    228515 -> 113379: because what
01-03 : 18:08    228515 -> 113379: oh my god you making me mad now ugh
01-05 : 15:05    113379 -> 228515: bc im looking for mexican latina or mix with mexican

01-03 : 20:19    152525 -> 113379: ?lol
01-05 : 15:04    113379 -> 152525: do u wanna date irl

01-04 : 03:49    45375 -> 113379: yes why
01-05 : 15:03    113379 -> 45375: bc i am

01-05 : 15:08    113379 -> 244943: hey girl what race r u
01-05 : 15:08    244943 -> 113379: im black
01-05 : 15:10    113379 -> 244943: r u mix
01-05 : 15:10    244943 -> 113379: im actually brownskin in rl
01-05 : 15:10    113379 -> 244943: wanna b sister with benefits
01-05 : 15:11    244943 -> 113379: sure
01-05 : 15:11    244943 -> 113379: wait what do you mean
01-05 : 15:12    113379 -> 244943: look it up
```

(b) A selection of conversations with the user

Figure 5.9: Chats with node 113379

Gaussian Mixture Models

Table 5.6 shows the clusterings from the GMM algorithm with 10, 7 and 5 clusters on the data set from month 4. Tables C.6, C.7, C.8, C.9, and C.10 in the Appendix C shows more extensive tables for all of the months.

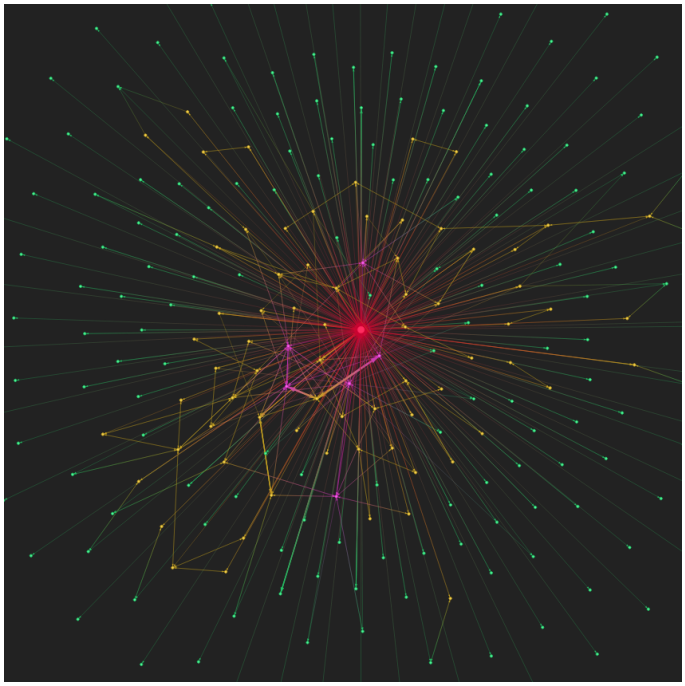


Figure 5.10: Ego graph for node 6255

```
03-06 : 01:59    6255 -> 53629: hey are u hornxxXy ?
03-06 : 18:08    53629 -> 6255: maybe
03-07 : 16:16    6255 -> 53629: how old are u
03-08 : 16:29    53629 -> 6255: 12
03-08 : 19:31    6255 -> 53629: wanna have funb

03-06 : 02:00    6255 -> 11702: hey are u hornxxXy ?
03-06 : 06:50    11702 -> 6255: y u wanna know

03-06 : 02:01    6255 -> 54666: hey are u hornxxXy ?
03-06 : 02:05    54666 -> 6255: why u asking ?
03-06 : 02:05    6255 -> 54666: if u are we could have fun
03-06 : 02:07    54666 -> 6255: have fun how ?? before I answer anything else
03-06 : 02:07    6255 -> 54666: trade pic
03-06 : 02:08    54666 -> 6255: I dont do that
03-06 : 02:08    6255 -> 54666: why not ?
03-06 : 02:09    54666 -> 6255: cause I dont . i prefer to be safe not sorry
03-06 : 02:11    6255 -> 54666: ok how old are u
03-06 : 02:11    54666 -> 6255: 15
03-06 : 02:11    6255 -> 54666: 17
```

Figure 5.11: Chats with node 6255

Table 5.6: GMM on data from the 4th month

Cluster	Number of nodes	Average degree	Average weighted degree	Average weighted out-degree	Average CC
Clustering 1: 10 Clusters					
0	21584	1.64	4.18	2.03	0.0
1	1640	80.72	2409.53	1202.3	0.03
2	2482	3.16	53.51	27.17	0.55
3	15347	4.22	28.15	13.85	0.04
4	2207	11.7	887.67	445.13	0.11
5	3	880.33	2761.33	1465.33	0.01
6	4520	14.35	167.77	84.7	0.05
7	4137	3.4	115.91	58.2	0.0
8	13119	1.13	1.17	0.0	0.0
9	3443	34.6	636.51	320.48	0.04
Clustering 2: 7 Clusters					
0	17290	1.74	4.88	2.21	0.0
1	7276	19.55	316.02	159.89	0.05
2	15952	4.39	28.82	14.19	0.04
3	17424	1.16	1.22	0.32	0.0
4	3111	4.14	231.23	114.96	0.45
5	4188	3.59	111.75	56.51	0.0
6	3241	59.07	1840.96	920.01	0.04
Clustering 3: 5 Clusters					
0	26924	2.13	11.89	5.78	0.0
1	7059	3.03	86.9	43.86	0.25
2	5510	44.79	1365.26	683.46	0.05
3	18468	1.22	1.38	0.35	0.0
4	10521	12.7	146.11	73.35	0.05
The entire data set					
	68482	7.04	146.3	73.15	0.04

Cluster 5 from clustering 1 stands out. The cluster has only three nodes, and the average degree is significantly higher than the other clusters. This particular cluster

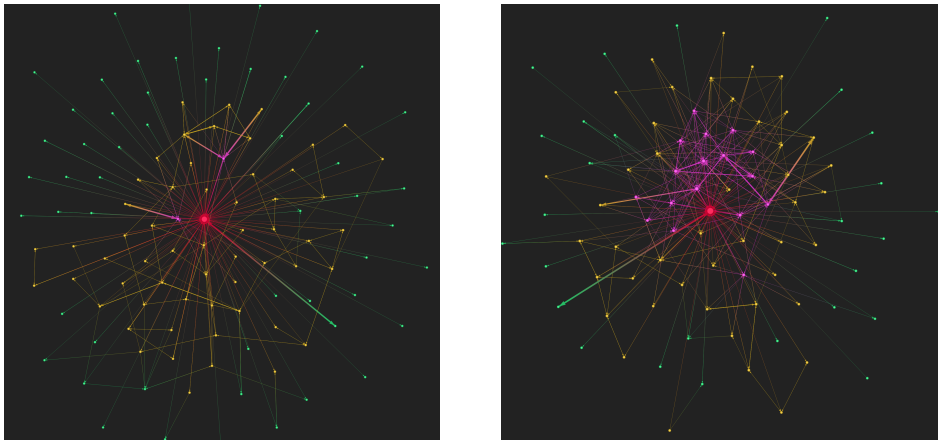
is found in other algorithms as well and will be reviewed later in this section.

The other clusters from month 4 are large, and no other cluster stands out. Therefore, the most interesting clusters from the first clustering were collected and further clustered to get a more helpful clustering. Clusters 0, 2, 3, 7, and 8 were not included for the new clustering since they all contained nodes with few neighbors, which is normal behavior. The nodes in the rest of the clusters were clustered again, aiming to get more useful clusters. Table 5.7 shows the key data from the new clustering with fewer nodes. Table C.11 shows an extended version of the table including more key values from the clustering.

Table 5.7: GMM on parts of data from the 4th month

Cluster	**Number of nodes**	**Average degree**	**Average weighted degree**	**Average weighted out-degree**	**Average CC**
10 Clusters					
0	2591	15.73	122.78	59.91	0.05
1	1420	44.82	637.04	319.04	0.04
2	909	7.02	439.51	221.32	0.16
3	482	117.96	1690.88	851.65	0.03
4	1126	67.72	2769.72	1382.51	0.04
5	544	12.7	1802.56	889.63	0.05
6	1885	11.61	250.89	129.15	0.07
7	1330	24.79	303.09	153.76	0.03
8	3	880.33	2761.33	1465.33	0.01
9	1523	23.98	950.72	480.54	0.06
The entire data set					
	11810	28.97	750.26	375.99	0.06

This clustering gave some smaller clusters; however, the algorithm still produces clusters of large sizes. Cluster 8 contains the same nodes given by the original run of the GMM algorithm. The next smallest cluster is cluster 3. This cluster seems to have relatively large average degree and weighted degree values. Five nodes from the cluster were randomly chosen and visualized. Figure 5.12 shows examples of two nodes, 40881 and 249130, from cluster 3. None of the users investigated showed predatory behavior in the text messages.



(a) Ego graph for node 40881 **(b)** Ego graph for node 249130

Figure 5.12: Two nodes from cluster 3 using GMM

None of the chosen nodes from cluster 3 showed predatory behavior in chats that were investigated. However, comparisons of this clustering with the k-means clustering show that 67% of the nodes in cluster 7 were in cluster 3 for the new GMM cluster. This shows that the clustering in two steps might be able to find similar users that the k-means algorithm found.

The other clusterings from other months gave similar results as in month 4, i.e., many large clusters where no cluster stands out. Therefore, no other months were analyzed in depth further.

BIRCH

Table 5.8 shows the clusterings from the BIRCH algorithm with 10, 7 and 5 clusters on the data set from month 4. Tables C.12, C.13, C.14, C.15, and C.16 in the Appendix C shows the extended tables for all months.

Table 5.8: BIRCH on data from the 4th month

Cluster	Number of nodes	Average degree	Average weighted degree	Average weighted out-degree	Average CC
Clustering 1: 10 Clusters					
0	63079	3.57	45.07	22.33	0.04

1	312	112.63	1303.2	653.83	0.03
2	39	172.64	9732.36	4847.46	0.03
3	46	219.26	2626.78	1308.67	0.02
4	2772	29.83	1263.99	638.58	0.05
5	3	880.33	2761.33	1465.33	0.01
6	9	2.11	829.22	412.56	0.96
7	279	91.8	3227.96	1632.78	0.04
8	110	78.13	6697.15	3287.8	0.03
9	1833	46.39	606.51	301.18	0.04
Clustering 2: 7 Clusters					
0	358	126.33	1473.27	737.97	0.03
1	389	87.94	4208.96	2100.78	0.04
2	39	172.64	9732.36	4847.46	0.03
3	63079	3.57	45.07	22.33	0.04
4	4605	36.42	1002.28	504.28	0.04
5	3	880.33	2761.33	1465.33	0.01
6	9	2.11	829.22	412.56	0.96
Clustering 3: 5 Clusters					
0	428	95.65	4712.26	2351.06	0.04
1	63088	3.57	45.19	22.39	0.04
2	3	880.33	2761.33	1465.33	0.01
3	358	126.33	1473.27	737.97	0.03
4	4605	36.42	1002.28	504.28	0.04
The entire data set					
	68482	7.04	146.3	73.15	0.04

From Table 5.8, cluster 5 stands out, having only three nodes. This cluster is also found in other clusterings from other algorithms and will be reviewed later in this section.

Cluster 6 is also small. However, the nodes of the clusters have few neighbors, which makes the nodes less suspicious. The average CC is significantly higher than the average, with 0.96 compared to 0.04 on average. However, when a node has such few neighbors, it is easier to achieve a higher CC value, so this does not make the cluster suspicious. This cluster was hence not studied further.

Clusters 1 and 2 are, on the other hand, more interesting. These have a high average degree, and both are relatively small clusters. From these clusters, seven nodes in total were investigated further. The nodes investigated from cluster 1 behaved normally and are considered normal users. None of the nodes from cluster 2 were deemed predators, but one of the users was having inappropriate conversations. The users wanted sexualized pictures from underage girls, and there was an exchange of personal information, which is against the game rules. However, the user's age was claimed to be 17, and the user mostly looked for girls of similar ages and was hence not deemed a predator. The claimed age of the user is not certain, but we found no other indication in the chats that the age was not correct. On the other hand, sexting in MovieStarPlanet is considered against the rules, so the user showed not-allowed behavior regardless of whether the user was a predator. Node 63953 is the related node to the user, and it is visualized in Figure 5.13. The node has no activity in the first three months and is most active in month 4. There is also some activity from month 5. The node has 151 neighbors, and there are 6451 messages sent to and from the user.

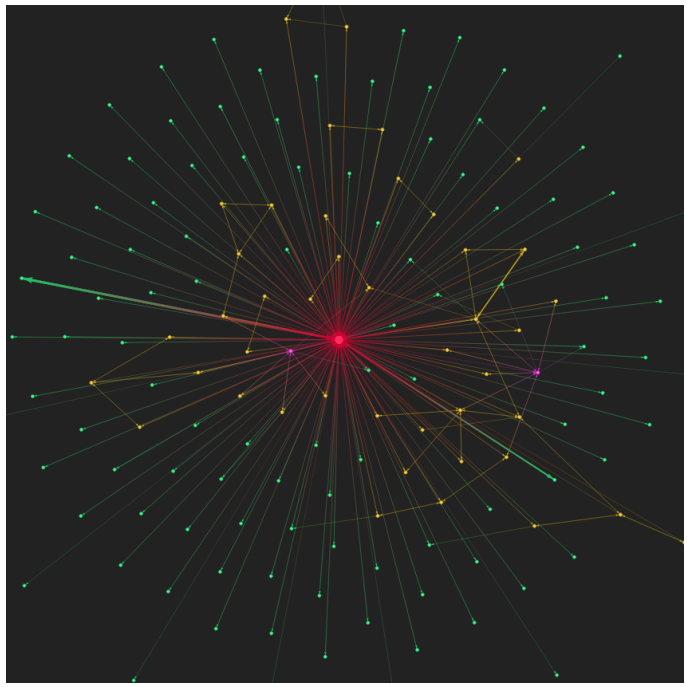


Figure 5.13: Ego graph for node 63953

BIRCH was also run with the other months. From month 1, there was one cluster containing one node. This node had many conversations, and an investigation of the node showed that the user was a spammer, but no illegal activity was found.

In addition, clusters 1 and 2 stood out. No illegal predatory activity was disclosed from nodes from cluster 1. However, in cluster 2, one node, node 261700, wanted sexualized pictures from girls. The user claimed to be a male, and he claims to have different ages with different users, so there is a significant probability that the user is an adult. In addition to the user's age being challenging to determine, it was also challenging to determine the age of the users he was communicating with, as he often did not ask for their age. This may indicate that the user is interested in girls of all ages, also younger ones, since it is a game aimed at children from 8 to 15 years old. Figure 5.14 shows the visualization of the ego graph for the node.

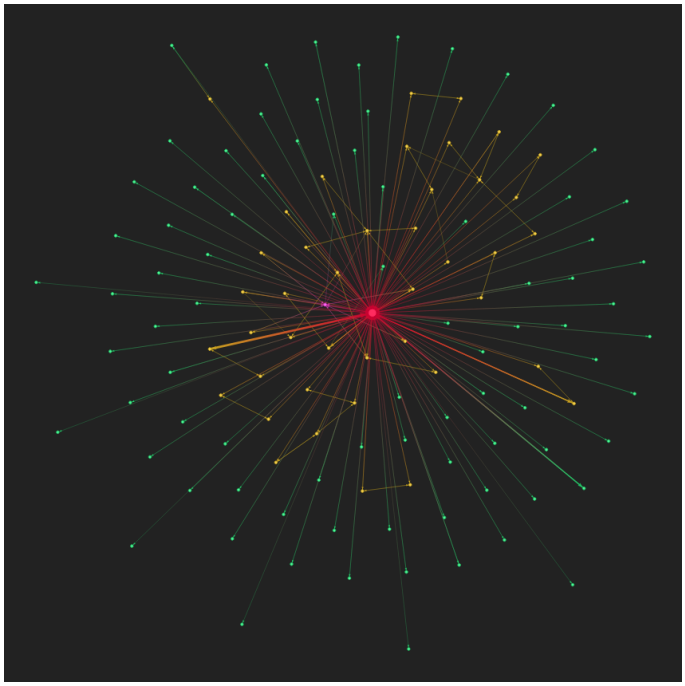


Figure 5.14: Ego graph for node 261700

In month 2, node 201279, from cluster 9 was discovered having inappropriate conversations. The user's age was not determined, and the user claimed to be of different ages towards different users. The user wanted to participate in sexual activity with girls of seemingly all ages. It was determined that the user could be a predator, but the user might also be a hormonal teenager. Figure 5.15 shows the visualization of the ego graph for the node, and Figure 5.16 shows some conversations the user has started. From the figure, we see that the user invites to sexual activities and that it wishess to switch media to Snapchat (here abbreviated to "sn"). This user was primarily active in month 2 and had little activity in the other four months. The node's properties changed substantially from month 2 to the other months,

and it was therefore clustered in larger clusters for the other months. 48% of the conversations with this users contained only one message, and 43% under 8 messages. The user seems to send the same message to a lot of users hoping some will reply.

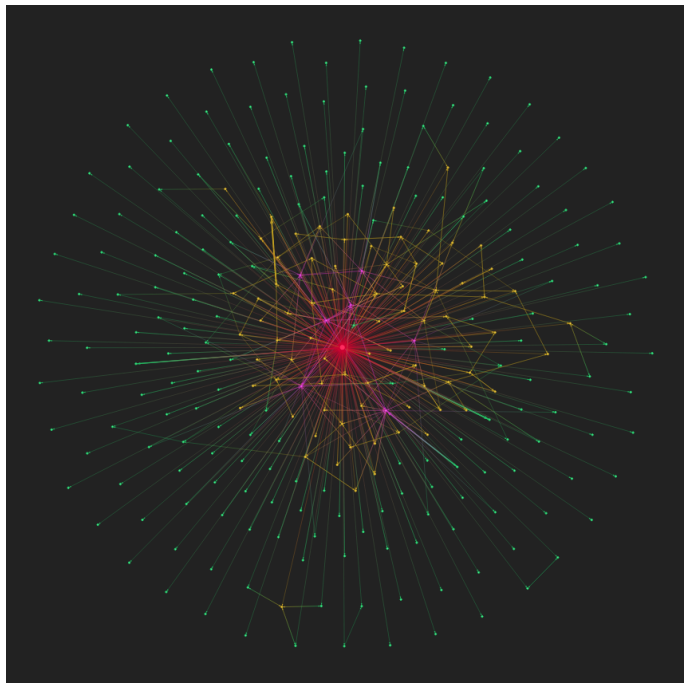


Figure 5.15: Ego graph for node 201279

```
01-06 : 02:04    201279 -> 732: heyy wanna mxaxstxurbate with me?

01-06 : 02:04    201279 -> 265606: heyy wanna mxaxstxurbate with me?
01-06 : 02:06    265606 -> 201279: yes
01-06 : 02:06    201279 -> 265606: whats ur sn

01-06 : 02:04    201279 -> 208623: heyy wanna mxaxstxurbate with me?

01-06 : 02:04    201279 -> 91887: heyy wanna mxaxstxurbate with me?

01-06 : 02:04    201279 -> 232001: heyy wanna mxaxstxurbate with me?

01-06 : 02:05    201279 -> 161053: heyy wanna mxaxstxurbate with me?
01-06 : 02:06    161053 -> 201279: yes
01-06 : 02:06    201279 -> 161053: whats ur sn
01-06 : 02:06    161053 -> 201279: I dont have one
01-06 : 02:06    161053 -> 201279: just talk dirty

01-06 : 02:08    201279 -> 252041: heyy wanna mxaxstxurbate with me?
01-06 : 02:09    252041 -> 201279: no. youre one sick mf

01-06 : 02:08    201279 -> 217614: heyy wanna mxaxstxurbate with me?
01-06 : 02:08    217614 -> 201279: yes sir teach me the ways u like best
01-06 : 02:09    201279 -> 217614: whats ur sn
```

Figure 5.16: Chats with node 201279

The clustering from month 3 also provided a node with illegal activity in the chats. The user represented by node 62508 wanted girls to join zoom calls, most likely for sexual reasons. The user sent out a short message containing just "hi" and approached

the users that replied. Figure 5.17 shows the visualization and Figure 5.18 shows some one of the chats with the user. As the chats also confirm, the user sends out many messages and get relatively few replies. Over half of the conversations that the user has, consists of only one message. The user has a behavioral pattern close to how we expect predators to behave, namely sending out many similar messages and trying to groom the users that reply. We considered thus this user a predator.

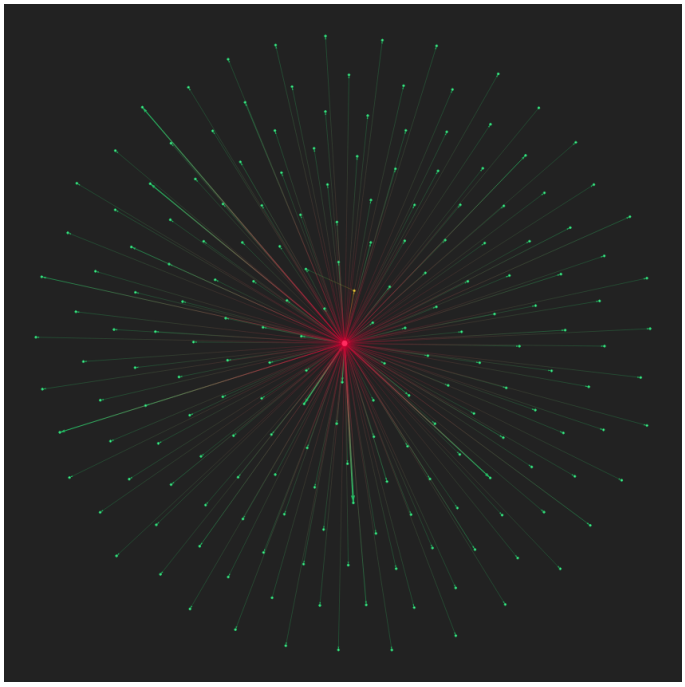


Figure 5.17: Ego graph for node 62508

```
03-04 : 01:02    62508 -> 215366: hi

03-04 : 01:02    62508 -> 218826: hi

03-04 : 01:02    62508 -> 109978: hi

03-04 : 08:23    62508 -> 34219: wyd
03-04 : 08:24    62508 -> 34219: bored wanna xzoom?
03-04 : 08:56    62508 -> 34219: when
03-04 : 08:57    34219 -> 62508: when yo want to do ittt
03-04 : 08:57    62508 -> 34219: make one
03-04 : 08:57    34219 -> 62508: no you
03-04 : 08:58    34219 -> 62508: give the code and stuff
03-04 : 08:58    34219 -> 62508: give me the c o d e
03-04 : 08:58    34219 -> 62508: c o d e
03-04 : 08:59    62508 -> 34219: [redacted]
03-04 : 08:59    62508 -> 34219: [redacted]
03-04 : 08:59    62508 -> 34219: and [redacted]
03-04 : 08:59    62508 -> 34219: psd is [redacted]
03-04 : 09:13    62508 -> 34219: ...
```

Figure 5.18: Chats with node 62508

Month 5 gave interesting results for the BIRCH algorithm. From cluster 4, three nodes disclosed illegal chats, namely nodes 113379, 32357, and 255808. Node 113379

is found from other months and other algorithms and is considered a predator. The two other nodes also included inappropriate conversation, but the age was difficult to determine for both. Investigation of node 32357 shows a user that directly asks for sexualized pictures from other players. The user states to be of different ages to different players, so the age is challenging to determine. The node might be a predator, but no evidence suggests that the user is significantly older than its peers.

Investigation of node 255808 showed that the user expresses interest in sexual activities with 5-year-old children or even younger. However, no users in the conversation with this user were this young; hence no proof that the user would engage in intimate conversations with children this young. Nevertheless, this user may be a predator. Figure 5.19 shows the visualization of the node from the fifth month. The node is active from the fourth month but is most active in the midst of month 5. The node has 407 different conversations where the user sent at least one message in 401. In comparison, the user received messages in only 227 of the conversations. Another interesting feature for the node is the percent of conversations containing one message, which is 43%. The high percentage may indicate that the user send out many messages to try to get answers from a few that will fulfill its wishes.

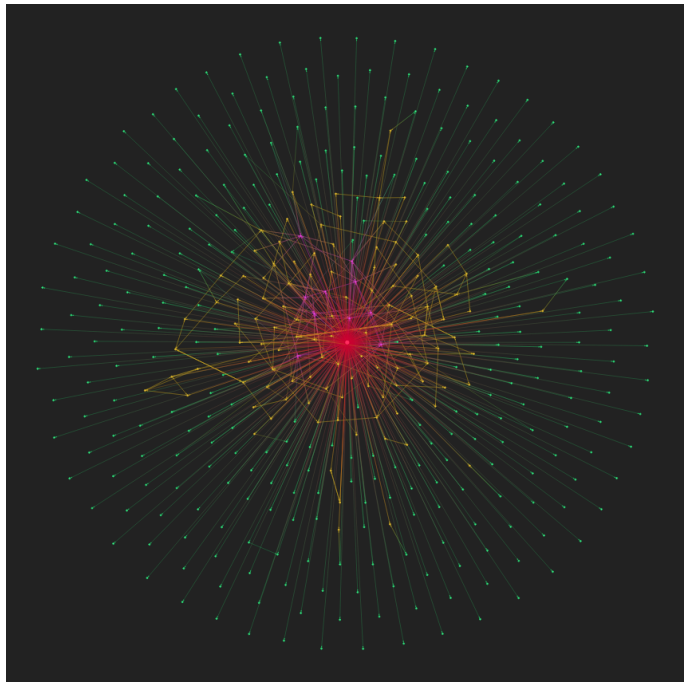


Figure 5.19: Ego graph for node 255808

Comparing the different clusterings, the algorithm uses hierarchical clustering,

which produces the same result every time it is run. The clusters from clustering 1 are subsets of the clusters from clustering 2 and 3. From month 4, the nodes from cluster 1 from clustering 1 were clustered in cluster 0 for clustering 3 and 2 in cluster 3.

Mean Shift

The mean shift algorithm calculates the number of clusters without it being given as input. Therefore, there are some different numbers of clusters for the different clusterings produced by the mean shift algorithm. For month 4, 22 clusters were created as shown in Table 5.9. More extencive tables for this and the other months, are found in Tables C.17, C.18, C.19, C.20, and C.21.

Table 5.9: Mean shift on data from the 4th month

Cluster	**Number of nodes**	**Average degree**	**Average weighted degree**	**Average weighted out-degree**	**Average CC**
22 Clusters					
0	68032	6.21	120.37	60.27	0.04
1	12	2.17	633.75	309.92	0.92
2	35	159.31	8055.43	3980.17	0.03
3	5	283.6	2950.2	1584.8	0.01
4	29	180.07	4157.03	2062.28	0.03
5	69	152.65	2052.84	1026.68	0.03
6	1	1121.0	4223.0	1921.0	0.01
7	1	806.0	2008.0	1133.0	0.0
8	1	714.0	2053.0	1342.0	0.01
9	1	524.0	3372.0	1768.0	0.0
10	1	399.0	7414.0	3665.0	0.04
11	1	397.0	10111.0	5962.0	0.04
12	1	331.0	3141.0	1912.0	0.01
13	1	308.0	913.0	150.0	0.01
14	4	259.5	1915.25	1015.0	0.01
15	1	294.0	9730.0	5289.0	0.02
16	6	202.0	1583.83	809.5	0.01
17	3	176.67	2406.67	1280.33	0.01

18	193	118.18	2709.59	1358.18	0.03
19	3	132.67	18677.67	9230.33	0.02
20	81	67.02	7602.04	3721.84	0.04
21	1	2.0	1074.0	603.0	1.0
The entire data set					
	68482	7.04	146.3	73.15	0.04

There are several clusters containing one node, 21 of the clusters consist of less than 200 nodes, and 20 of them consist of less than 100. This means that only one cluster is large compared with other clusterings from other algorithms. Clusters 6, 7, and 8 consist of the three nodes clustered alone in k-means, GMM and BIRCH. These three nodes will be studied later in this section.

From the small clusters, four nodes from three clusters disclosed interesting users. Firstly, cluster 14 contained four nodes, and one of them was node 113379, which was previously identified as a predator. One other node, node 232499, from the cluster was interesting. The text messages written by the user indicated that he wanted sexualized pictures from other users down to 13 years old. The user might be a predator, but its age is challenging to determine as it changes between conversations. The user had conversations with 298 different users, and most of the activity from the user was in months three and four. Over half of the conversations with the user contains less than eight messages. Figure 5.20 shows the visualization of the node and Figure 5.21 shows chats with the user. With "pictures", we interpret that the user refers to sexualized pictures. The chats disclose that the user wants pictures of girls as young as 13.

Two other interesting nodes are two nodes from clusters 3 and 16. Both of them have similar behavior to node 232499, where the user's main goal seems to be to get girls to send pictures of themselves. However, the age of the users is challenging to determine, and it is hence difficult to determine if the users are predators or not. The two nodes are visualized in Figure 5.22.

The other small clusters had normal and often highly active users or spam users aiming to achieve progress in the game.

DBSCAN

Table 5.10 shows key data from the DBSCAN clustering on month 4. The clustering for this month only gave one cluster, cluster 0, and outliers. The outliers are the

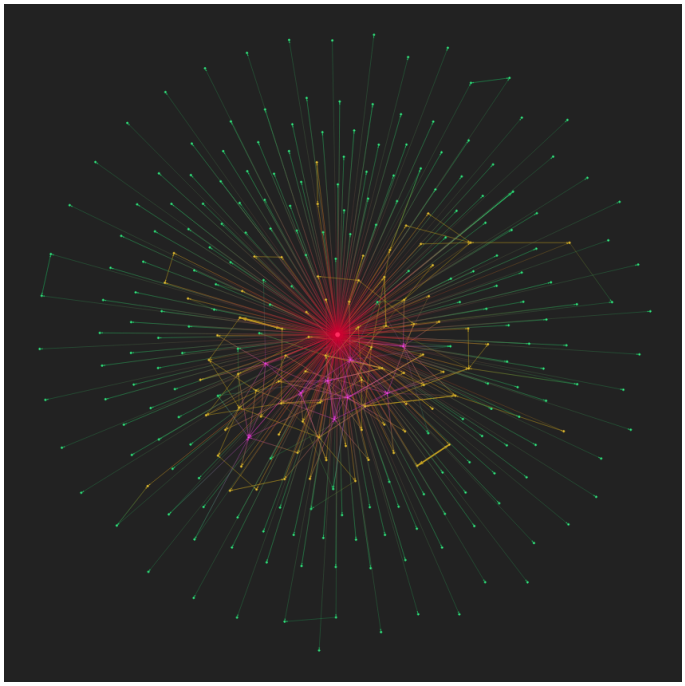


Figure 5.20: Ego graph for node 232499

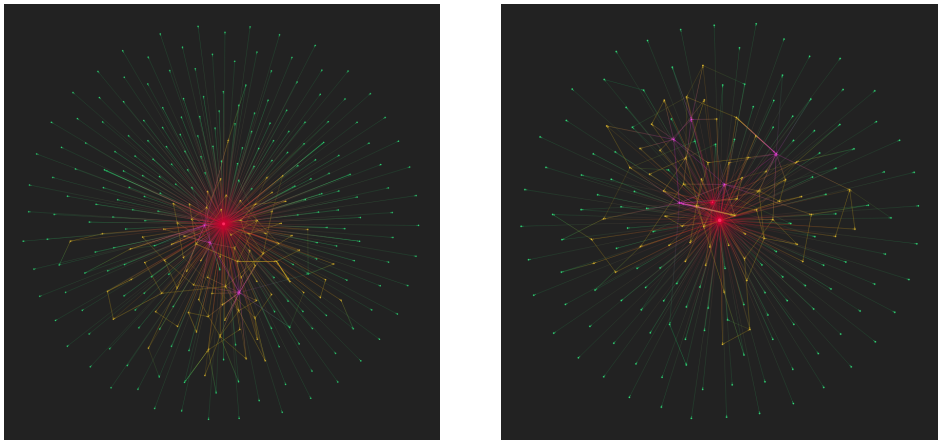
```
03-07 : 00:12    232499 -> 79597: how old?
03-07 : 00:14    79597 -> 232499: why ?

03-07 : 00:13    232499 -> 9720: heyy
03-07 : 00:24    9720 -> 232499: hiii
03-07 : 02:15    232499 -> 9720: how old?
03-07 : 02:15    9720 -> 232499: Im 13
03-07 : 02:15    232499 -> 9720: do u send pictures?
03-07 : 02:18    9720 -> 232499: nah

03-07 : 00:13    232499 -> 138086: how old?
03-07 : 00:13    138086 -> 232499: 18
03-07 : 00:13    232499 -> 138086: nice
03-07 : 00:14    232499 -> 138086: do u send pictures?
03-07 : 00:14    138086 -> 232499: no
```

Figure 5.21: Chats with node node 232499

nodes that were given cluster -1. The algorithm were run on the other months as well. The result from this can be found in Tables C.22, C.23, C.24, C.25, and C.26.



(a) Ego graph for node 210526 from cluster 3 **(b)** Ego graph for node 5444 from cluster 16

Figure 5.22: Nodes from cluster 3 and 16 using mean shift

Table 5.10: DBSCAN on data from the 4th month

Cluster	Number of nodes	Average degree	Average weighted degree	Average weighted out-degree	Average CC
Clustering 1: 1 Cluster plus outliers					
-1	1131	87.0	2879.1	1435.84	0.11
0	67351	5.69	100.41	50.27	0.04
The entire data set					
	68482	7.04	146.3	73.15	0.04

For this clustering, the outliers are the most interesting to examine. The number of outliers is high, and it is too time-consuming to investigate all nodes manually. To further determine if the clustering algorithm is possible to use for predator detection, it will be compared with the k-means and BIRCH clustering algorithms.

Firstly, the algorithm was compared with the k-means algorithm for month 4 from Table 5.5. All nodes from clusters 1 and 9 were in the outlier group. 93 % of nodes from cluster 7 were also considered outliers with the DBSCAN algorithm. The rest of the clusters from the k-means algorithms were primarily clustered in cluster 0. This shows that the clusters that were small and interesting from the k-means algorithm to a large extent were labeled outliers when using DBSCAN

Comparing the clusters from the BIRCH algorithm (Table 5.8) also gave results indicating that the DBSCAN can give some relevant information. From the BIRCH algorithm, the nodes from cluster 1, 2, 3, 5, 6, 7, 8 were mostly considered outlier by the DBSCAN algorithm. These clusters were the ones considered interesting when investigating the BIRCH algorithm.

The comparison of DBSCAN with k-means and BIRCH shows that the algorithm could be helpful in predator detection as many predators will be considered outliers of the graph. However, more adjustments to the algorithm or other feature sets might produce clusters that give more helpful information. Another solution could be to combine the algorithm with another one by using, for example, k-means on the outliers to sort them into helpful clusters further.

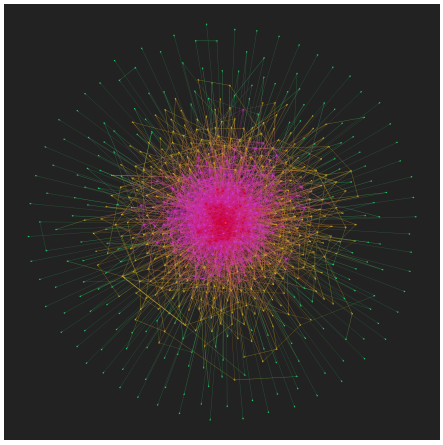
Investigation and comparison of the clusterings

One cluster containing three nodes stood out from multiple clusterings from the fourth-month data set. The cluster was present in k-means with 10 and 7 clusters, GMM with 10 clusters, and in BIRCH in all three clusterings. Mean shift clustered the three nodes in three clusters with no other nodes. The three nodes in the clusters were nodes 21889, 52855, and 252962. The average degree value for the cluster is 880.33, more significant than the average for the data set. The average number of messages is 2761.33, which is not too high compared with other clusters. On average, the conversations consist of only 3 messages. Figure 5.23 shows the visualizations of the three nodes.

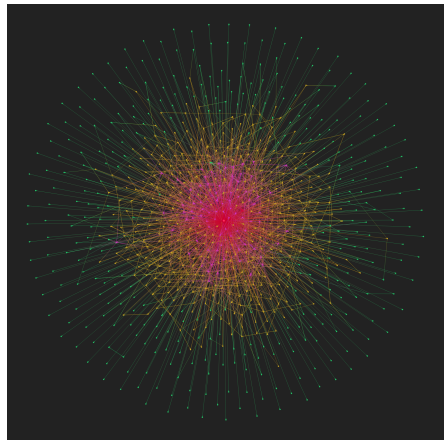
We observed that the three nodes have similar patterns by looking at the visualizations. The nodes have many neighbors, and many of the neighbors are also connected. They were also investigated in the other months to see if they had similar activity in all months. Nodes 21889 and 252962 did not have much activity other than month 4. On the other hand, 52855 had much activity all months. The k-means algorithm with 10 clusters clustered the node in relatively small clusters in the first and fifth months. The node were in cluster 5 with 273 nodes from month 1 and cluster 7 with 219 nodes from month 5. From the first month, cluster 5 is the smallest cluster where the average degree is the highest. The same was the case for month 5.

The chats of the users related to the three nodes were investigated. All three users turned out to be spammers. The users sent out the same message to as many users as possible, asking for favors that would help them progress in the game. This activity is not considered against the rules by MovieStarPlanet.

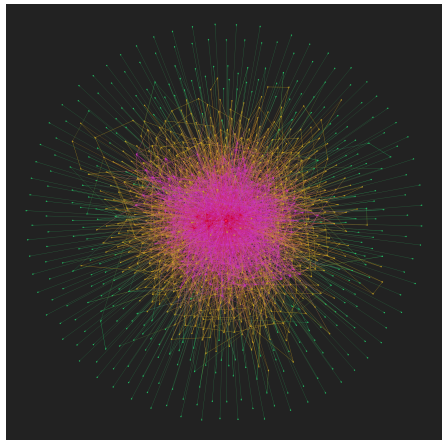
To further compare the different clusters, we looked at three predator nodes and where they were clustered in the different algorithms and months.



(a) Ego graph for node 21889



(b) Ego graph for node 252962



(c) Ego graph for node 52855

Figure 5.23: Ego graph for nodes 21889, 252962, and 52855

The user represented by node 113379 was concluded to be a predator. The node was initially found by using k-means in cluster 5 from month 1, but also with BIRCH from month 5 and month 4 from mean shift. Table 5.11 shows where the node was clustered all month for all clustering algorithms. The BIRCH algorithm clustered the node in small clusters for all months. Especially, month 5 gave small clusters. k-means and mean shift clustered the node in relatively small clusters for all months except month 2. Mean shift clusters the node in clusters containing only four and seven nodes for months 4 and 5, respectively. Finally, DBSCAN clustered the node as an outlier in all months, and GMM clustered the node in a relatively small cluster for the new run of the algorithm in month 4. This shows that multiple algorithms

consider this predator node as abnormal and could potentially be used to detect the predator.

Table 5.11: Node 113379 in different months and clustering algorithms

Algorithm	Month	Cluster (size)
k-means	1	5 (273)
	2	9 (2105)
	3	7 (137)
	4	7 (459)
	5	7 (219)
GMM	1	6 (1528)
	2	4 (2110)
	3	4 (1666)
	4	3 (482)
	5	5 (2177)
BIRCH	1	6 (131)
	2	4 (167)
	3	4 (32)
	4	3 (46)
	5	4 (27)
Mean shift	1	1 (111)
	2	0 (83795)
	3	1 (36)
	4	14 (4)
	5	4 (7)
DBSCAN	1	-1 (2078)
	2	-1 (1398)
	3	-1 (1204)
	4	-1 (1131)
	5	-1 (1329)

Node 62508 is also considered a predator node. The node had either 1 or 0 messages sent or received in months 1, 2, and 4; hence these months were not interesting to analyze. The user was more active in months 3 and 5, and we studied the node in clusters for different algorithms for these two months. Table 5.12 shows the overview of the node in the cluster. All algorithms except GMM cluster the predator node in small clusters in the third month. Only BIRCH and DBSCAN seem to view the node as an outlier for the fifth month. The node is less active in month 5 compared with month 3, which may cause the node to be clustered in larger groups

in the last month.

Table 5.12: Node 62508 in different months and clustering algorithms

Algorithm	Month	Cluster (size)
k-means	3	7 (137)
	5	8 (2030)
GMM	3	4 (1666)
	5	5 (2177)
BIRCH	3	3 (67)
	5	5 (487)
Mean shift	3	6 (5)
	5	0 (84628)
DBSCAN	3	-1 (1204)
	5	-1 (1329)

6255 is the last node that was compared for different clusterings. The node had no activity before the third month and little activity in month 4. Thus, only clusterings from months 3 and 5 will be shown in Table 5.13. The table shows that the node is clustered mainly in small clusters for the different algorithms.

Table 5.13: Node 6255 in different months and clustering algorithms

Algorithm	Month	Cluster (size)
k-means	3	0 (1436)
	5	7 (219)
GMM	3	4 (1666)
	5	5 (2177)
BIRCH	3	6 (2884)
	5	2 (167)
Mean shift	3	0 (80933)
	5	5 (10)
DBSCAN	3	-1 (1204)
	5	-1 (1329)

By looking at the three nodes in Tables 5.11, 5.12, and 5.13, we see that the nodes are generally clustered in small clusters in different algorithms. This shows that several different clustering algorithms can be used for predator detection.

Chapter 6
Discussion

This chapter will discuss the results related to the research questions. Firstly, the discussion around research questions will be presented, followed by some discussion on the limitations of the thesis.

6.1 Research questions

6.1.1 RQ a: Which graph features can be used to detect predators in an online chat network for children?

Table 4.1 shows all features that were used to do the clustering in the thesis. This set was the only one tested out during the study. The feature set used did contribute to detecting predators.

The different degree measures were a large part of the set and seemed to be an essential part of the clustering. The different clusters were often characterized by the number of neighbors and messages of the different users. However, including both in-degree, out-degree, weighted in-degree, and weighted out-degree may have been unnecessary. For instance, the weighted in-degree value is already represented by the weighted degree and weighted out-degree, and excluding it from the feature set could influence the clustering results.

One feature that could be interesting to include is the number of connections between a user's neighbors. The CC value gives the connection between neighbors related to the number of possible connections. However, when the number of neighbors gets large, the possible number of connections between neighbors grows aggressively. For example, if a node has 30 neighbors, there are 870 possible connections between the neighbors. Therefore, the feature will often be relatively low for nodes with many neighbors despite there might be many connections between the neighbors. Nodes with two or three neighbors will achieve larger CC values, although those nodes might be less interesting. If a feature only calculates the connection between

the neighbors, nodes with more neighbors can get larger values. Additionally, the weighted CC value seems to influence the clustering minimally as all the values from the features are neglectable.

The feature called "Out/Total", calculating $\frac{weighted\ out-degree}{weighted\ degree}$, seemed to contribute little to the clusterings. The value is mostly around 0.5 for most clusters. The clusters with deviating values are often too large to be interesting. However, nodes with a large portion of messages sent out are abnormal and interesting for further investigations. A similar feature that could give interesting results is $\frac{out-degree}{degree}$ instead of $\frac{weighted\ out-degree}{weighted\ degree}$. $\frac{out-degree}{degree}$ could detect users who have many conversations where the other user has not replied. This is typical for users who spam a lot, which some predators do.

Non-common neighbors seem to not influence the clustering too much as the standard deviation for that value is generally high, and the different clusters have a similar number of non-common neighbors values on average.

The distribution of messages, the last 12 features from Table 4.1, seems to influence the clustering to some degree. The different distribution values inside the different features vary significantly for different clusters.

Despite some features possibly being unnecessary for predator detection in online chat platforms for children, the feature set successfully detected some predators and other users committing illegal activities.

6.1.2 RQ b: Can unsupervised clustering algorithms be used to detect predators in an online chat network for children?

Mostly 5 clustering algorithms were tested for predator detection in the thesis; k-means, GMM, BIRCH, mean shift and DBSCAN. Agglomerative clustering was partially tested out for the smallest data set but was not included further due to its slow run time.

Using k-means, we discovered one predator and several other suspicious users that might be predators. The clustering algorithm gave clusters of different sizes, some small and interesting, to investigate further, which led us to a predator.

The GMM algorithm gave few interesting results. The clustering contained relatively equal-sized clusters, making them challenging to interpret. Some of the clusters were further clustered to make the results easier to analyze. This gave smaller clusters than the initial clustering, but no predators or users with other illegal or suspect behavior were found. On the other hand, when we searched for predator nodes in the GMM clustering, the nodes were found in smaller clusters.

The BIRCH clusterings lead to multiple predator users for multiple months. In addition, the clustering gave a wide spread of clusters of different lengths, making it more intuitive to choose clusters to investigate further.

Mean shift calculates the number of clusters without having it as a parameter, making the clusterings different for the various months. The clusterings from the individual months consisted of 22 to 29 clusters. This is many clusters compared with the other algorithms. However, it was not too challenging to analyze the clustering as many of the clusters only contained one or just a few nodes. This prevented the workload on assessing the different clusters from increasing extensively. Multiple suspect users were found in some of the small clusters, where one of them was a predator. One downside of the clustering algorithm was that it might exclude many interesting nodes by creating small clusters. Another disadvantage is that the disclosure of a predator will not lead to any other predators if it is clustered alone. If the predator is clustered in a cluster containing other users, the cluster could be further analyzed to find additional predators.

Clusterings from the DBSCAN stand out from clusterings from other algorithms as the algorithm created only one to three clusters plus groups of outliers. The groups of outliers from the various months were interesting, but they were also large groups. The outliers were not investigated extensively, but comparing the different clusters showed that the algorithm categorized the predator users and other suspicious users as outliers.

Generally, clustering algorithms can be used for predator detection. We used multiple clustering algorithms, and several of them did disclose predators. Especially BIRCH disclosed many different users that were predators or that had other illegal activities in the chats. Predators were found by examine small clusters and assess nodes within these clusters. The small clusters might detect normal users in addition to unwanted behavior. Nonetheless, if some unwanted behavior is detected, it can be useful for a platform such as MoviStarPlanet.

6.1.3 RQ c: How does the length of the time frame for the data set influence predator detection in an online chat network for children?

When testing the algorithms, we used a data set consisting of five shorter time frames studied individually and concatenated to a large set. We also had one data set from one day that was used to study the properties of the graph.

It was challenging to analyze the results when using the complete data set. Firstly, the clusters were larger when using a large data set, making the investigation of the

nodes more time-consuming. Simultaneously, BIRCH created several clusters that were small enough to investigate. The positive aspect of using the more extensive data set is that the set gives a more stable impression of the different users. If a user, for instance, has a very active or inactive day, this will not affect the overall impression of the user much. However, this might also affect the analysis negatively. For example, popular users will have many chats with many different users over time, and the chats may contain only lawful conversations. This is not a behavior that we want to detect with the clustering algorithms. However, the pattern of a popular user over time may be very similar to a user with much activity only in short periods, which is abnormal behavior. The results also show that some predator users were most active in shorter periods. The clustering on the more extensive data set may struggle to detect these kinds of abnormal behavior.

The one-day data set was also studied to some extent in the beginning. The downside of such small data sets is that they may give inaccurate graph patterns. It may be a bit arbitrary who is talking with each other on a particular day compared to a more extended period. For example, if a user is inactive for a day, it may cause a lot of received messages and none sent. This pattern can also be the opposite. However, a user with much activity will still be suspicious as sending out a significant number of messages in a short period is unusual. Simultaneously, one might miss multiple interesting users by only looking at data from one day.

The clusterings from the individual months gave the most valuable clusterings. The clusters were often not too large to look into, and the data sets were better at capturing abnormal behavior.

6.1.4 RQ: Can we detect predators in online chats for children by using a graph-theoretical approach?

Throughout the thesis, we have studied many different clustering results created by different clustering algorithms. From the study, some users have been analyzed through visualizations of the nodes' ego graphs and a thorough assessment of anonymized text messages sent to and from the relevant users. Some of these users are regular users, some have behaved like spammers, some have predatory behavior, and others performed other illegal activities. All the different users are categorized based on our interpretation of the text messages.

The users we found most likely to be predators had similar approaches when talking with other users. They would start the conversations differently, but eventually, they would try to convince the other party to join them on another social media platform. We believe that are mainly two reasons for this. Firstly, MovieStarPlanet censors many words that could be interpreted as sexual or offensive, so changing

the chat platform would make it possible to have conversations without words being censored. Secondly, other social media platforms facilitate sharing pictures and videos, which might be a goal for a predator or other users. The switching of media makes it challenging to determine the real intentions of a user. Hence, categorizing the different users was, to some extent, based on intuition and interpretation of the conversations before switching media.

Although we cannot be entirely sure if a user is a predator or not, we find several users that we believe to be predators. The users were found using the graph-theoretical approach described in Chapter 4. In addition to predators, we found other users that broke the game's rules. For example, many users were sexting with each other, and some users were giving out personal information, such as phone numbers and email addresses, which are rule breaks in MovieStarPlanet.

6.2 Limitations

As mentioned in Chapter 4, the main limitation of the thesis is the lack of ground truth. When beginning the research, the plan was to use a list of known predators as the base for predator detection. We planned to use known predators to train supervised learning algorithms and learn which graph patterns predators typically have. The ground truth could also contribute to the work on clustering algorithms. Instead of looking into interesting clusters, we could look for the known predators and see where they were clustered and if they were clustered together. This method could further be used to find similar nodes. However, without the set of known predators, this was not possible, and supervised algorithms were dropped from the study. Therefore, the predator detection was solely based on clustering and investigations heavily based on assumptions about predators and expected behavior.

The lack of ground truth also made it challenging to say anything about the performance of the algorithms. Initially, the plan was to classify the nodes as regular and abnormal users. This kind of classification is typically the result of supervised learning algorithms. By using classification algorithms, it would be interesting to measure the performance of the algorithms. However, this measurement was challenging when only using clustering algorithms. Few clusters contained only predator nodes, and none of the clusters were concluded only to contain normal users. The clustering in this research consisted of finding valuable clusters, finding interesting nodes within these clusters, and investigating whether there is any illegal or predatory behavior in the chats. It did not make sense to measure any performance on this, as we only looked into a small part of the nodes.

Another limitation to the thesis is the degree of assumptions and subjective perspective throughout the study. Especially when the clusters were investigated,

the nodes and clusters for further analysis were chosen based on a hypothesis about how predators might behave. Results from other research projects supported the hypothesis. However, we have likely missed out on some predators' features and properties because of this limitation. Also, due to the limited time we had on the research, we could not investigate all clusters from all clusterings, which could have disclosed properties of predators that we did not expect. In addition to basing the choice of interesting nodes and clusters on assumptions, we concluded if the users were predators based on our interpretation of the chats between the users. A part of the game is to form relationships with other players, and hence much of the conversations between the users revolved around romance, relations, and similar topics. Also, lying about age and other personal information is easy since it is a virtual world. Separating who are regular users and who are predators was hence sometimes challenging. The users that were said to be predators or normal users from this research might be wrongfully classified, as it is based on our interpretation of the conversations. However, the classification of the chats done in this thesis was in line with experts that have already seen many predatory conversations in the past. Other activities against the game rules were easier to recognize, such as sharing private information and sexual conversations.

Chapter 7
Conclusion and future work

This chapter concludes the thesis and presents some areas that should be studied further in future projects.

7.1 Conclusion

Throughout the thesis, we used graph theory and clustering algorithms aiming to detect predators. We did find several users likely to be predators, and hence the answer to the main research question is yes, it is possible to use graph theory to detect predators in online chat networks for children.

The feature set described in Table 4.1 was used in the clustering algorithms to detect predators. This set is hence usable for predator detection. However, the set was static and not adjusted throughout the testing of the clustering algorithms, and iterations with different feature sets could have given more valuable results. With useful results, we refer to clusterings with separate clusters containing, for example, primarily predators, regular users, and spammers. The clusterings in this thesis led us to predators, but few clusters seemed to contain only predators.

This thesis shows that it is possible to use clustering algorithms to find predators in a social network created for children. We used k-means, GMM, BIRCH, mean shift, and DBSCAN aiming to detect predators. k-means and BIRCH were the two algorithms that gave the most valuable results. These clusterings gave some clusters of small sizes which were possible to investigate. Both of the algorithms led to predators. The mean shift algorithm also clustered the data in valuable clusters, with many small clusters often containing only one node. Some of the small clusters did contain predator nodes. We did not find any predators using the GMM algorithm. This clustering gave relatively large clusters, making examining the results challenging. Nevertheless, when we further clustered a subset of the initial clusters, we got one smaller cluster that contained multiple of the predator nodes found in other algorithms. The DBSCAN gave results that were demanding to analyze due

to the few clusters generated. The algorithm gave a set of outliers that were too large to investigate manually, but it did contain all predator nodes found in other clusterings. Overall, clustering was used to divide the users of the data set based on the users' behavior, and the different clusters did lead us to users that are likely to be predators.

The time frame of the length of the data set did influence how well the clustering performed at disclosing predators. We found using data sets collected over approximately seven days most beneficial. A data set collected over a long period made separating the different behavior of more users challenging. The clusters got extensive, and it was challenging to analyze the results. The smallest data sets were challenging to analyze as there was too little data on the different users to conclude their behavior.

This thesis is the first study using this approach to detect predators, and there is still much research that should be done further to utilize this research on a social media platform.

7.2 Future work

Several different research topics could be interesting for future studies. First of all, supervised learning algorithms would be interesting to investigate. The plan was initially to include the supervised learning algorithms in the thesis, but we did not have any data to train the supervised learning algorithms due to the lack of ground truth. However, it would be interesting to study how supervised learning could detect predators. The performance of the algorithms would then be possible to calculate, and the method could be compared with state-of-the-art approaches.

In this thesis, only predator detection in static graphs has been studied. However, it could be interesting to study if anomaly detection in dynamic graphs could contribute to detecting predators. Dynamic graphs describe how components change over time and in different periods, unlike static graphs that utilize only one period. There is a chance that predators behave differently than normal users over time. For instance, a predator might send many messages at once, which might be unusual for non-predatory users. They might also use platforms in untypical time slots for a child, such as during school or at night. These features are impossible to detect in a static graph but may help detect predators.

When creating the feature sets, we only looked at features that can be calculated with only the knowledge of neighbors and neighbors' neighbors. Centrality features were not included as they were too computationally slow to calculate for the extensive network. However, they might give helpful information about the data, and it could

be interesting to test out the ML algorithms with centrality features as a part of the feature set.

Only one feature set was tested out in this study. Other features or subsets of the used feature set should be studied to get more valuable clusters. A more valuable way to calculate how tightly neighbors are connected could, for instance, replace or add to the CC features for a more relevant result. Also, using subsets of the feature set used in this study could be interesting. Studying different feature sets could indicate which features perform best at predator detection.

Exploring a broader spectrum of ML algorithms could also be interesting. As this thesis is the first study to use a graph-theoretical approach for predator detection, only well-known ML algorithms were tested. However, different algorithms and optimizations of the used algorithms could be investigated to get a more accurate result. Also, the parameters of the algorithms could be adjusted further to optimize the results. The algorithms could also be used together to find valuable clusters. For instance, the outliers of DBSCAN could be clustered with another algorithm to utilize it better. Multiple iterations of the clusterings with the same algorithm could also be tested further. This was studied to some degree when using GMM, and the results showed that using multiple iterations could sometimes be valuable.

Future work should also include research on determining a suitable timeframe for predator detection. A suitable timeframe would be a long enough period to have representative data of the users. However, the period should not be so long that predator detection is too late to prevent online grooming.

References

[AD15] P. H. Ahmad and S. Dang, «Performance evaluation of clustering algorithm using different datasets», *International Journal of Advance Research in Computer Science and Management Studies*, vol. 3, no. 1, pp. 167–173, 2015.

[AIP+18] A. Akabane, R. Immich, R. Pazzi, E. Madeira, and L. Villas, «Distributed egocentric betweenness measure as a vehicle selection mechanism in vanets: A performance evaluation study», *Sensors*, vol. 18, p. 2731, Aug. 2018.

[ASN+16] A. Almaatouq, E. Shmueli, M. Nouh, A. Alabdulkareem, V. K. Singh, M. Alsaleh, A. Alarifi, A. Alfaris, and A. Pentland, «If it looks like a spammer and behaves like a spammer, it must be a spammer: Analysis and detection of microblogging spam accounts», *International Journal of Information Security*, vol. 15, no. 5, pp. 475–491, 2016.

[BK19] P. Bours and H. Kulsrud, «Detection of cyber grooming in online conversation», in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2019, pp. 1–6.

[BKH11] K. M. Babchishin, R. Karl Hanson, and C. A. Hermann, «The characteristics of online sex offenders: A meta-analysis», *Sexual Abuse*, vol. 23, no. 1, pp. 92–123, 2011.

[Bou11] S. R. Boutwell, «Authorship attribution of short messages using multimodal features», Naval Postgraduate School Monterey CA, Tech. Rep., 2011.

[BvLR11] O. Bousquet, U. von Luxburg, and G. Rätsch, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*. Springer, 2011, vol. 3176.

[BW10] E. A. Bender and S. G. Williamson, *Lists, decisions and graphs*. S. Gill Williamson, 2010.

[Car15] M. A. Carreira-Perpinán, «A review of mean-shift algorithms for clustering», *arXiv preprint arXiv:1503.00687*, 2015.

[CBG06] S. Craven, S. Brown, and E. Gilchrist, «Sexual grooming of children: Review of literature and theoretical considerations», *Journal of sexual aggression*, vol. 12, no. 3, pp. 287–299, 2006.

[CFA14] A. E. Cano, M. Fernandez, and H. Alani, «Detecting child grooming behaviour patterns on social media», in *International conference on social informatics*, Springer, 2014, pp. 412–427.

[CJG+15] Y.-G. Cheong, A. K. Jensen, E. R. Guðnadóttir, B.-C. Bae, and J. Togelius, «Detecting predatory behavior in game chats», *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 7, no. 3, pp. 220–232, 2015.

[CM02] D. Comaniciu and P. Meer, «Mean shift: A robust approach toward feature space analysis», *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[CR17] C. Cardei and T. Rebedea, «Detecting sexual predators in chats using behavioral features and imbalanced learning», *Natural Language Engineering*, vol. 23, no. 4, pp. 589–616, 2017.

[DW10] D. DeBarr and H. Wechsler, «Using social network analysis for spam detection», in *Advances in Social Computing*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 62–69.

[EKSX96] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, «A density-based algorithm for discovering clusters in large spatial databases with noise.», in *kdd*, vol. 96, 1996, pp. 226–231.

[FB20] M. A. Fauzi and P. Bours, «Ensemble method for sexual predators identification in online chats», in *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, 2020, pp. 1–6.

[FKE12] M. Fire, G. Katz, and Y. Elovici, «Strangers intrusion detection-detecting spammers and fake profiles in social networks based on topology anomalies», *Human Journal*, vol. 1, no. 1, pp. 26–39, 2012.

[GACS16] F. E. Gunawan, L. Ashianti, S. Candra, and B. Soewito, «Detecting online child grooming conversation», in *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, IEEE, 2016, pp. 1–6.

[HW79] J. A. Hartigan and M. A. Wong, «Algorithm as 136: A k-means clustering algorithm», *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[IC12] G. Inches and F. Crestani, «Overview of the international sexual predator identification competition at pan-2012.», in *CLEF (Online working notes/labs/workshop)*, vol. 30, 2012.

[Joh16] J. W. Johnsen, «Algorithms and methods for organised cybercrime analysis», M.S. thesis, Norwegian University of Science and Technology, 2016.

[Kon09] A. Kontostathis, «Chatcoder: Toward the tracking and categorization of internet predators», in *PROC. TEXT MINING WORKSHOP 2009 HELD IN CONJUNCTION WITH THE NINTH SIAM INTERNATIONAL CONFERENCE ON DATA MINING (SDM 2009). SPARKS, NV. MAY 2009.*, 2009.

[KZP+07] S. B. Kotsiantis, I. Zaharakis, P. Pintelas, *et al.*, «Supervised machine learning: A review of classification techniques», *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.

[Mal07] L. A. Malesky, «Predatory online behavior: Modus operandi of convicted sex offenders in identifying potential victims and contacting minors over the internet», *Journal of child sexual abuse*, vol. 16, pp. 23–32, Feb. 2007.

[Mat21] R. Matteini Palmerini, «Graph theoretical approach to sexual predator detection», M.S. thesis, Norwegian University of Science and Technology, 2021.

[MDRR19] K. Misra, H. Devarapalli, T. R. Ringenberg, and J. T. Rayz, «Authorship analysis of online predatory conversations using character level convolution neural networks», in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, pp. 623–628.

[Mor13] C. Morris, «Identifying online sexual predators by svm classification with lexical and behavioral features», *Master of Science Thesis, University Of Toronto, Canada*, 2013.

[MP13] G. K. Mikros and K. Perifanos, «Authorship attribution in greek tweets using author's multilevel n-gram profiles», in *AAAI Spring Symposium: Analyzing Microtext*, 2013, pp. 17–23.

[MRT18] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[OCo03] R. O'Connell, «A typology of child cybersexploitation and online grooming practices», *Cyberspace Research Unit, University of Central Lancashire*, 2003.

[ODER07] L. Olson, J. Daggs, B. Ellevold, and T. Rogers, «Entrapping the innocent: Toward a theory of child sexual predators' luring communication», *Communication Theory*, vol. 17, pp. 231–251, Jul. 2007.

[PGS15] H. Pranoto, F. E. Gunawan, and B. Soewito, «Logistic models for classifying online grooming conversation», *Procedia Computer Science*, vol. 59, pp. 357–365, 2015.

[POC09] P. Panzarasa, T. Opsahl, and K. Carley, «Patterns and dynamics of users' behavior and interaction: Network analysis of an online community», *JASIST*, vol. 60, pp. 911–932, May 2009.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, «Scikit-learn: Machine learning in Python», *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[Rey09] D. A. Reynolds, «Gaussian mixture models.», *Encyclopedia of biometrics*, vol. 741, no. 659-663, 2009.

[SAIR11] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, «Finding a "kneedle" in a haystack: Detecting knee points in system behavior», in *2011 31st International Conference on Distributed Computing Systems Workshops*, 2011, pp. 166–171.

[SB13] K. Sasirekha and P. Baby, «Agglomerative hierarchical clustering algorithm- a review», *International Journal of Scientific and Research Publications*, vol. 83, no. 3, p. 83, 2013.

[Tru13] R. J. Trudeau, *Introduction to graph theory*. Courier Corporation, 2013.

[VJE+12] E. Villatoro-Tello, A. Juárez-González, H. J. Escalante, M. Montes-y-Gómez, and L. V. Pineda, «A two-step approach for effective detection of misbehaving users in chats.», in *CLEF (Online Working Notes/Labs/Workshop)*, vol. 1178, 2012.

[WBKP08] B. Walter, K. Bala, M. Kulkarni, and K. Pingali, «Fast agglomerative clustering for rendering», in *2008 IEEE Symposium on Interactive Ray Tracing*, IEEE, 2008, pp. 81–86.

[WFMY10] J. Wolak, D. Finkelhor, K. Mitchell, and M. Ybarra, «Online "predators" and their victims: Myths, realities, and implications for prevention and treatment», *The American psychologist*, vol. 63, pp. 111–28, Aug. 2010.

[Wu12] J. Wu, «Cluster analysis and k-means clustering: An introduction», in *Advances in K-means Clustering: A Data Mining Thinking*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 1–16. [Online]. Available: https://doi.org/10.1007/978-3-642-29807-3_1.

[ZLA+18] Z. Zuo, J. Li, P. Anderson, L. Yang, and N. Naik, «Grooming detection using fuzzy-rough feature selection and text classification», in *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2018, pp. 1–8.

[ZRL97] T. Zhang, R. Ramakrishnan, and M. Livny, «Birch: A new data clustering algorithm and its applications», *Data mining and knowledge discovery*, vol. 1, no. 2, pp. 141–182, 1997.

[Aar21] A. F. Aarekol, «Graph theoretical approach to online predator detection», Pre-project report, Norwegian University of Science and Technology, 2021.

Appendix

A Results from the clustering algorithms from the one-day data set

A.1 *k*-means

Table A.1: k-means on data from the 1st month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 nmsgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	2294	2.68 (2.3)	1.06 (1.69)	2.28 (1.94)	8.07 (15.56)	3.13 (7.71)	4.95 (8.33)	0.82 (0.22)	0.01 (0.05)	0.0 (0.0)	8.4 (12.98)	1.84 (1.35)	0.61 (0.92)	0.16 (0.4)	0.06 (0.25)	1.74 (1.44)	0.4 (0.74)	0.11 (0.33)	0.03 (0.17)	0.8 (0.23)	0.15 (0.2)	0.03 (0.08)	0.01 (0.05)
1	1678	14.46 (4.96)	12.12 (4.52)	12.9 (4.66)	202.79 (131.07)	100.03 (66.81)	102.77 (69.03)	0.51 (0.09)	0.04 (0.04)	0.0 (0.0)	16.15 (7.74)	2.93 (2.74)	5.27 (2.91)	4.03 (2.28)	2.24 (1.57)	4.19 (3.07)	4.86 (2.74)	2.75 (1.85)	1.1 (1.1)	0.19 (0.14)	0.36 (0.15)	0.28 (0.13)	0.17 (0.12)
2	1682	3.24 (2.2)	2.9 (2.01)	2.89 (2.03)	132.75 (139.07)	66.92 (71.63)	65.83 (71.75)	0.5 (0.13)	0.03 (0.08)	0.0 (0.0)	13.46 (15.55)	0.44 (0.73)	0.73 (0.94)	0.51 (0.74)	1.55 (0.87)	0.55 (0.81)	0.63 (0.87)	0.82 (0.89)	0.89 (0.8)	0.11 (0.17)	0.17 (0.2)	0.12 (0.17)	0.61 (0.28)
3	3886	1.33 (0.95)	1.31 (0.92)	0.04 (0.24)	1.85 (4.51)	1.76 (4.04)	0.09 (0.95)	0.01 (0.05)	0.0 (0.03)	0.0 (0.0)	16.42 (31.49)	1.21 (0.66)	0.09 (0.35)	0.02 (0.14)	0.0 (0.06)	0.03 (0.19)	0.01 (0.1)	0.0 (0.05)	0.0 (0.02)	0.96 (0.12)	0.03 (0.11)	0.01 (0.05)	0.0 (0.02)
4	4759	2.3 (1.99)	1.98 (1.78)	1.79 (1.85)	9.62 (15.42)	5.21 (8.83)	4.4 (7.81)	0.42 (0.28)	0.01 (0.05)	0.0 (0.0)	13.91 (21.18)	0.35 (0.66)	1.74 (1.23)	0.15 (0.4)	0.06 (0.24)	0.95 (1.18)	0.74 (0.98)	0.09 (0.3)	0.02 (0.14)	0.09 (0.16)	0.87 (0.2)	0.03 (0.08)	0.01 (0.05)
5	224	2.44 (0.9)	2.2 (0.87)	2.06 (1.11)	75.18 (139.07)	38.27 (73.08)	36.92 (67.32)	0.48 (0.19)	0.88 (0.21)	0.01 (0.01)	18.46 (23.68)	0.44 (0.68)	0.73 (0.79)	0.68 (0.73)	0.59 (0.81)	0.42 (0.68)	0.71 (0.81)	0.56 (0.72)	0.37 (0.62)	0.18 (0.28)	0.31 (0.33)	0.28 (0.31)	0.23 (0.31)
6	349	23.74 (10.35)	20.88 (9.28)	22.18 (9.92)	891.21 (422.52)	439.34 (221.1)	451.87 (219.25)	0.51 (0.07)	0.04 (0.04)	0.0 (0.0)	16.36 (7.24)	2.97 (2.74)	6.48 (4.01)	6.59 (3.98)	7.7 (3.47)	4.69 (3.41)	6.77 (4.25)	5.73 (3.77)	4.99 (2.38)	0.12 (0.09)	0.26 (0.11)	0.27 (0.11)	0.35 (0.13)
7	176	45.82 (17.87)	34.52 (13.44)	42.41 (16.89)	401.49 (265.62)	194.22 (136.95)	207.28 (133.07)	0.53 (0.07)	0.02 (0.01)	0.0 (0.0)	13.97 (4.34)	12.32 (10.39)	17.79 (8.19)	11.82 (6.71)	3.89 (3.48)	17.27 (12.78)	16.52 (7.6)	7.1 (5.29)	1.52 (1.57)	0.25 (0.14)	0.39 (0.12)	0.27 (0.13)	0.09 (0.06)
8	3314	3.33 (2.4)	2.91 (2.13)	2.83 (2.18)	28.51 (27.07)	14.75 (15.2)	13.76 (13.99)	0.48 (0.18)	0.03 (0.08)	0.0 (0.0)	15.12 (17.25)	0.58 (0.88)	0.88 (1.07)	1.68 (0.99)	0.19 (0.4)	0.7 (0.97)	1.33 (1.25)	0.75 (0.86)	0.05 (0.23)	0.13 (0.18)	0.2 (0.21)	0.64 (0.28)	0.04 (0.08)
9	2	252.0 (27.0)	141.5 (20.5)	251.0 (28.0)	1198.0 (49.0)	521.5 (31.5)	676.5 (17.5)	0.57 (0.01)	0.0 (0.0)	0.0 (0.0)	9.68 (1.03)	94.5 (5.5)	114.5 (22.5)	35.5 (0.5)	7.5 (0.5)	117.0 (12.0)	115.5 (45.5)	15.0 (5.0)	3.5 (0.5)	0.38 (0.02)	0.45 (0.04)	0.14 (0.02)	0.03 (0.01)
The entire data set																							
0	18364	4.38 (7.36)	3.59 (5.84)	3.59 (6.98)	61.56 (162.79)	30.78 (81.41)	30.78 (83.19)	0.41 (0.31)	0.03 (0.11)	0.0 (0.0)	14.24 (21.04)	1.18 (2.31)	1.57 (2.99)	1.03 (2.18)	0.6 (1.52)	1.3 (2.99)	1.29 (2.9)	0.68 (1.62)	0.32 (0.95)	0.38 (0.41)	0.35 (0.37)	0.18 (0.28)	0.09 (0.21)

Appendix B

Results from the clustering algorithms the five-months data set

Table B.1: k-means on data from all five months

Cluster	# of nodes	Degree	Out-degree	In-degree	W. de-gree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neigh-bors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 mes-sage out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 mes-sage (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	60477	2.13 (2.26)	1.9 (2.11)	1.71 (2.12)	10.78 (28.83)	5.84 (15.99)	4.94 (13.81)	0.42 (0.28)	0.01 (0.05)	0.0 (0.0)	70.83 (111.39)	0.15 (0.44)	1.82 (1.65)	0.13 (0.42)	0.03 (0.18)	0.72 (1.15)	0.92 (1.3)	0.06 (0.27)	0.01 (0.1)	0.03 (0.09)	0.94 (0.12)	0.03 (0.08)	0.01 (0.04)
1	3750	161.02 (57.66)	137.99 (49.06)	145.34 (52.64)	3975.94 (2352.25)	1984.54 (1194.8)	1991.4 (1194.26)	0.5 (0.05)	0.04 (0.03)	0.0 (0.0)	109.35 (35.93)	30.83 (23.82)	66.7 (31.43)	43.07 (18.24)	20.42 (10.21)	42.5 (27.41)	63.79 (28.63)	28.51 (13.14)	10.54 (6.81)	0.18 (0.09)	0.4 (0.09)	0.27 (0.08)	0.14 (0.09)
2	33241	2.84 (3.13)	2.61 (2.92)	2.6 (3.02)	39.87 (81.84)	19.83 (42.42)	20.05 (40.97)	0.49 (0.17)	0.02 (0.08)	0.0 (0.0)	75.62 (101.05)	0.29 (0.58)	0.62 (1.09)	1.82 (1.7)	0.1 (0.41)	0.37 (0.71)	1.35 (1.73)	0.84 (1.2)	0.04 (0.23)	0.08 (0.16)	0.15 (0.2)	0.76 (0.23)	0.01 (0.06)
3	790	212.79 (89.26)	194.81 (81.77)	195.08 (82.68)	15067.58 (6650.52)	7583.69 (3469.46)	7483.89 (3324.04)	0.5 (0.04)	0.04 (0.03)	0.0 (0.0)	112.69 (33.23)	27.55 (17.4)	73.05 (37.56)	60.53 (29.03)	51.66 (20.1)	40.06 (22.28)	76.51 (38.77)	45.6 (21.71)	32.91 (13.27)	0.13 (0.06)	0.33 (0.07)	0.28 (0.05)	0.26 (0.09)
4	56312	1.49 (1.63)	1.44 (1.59)	0.08 (0.34)	2.08 (6.6)	1.97 (6.29)	0.12 (0.77)	0.02 (0.09)	0.0 (0.04)	0.0 (0.0)	80.01 (160.98)	1.34 (1.22)	0.13 (0.5)	0.02 (0.13)	0.0 (0.05)	0.07 (0.3)	0.01 (0.12)	0.0 (0.03)	0.0 (0.01)	0.96 (0.13)	0.04 (0.12)	0.0 (0.04)	0.0 (0.01)
5	87363	8.74 (8.33)	7.02 (7.47)	7.36 (7.31)	111.86 (197.58)	55.16 (100.04)	56.7 (100.88)	0.59 (0.21)	0.04 (0.08)	0.0 (0.0)	87.82 (106.7)	2.53 (2.56)	3.65 (4.15)	1.87 (2.43)	0.68 (1.17)	2.79 (2.98)	3.1 (3.72)	1.16 (1.7)	0.3 (0.7)	0.43 (0.3)	0.35 (0.2)	0.16 (0.15)	0.06 (0.1)
6	248	538.91 (212.13)	441.59 (168.3)	490.49 (203.29)	10645.54 (8256.25)	5322.71 (4159.36)	5322.83 (4164.7)	0.5 (0.05)	0.02 (0.02)	0.0 (0.0)	97.31 (28.83)	120.99 (94.6)	229.77 (118.6)	133.11 (59.94)	55.04 (37.89)	166.74 (144.13)	214.77 (94.58)	82.55 (44.86)	26.42 (23.3)	0.22 (0.11)	0.42 (0.08)	0.25 (0.09)	0.11 (0.07)
7	9323	2.73 (3.11)	2.53 (2.87)	2.49 (2.91)	262.7 (622.07)	132.93 (325.6)	129.77 (313.09)	0.5 (0.14)	0.03 (0.08)	0.0 (0.0)	62.89 (92.67)	0.27 (0.6)	0.5 (0.96)	0.37 (0.84)	1.59 (1.37)	0.31 (0.69)	0.47 (0.96)	0.93 (1.1)	0.79 (1.1)	0.07 (0.16)	0.12 (0.19)	0.08 (0.16)	0.73 (0.25)
8	14855	56.93 (22.81)	49.35 (20.21)	50.74 (20.52)	1429.15 (1404.14)	710.86 (712.98)	718.29 (709.57)	0.5 (0.06)	0.04 (0.04)	0.0 (0.0)	106.43 (40.55)	10.84 (8.76)	23.79 (12.32)	15.17 (7.52)	7.13 (4.81)	14.74 (9.83)	22.37 (11.16)	9.89 (5.47)	3.74 (3.35)	0.18 (0.1)	0.41 (0.1)	0.27 (0.09)	0.14 (0.1)
9	6340	2.77 (1.64)	2.44 (1.57)	2.24 (1.7)	104.66 (536.15)	53.21 (283.49)	51.45 (258.84)	0.47 (0.21)	0.86 (0.22)	0.0 (0.0)	100.71 (129.63)	0.58 (0.76)	1.04 (0.96)	0.72 (0.87)	0.43 (0.81)	0.55 (0.76)	0.92 (0.97)	0.54 (0.8)	0.24 (0.61)	0.22 (0.29)	0.39 (0.33)	0.26 (0.29)	0.13 (0.23)
Clustering 2: 7 Clusters																							
0	76328	1.83 (2.45)	1.41 (2.17)	0.6 (1.47)	4.48 (23.13)	2.88 (13.08)	1.6 (11.05)	0.24 (0.39)	0.01 (0.04)	0.0 (0.0)	73.61 (148.05)	1.55 (1.65)	0.22 (0.76)	0.05 (0.24)	0.02 (0.14)	0.48 (1.16)	0.09 (0.4)	0.02 (0.16)	0.01 (0.08)	0.94 (0.14)	0.05 (0.12)	0.01 (0.04)	0.0 (0.03)
1	3708	164.89 (62.84)	142.54 (52.51)	149.29 (57.56)	5392.81 (3788.77)	2692.05 (1926.31)	2700.76 (1909.28)	0.5 (0.05)	0.04 (0.03)	0.0 (0.0)	109.44 (35.87)	30.09 (25.24)	66.25 (34.02)	44.42 (19.14)	24.13 (12.02)	41.58 (29.79)	64.21 (30.9)	30.23 (13.7)	13.28 (8.53)	0.17 (0.09)	0.39 (0.09)	0.27 (0.07)	0.17 (0.1)
2	90250	3.98 (5.19)	3.3 (4.47)	3.22 (4.44)	27.32 (76.09)	13.98 (39.59)	13.34 (37.97)	0.46 (0.25)	0.02 (0.06)	0.0 (0.0)	79.12 (119.92)	0.87 (1.69)	2.61 (2.99)	0.36 (0.82)	0.13 (0.45)	1.44 (2.27)	1.52 (2.22)	0.21 (0.57)	0.05 (0.26)	0.14 (0.19)	0.8 (0.22)	0.04 (0.09)	0.02 (0.06)
3	14834	60.51 (25.73)	52.19 (22.53)	53.85 (23.3)	1435.46 (1335.85)	714.56 (680.6)	720.89 (673.98)	0.5 (0.06)	0.04 (0.04)	0.0 (0.0)	106.51 (40.29)	11.86 (9.74)	25.52 (13.71)	15.85 (8.31)	7.27 (4.93)	16.06 (11.19)	23.76 (12.42)	10.26 (5.92)	3.77 (3.37)	0.19 (0.1)	0.41 (0.1)	0.27 (0.09)	0.14 (0.1)
4	591	387.21 (194.2)	335.09 (149.02)	352.96 (181.63)	15119.6 (8808.57)	7627.66 (4512.93)	7491.94 (4406.12)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	108.57 (30.64)	69.68 (73.9)	151.8 (103.19)	103.69 (49.47)	62.04 (27.82)	97.9 (108.38)	149.26 (85.7)	70.39 (33.6)	35.41 (19.21)	0.16 (0.09)	0.38 (0.08)	0.28 (0.06)	0.19 (0.09)
5	69957	8.2 (8.54)	7.06 (7.44)	7.22 (7.64)	133.35 (221.87)	65.72 (112.24)	67.63 (113.28)	0.5 (0.14)	0.04 (0.07)	0.0 (0.0)	87.06 (89.28)	1.64 (2.36)	2.92 (3.86)	2.88 (2.57)	0.76 (1.3)	1.92 (2.72)	3.31 (3.8)	1.65 (1.9)	0.34 (0.77)	0.16 (0.17)	0.25 (0.2)	0.52 (0.29)	0.06 (0.11)
6	17031	3.19 (3.24)	2.85 (2.97)	2.73 (2.98)	193.83 (537.74)	97.61 (280.19)	96.21 (270.24)	0.49 (0.17)	0.38 (0.41)	0.0 (0.0)	82.45 (111.79)	0.54 (0.87)	0.96 (1.3)	0.57 (1.02)	1.12 (1.31)	0.55 (0.93)	0.82 (1.25)	0.79 (1.06)	0.57 (0.98)	0.15 (0.23)	0.26 (0.28)	0.14 (0.22)	0.44 (0.38)
Clustering 3: 5 Clusters																							
0	9557	101.85 (45.72)	87.97 (39.36)	91.47 (41.53)	2873.75 (2542.08)	1432.21 (1284.74)	1441.55 (1284.22)	0.5 (0.05)	0.04 (0.04)	0.0 (0.0)	108.21 (37.92)	19.26 (16.49)	42.02 (23.31)	27.12 (13.97)	13.46 (8.62)	26.37 (19.13)	39.94 (21.27)	17.94 (10.11)	7.22 (5.82)	0.18 (0.09)	0.4 (0.09)	0.27 (0.08)	0.15 (0.1)
1	77453	1.84 (2.47)	1.42 (2.24)	0.6 (1.37)	4.13 (17.09)	2.73 (10.34)	1.4 (7.92)	0.24 (0.39)	0.01 (0.07)	0.0 (0.0)	73.86 (147.95)	1.55 (1.68)	0.22 (0.77)	0.06 (0.26)	0.01 (0.12)	0.48 (1.1)	0.09 (0.4)	0.02 (0.16)	0.0 (0.06)	0.94 (0.14)	0.05 (0.12)	0.01 (0.06)	0.0 (0.03)
2	85774	3.3 (3.94)	2.74 (3.42)	2.66 (3.41)	17.48 (40.76)	9.0 (21.45)	8.48 (20.23)	0.45 (0.26)	0.03 (0.12)	0.0 (0.0)	77.78 (122.1)	0.69 (1.3)	2.29 (2.41)	0.26 (0.63)	0.06 (0.26)	1.21 (1.84)	1.29 (1.81)	0.13 (0.4)	0.02 (0.16)	0.13 (0.19)	0.82 (0.21)	0.04 (0.09)	0.01 (0.04)
3	1313	303.37 (159.39)	262.16 (126.6)	276.73 (147.83)	11322.85 (7672.13)	5699.47 (3929.02)	5623.38 (3828.19)	0.5 (0.05)	0.03 (0.03)	0.0 (0.0)	108.94 (32.23)	54.46 (57.13)	119.36 (81.62)	81.59 (42.0)	47.97 (24.83)	76.47 (81.15)	117.49 (70.13)	55.64 (28.6)	27.14 (16.91)	0.16 (0.09)	0.38 (0.08)	0.28 (0.07)	0.18 (0.1)
4	98602	10.82 (13.27)	9.31 (11.54)	9.51 (11.77)	218.18 (436.13)	108.29 (221.66)	109.89 (220.89)	0.5 (0.13)	0.09 (0.2)	0.0 (0.0)	88.58 (89.11)	2.18 (3.55)	4.2 (6.23)	3.2 (3.72)	1.23 (1.94)	2.71 (4.32)	4.2 (5.65)	2.01 (2.59)	0.59 (1.19)	0.16 (0.17)	0.28 (0.21)	0.42 (0.3)	0.14 (0.23)

Table B.2: GMM on data from all five months

Cluster	# of nodes	Degree	Out-degree	In-degree	W. de-gree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neigh-bors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 mes-sage out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 mes-sage (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 15 Clusters																							
0	11100	6.1 (1.61)	5.25 (1.69)	5.36 (1.7)	109.68 (53.94)	53.48 (27.8)	56.2 (28.86)	0.51 (0.09)	0.07 (0.1)	0.0 (0.0)	91.88 (76.05)	1.25 (1.15)	2.25 (1.32)	1.38 (1.13)	1.22 (0.48)	1.47 (1.2)	2.03 (1.32)	1.43 (1.02)	0.43 (0.55)	0.2 (0.18)	0.37 (0.19)	0.22 (0.18)	0.21 (0.07)
1	16161	1.3 (0.55)	1.3 (0.55)	0.06 (0.24)	3.12 (1.45)	3.06 (1.48)	0.06 (0.24)	0.02 (0.08)	0.0 (0.0)	0.0 (0.0)	56.84 (104.95)	0.27 (0.51)	1.04 (0.19)	0.0 (0.0)	0.0 (0.0)	0.06 (0.24)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.12 (0.22)	0.88 (0.22)	0.0 (0.0)	0.0 (0.0)
2	4528	5.9 (3.5)	5.33 (3.3)	4.9 (3.39)	793.11 (1667.58)	401.73 (848.96)	391.38 (836.28)	0.47 (0.17)	0.22 (0.28)	0.0 (0.0)	82.68 (77.9)	0.92 (1.12)	1.57 (1.5)	1.25 (1.31)	2.17 (1.57)	0.89 (1.11)	1.37 (1.41)	1.17 (1.29)	1.48 (1.28)	0.15 (0.17)	0.25 (0.2)	0.2 (0.2)	0.4 (0.21)
3	48321	1.13 (0.42)	1.13 (0.42)	0.0 (0.0)	1.13 (0.42)	1.13 (0.42)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	78.92 (168.41)	1.13 (0.42)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
4	38375	2.48 (1.58)	2.15 (1.39)	2.13 (1.46)	21.47 (11.96)	10.76 (6.59)	10.71 (7.02)	0.49 (0.18)	0.0 (0.0)	0.0 (0.0)	75.83 (100.28)	0.48 (0.74)	0.77 (1.02)	1.23 (0.53)	0.0 (0.0)	0.54 (0.84)	1.1 (1.03)	0.49 (0.59)	0.0 (0.0)	0.14 (0.21)	0.21 (0.25)	0.65 (0.31)	0.0 (0.0)
5	10382	31.08 (9.28)	26.09 (8.79)	27.04 (9.23)	297.89 (140.0)	148.54 (70.99)	149.35 (73.88)	0.5 (0.07)	0.03 (0.03)	0.0 (0.0)	101.34 (42.15)	7.19 (4.58)	14.27 (5.85)	7.73 (4.18)	1.88 (1.45)	9.44 (5.23)	12.59 (5.63)	4.37 (2.9)	0.63 (0.81)	0.23 (0.13)	0.46 (0.12)	0.25 (0.12)	0.06 (0.05)
6	8251	81.43 (29.63)	70.35 (26.61)	73.35 (28.39)	1438.16 (942.65)	715.65 (472.08)	722.51 (479.88)	0.5 (0.05)	0.04 (0.03)	0.0 (0.0)	108.15 (37.05)	15.23 (8.48)	35.15 (15.14)	22.23 (10.63)	8.82 (5.96)	21.67 (11.2)	33.43 (14.72)	14.08 (7.91)	4.17 (3.65)	0.19 (0.09)	0.43 (0.09)	0.27 (0.09)	0.11 (0.06)
7	4469	178.8 (128.64)	153.67 (107.59)	158.95 (120.92)	6498.82 (5900.1)	3267.34 (2998.77)	3231.48 (2957.73)	0.49 (0.11)	0.04 (0.03)	0.0 (0.0)	106.76 (36.84)	35.43 (38.95)	70.98 (60.41)	45.86 (36.24)	26.52 (21.58)	45.84 (53.56)	67.07 (56.03)	31.0 (25.09)	15.04 (13.67)	0.2 (0.14)	0.38 (0.1)	0.25 (0.09)	0.17 (0.11)
8	9830	2.37 (0.53)	2.09 (0.68)	2.0 (0.82)	51.17 (39.82)	25.11 (20.48)	26.07 (21.23)	0.49 (0.18)	0.45 (0.47)	0.0 (0.0)	98.29 (162.16)	0.46 (0.61)	0.76 (0.69)	0.48 (0.62)	0.67 (0.55)	0.45 (0.6)	0.66 (0.69)	0.66 (0.68)	0.22 (0.42)	0.2 (0.27)	0.33 (0.31)	0.2 (0.27)	0.27 (0.22)
9	29654	8.23 (4.91)	6.8 (4.37)	6.5 (4.52)	44.23 (32.77)	22.45 (16.92)	21.78 (17.42)	0.47 (0.16)	0.14 (0.16)	0.0 (0.0)	102.19 (82.0)	2.46 (2.32)	3.91 (2.97)	1.85 (1.75)	0.0 (0.0)	2.62 (2.49)	3.15 (2.63)	0.73 (0.98)	0.0 (0.0)	0.3 (0.23)	0.47 (0.22)	0.23 (0.2)	0.0 (0.0)
10	7992	31.12 (12.61)	27.9 (11.77)	28.2 (12.0)	1324.23 (1075.88)	653.12 (543.5)	671.1 (541.94)	0.51 (0.06)	0.06 (0.06)	0.0 (0.0)	104.77 (47.96)	4.77 (3.08)	11.22 (5.48)	8.53 (4.69)	6.6 (3.9)	6.55 (3.71)	11.07 (5.63)	6.6 (3.96)	3.98 (2.78)	0.16 (0.09)	0.36 (0.11)	0.27 (0.1)	0.21 (0.09)
11	33116	1.25 (0.6)	1.06 (0.69)	1.25 (0.6)	4.47 (2.84)	1.99 (1.68)	2.47 (1.7)	0.57 (0.21)	0.0 (0.0)	0.0 (0.0)	75.66 (155.52)	0.0 (0.0)	1.25 (0.6)	0.0 (0.0)	0.0 (0.0)	0.53 (0.62)	0.72 (0.66)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)
12	12126	13.4 (3.29)	11.56 (3.3)	11.9 (3.46)	219.99 (115.35)	107.55 (58.28)	112.44 (60.58)	0.51 (0.07)	0.05 (0.06)	0.0 (0.0)	97.16 (54.62)	2.58 (1.81)	5.41 (2.34)	3.62 (2.1)	1.78 (0.98)	3.33 (1.99)	5.14 (2.39)	2.67 (1.71)	0.76 (0.81)	0.19 (0.13)	0.4 (0.15)	0.27 (0.14)	0.13 (0.07)
13	4092	1.06 (0.26)	1.05 (0.27)	1.02 (0.32)	134.23 (250.59)	67.61 (127.04)	66.62 (125.96)	0.49 (0.15)	0.0 (0.0)	0.0 (0.0)	50.05 (96.71)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.06 (0.26)	0.0 (0.0)	0.01 (0.08)	0.53 (0.52)	0.48 (0.56)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)
14	34302	2.38 (1.5)	1.33 (1.53)	1.77 (1.21)	4.69 (4.23)	2.17 (2.7)	2.52 (2.21)	0.7 (0.3)	0.0 (0.0)	0.0 (0.0)	61.9 (95.15)	1.43 (0.9)	0.95 (1.12)	0.0 (0.0)	0.0 (0.0)	1.3 (0.98)	0.46 (0.73)	0.0 (0.0)	0.0 (0.0)	0.7 (0.3)	0.3 (0.3)	0.0 (0.0)	0.0 (0.0)
Clustering 2: 10 Clusters																							
0	30262	8.08 (3.56)	6.75 (3.33)	6.49 (3.44)	55.56 (36.08)	28.17 (18.73)	27.4 (19.03)	0.48 (0.14)	0.1 (0.1)	0.0 (0.0)	100.55 (74.23)	2.28 (2.0)	3.71 (2.35)	1.8 (1.54)	0.29 (0.51)	2.46 (2.06)	3.02 (2.16)	1.01 (1.09)	0.0 (0.0)	0.28 (0.21)	0.46 (0.21)	0.22 (0.18)	0.04 (0.07)
1	8267	6.85 (4.58)	6.1 (4.24)	5.95 (4.34)	585.63 (1021.23)	293.3 (521.29)	292.34 (512.26)	0.5 (0.14)	0.18 (0.26)	0.0 (0.0)	87.82 (82.23)	1.1 (1.31)	2.0 (1.92)	1.6 (1.67)	2.14 (1.65)	1.2 (1.38)	1.89 (1.89)	1.38 (1.52)	1.48 (1.24)	0.14 (0.17)	0.26 (0.2)	0.21 (0.2)	0.38 (0.26)
2	81539	1.71 (1.18)	1.19 (1.1)	1.26 (1.05)	4.3 (3.39)	2.23 (2.17)	2.07 (2.02)	0.53 (0.34)	0.0 (0.0)	0.0 (0.0)	66.61 (127.29)	0.62 (0.89)	1.09 (0.84)	0.0 (0.0)	0.0 (0.0)	0.77 (0.88)	0.49 (0.68)	0.0 (0.0)	0.0 (0.0)	0.31 (0.4)	0.69 (0.4)	0.0 (0.0)	0.0 (0.0)
3	14304	45.27 (17.69)	39.3 (15.94)	40.69 (16.67)	864.7 (553.64)	427.73 (276.81)	436.97 (283.61)	0.5 (0.05)	0.04 (0.04)	0.0 (0.0)	105.26 (41.03)	8.32 (5.11)	19.24 (9.28)	12.35 (6.5)	5.35 (3.7)	11.75 (6.63)	18.23 (8.97)	8.04 (4.9)	2.66 (2.34)	0.18 (0.09)	0.42 (0.11)	0.27 (0.1)	0.13 (0.08)
4	50254	1.16 (0.46)	1.16 (0.46)	0.0 (0.0)	1.22 (0.65)	1.22 (0.65)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	78.76 (166.72)	1.13 (0.43)	0.04 (0.19)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.98 (0.1)	0.02 (0.1)	0.0 (0.0)	0.0 (0.0)
5	19232	17.51 (6.83)	14.92 (6.23)	15.37 (6.32)	229.18 (114.21)	112.49 (57.86)	116.69 (59.92)	0.51 (0.07)	0.04 (0.05)	0.0 (0.0)	97.04 (50.08)	3.74 (2.92)	7.57 (4.22)	4.51 (2.83)	1.69 (1.25)	4.87 (3.41)	6.87 (3.83)	2.87 (2.12)	0.77 (0.79)	0.21 (0.13)	0.42 (0.14)	0.26 (0.14)	0.11 (0.08)
6	18599	2.44 (1.01)	2.16 (1.0)	2.05 (1.1)	53.04 (40.84)	26.15 (21.01)	26.9 (21.9)	0.49 (0.17)	0.32 (0.4)	0.0 (0.0)	88.21 (126.96)	0.5 (0.7)	0.77 (0.81)	0.49 (0.7)	0.68 (0.55)	0.47 (0.66)	0.68 (0.8)	0.65 (0.69)	0.24 (0.44)	0.18 (0.27)	0.28 (0.3)	0.17 (0.25)	0.36 (0.36)
7	5686	115.1 (88.64)	104.07 (80.29)	103.74 (81.71)	5886.81 (5335.8)	2951.78 (2717.12)	2935.03 (2668.89)	0.5 (0.07)	0.05 (0.04)	0.0 (0.0)	110.47 (39.21)	17.19 (16.07)	42.14 (34.85)	32.2 (27.02)	23.58 (18.92)	23.78 (20.51)	42.09 (35.61)	23.71 (20.24)	14.16 (11.9)	0.15 (0.08)	0.36 (0.09)	0.27 (0.07)	0.22 (0.1)
8	40681	2.76 (1.96)	2.37 (1.69)	2.36 (1.78)	22.73 (13.59)	11.4 (7.38)	11.33 (7.8)	0.49 (0.18)	0.0 (0.0)	0.0 (0.0)	76.54 (100.81)	0.57 (0.88)	0.92 (1.23)	1.27 (0.61)	0.0 (0.0)	0.64 (0.99)	1.21 (1.17)	0.51 (0.61)	0.0 (0.0)	0.15 (0.21)	0.23 (0.26)	0.63 (0.31)	0.0 (0.0)
9	3875	146.92 (121.33)	119.0 (97.92)	128.88 (114.49)	1694.97 (1747.94)	850.48 (878.27)	844.49 (884.6)	0.48 (0.13)	0.02 (0.02)	0.0 (0.0)	100.96 (36.86)	37.05 (38.79)	65.8 (58.84)	34.48 (32.42)	9.58 (11.06)	46.91 (54.77)	58.63 (53.54)	19.53 (20.37)	3.81 (5.07)	0.28 (0.17)	0.44 (0.11)	0.22 (0.11)	0.06 (0.04)
The entire data set																							

Table B.3: BIRCH on data from all five months

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	268	382.76 (144.22)	345.82 (126.21)	353.34 (135.17)	20308.22 (9014.02)	10257.29 (4666.43)	10050.93 (4499.98)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	111.81 (28.31)	51.3 (32.83)	135.99 (65.19)	114.07 (48.84)	81.4 (25.39)	75.23 (39.96)	145.69 (67.76)	83.97 (34.92)	48.45 (17.76)	0.13 (0.05)	0.34 (0.06)	0.3 (0.05)	0.23 (0.08)
1	21189	56.09 (29.28)	48.44 (26.27)	49.86 (26.86)	1399.7 (1770.98)	696.89 (894.14)	702.81 (891.8)	0.5 (0.07)	0.04 (0.04)	0.0 (0.0)	105.08 (40.41)	10.99 (8.78)	23.63 (14.08)	14.69 (9.34)	6.78 (6.38)	14.87 (10.41)	21.97 (13.29)	9.44 (6.83)	3.58 (4.22)	0.2 (0.12)	0.42 (0.11)	0.26 (0.1)	0.12 (0.1)
2	502	317.77 (60.17)	263.39 (63.0)	291.42 (55.54)	6096.02 (3370.09)	3079.11 (1764.01)	3016.91 (1652.85)	0.5 (0.05)	0.03 (0.02)	0.0 (0.0)	104.68 (33.35)	66.2 (39.24)	137.2 (40.0)	80.9 (26.24)	33.48 (18.15)	94.99 (48.79)	130.46 (37.88)	49.89 (20.28)	16.08 (11.22)	0.2 (0.11)	0.43 (0.08)	0.26 (0.08)	0.11 (0.07)
3	1860	181.44 (38.15)	155.83 (38.52)	162.72 (35.24)	3931.03 (2199.04)	1961.05 (1111.59)	1969.98 (1119.51)	0.5 (0.06)	0.04 (0.03)	0.0 (0.0)	108.77 (35.63)	35.18 (23.3)	74.61 (23.86)	49.33 (16.51)	22.32 (11.48)	47.01 (22.55)	71.56 (21.45)	33.19 (13.52)	10.96 (7.49)	0.19 (0.1)	0.41 (0.08)	0.28 (0.08)	0.13 (0.07)
4	59	701.44 (184.66)	527.1 (176.26)	625.03 (207.77)	7556.44 (3917.77)	3761.71 (2025.61)	3794.73 (1982.79)	0.5 (0.07)	0.02 (0.01)	0.0 (0.0)	81.41 (28.01)	210.9 (96.2)	316.14 (106.92)	138.81 (82.01)	35.59 (25.31)	261.1 (125.65)	277.59 (141.31)	71.71 (50.69)	14.63 (12.45)	0.31 (0.12)	0.45 (0.08)	0.19 (0.09)	0.05 (0.04)
5	387	158.66 (54.63)	146.12 (50.28)	143.63 (49.02)	14801.19 (4555.62)	7402.1 (2409.12)	7399.09 (2371.87)	0.5 (0.04)	0.04 (0.02)	0.0 (0.0)	111.99 (33.9)	21.25 (13.9)	51.85 (23.53)	42.46 (16.65)	43.1 (14.3)	28.77 (14.15)	52.97 (22.55)	32.89 (13.39)	28.98 (9.82)	0.13 (0.06)	0.32 (0.07)	0.26 (0.06)	0.29 (0.1)
6	2	2.5 (0.5)	2.0 (0.0)	2.5 (0.5)	2399.0 (145.0)	1086.5 (76.5)	1312.5 (68.5)	0.55 (0.0)	0.75 (0.25)	0.04 (0.01)	38.08 (35.42)	0.0 (0.0)	0.5 (0.5)	0.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.5 (0.5)	0.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.17 (0.17)	0.0 (0.0)	0.83 (0.17)
7	1	2701.0 (0.0)	1906.0 (0.0)	2493.0 (0.0)	13738.0 (0.0)	7592.0 (0.0)	6146.0 (0.0)	0.45 (0.0)	0.01 (0.0)	0.0 (0.0)	57.24 (0.0)	917.0 (0.0)	1553.0 (0.0)	186.0 (0.0)	45.0 (0.0)	1872.0 (0.0)	552.0 (0.0)	54.0 (0.0)	15.0 (0.0)	0.34 (0.0)	0.57 (0.0)	0.07 (0.0)	0.02 (0.0)
8	248344	3.91 (4.95)	3.29 (4.37)	3.04 (4.47)	56.05 (192.41)	28.12 (99.01)	27.93 (97.85)	0.4 (0.3)	0.04 (0.15)	0.0 (0.0)	79.48 (122.67)	1.14 (1.51)	1.62 (2.38)	0.84 (1.56)	0.31 (0.86)	1.06 (1.71)	1.32 (2.2)	0.52 (1.13)	0.14 (0.52)	0.39 (0.41)	0.39 (0.38)	0.17 (0.28)	0.05 (0.16)
9	87	2.55 (0.89)	2.43 (0.83)	2.26 (0.99)	906.17 (1676.95)	491.08 (951.48)	415.09 (755.19)	0.46 (0.15)	0.85 (0.25)	0.01 (0.0)	57.41 (74.42)	0.09 (0.29)	0.26 (0.49)	0.43 (0.65)	1.77 (0.72)	0.06 (0.23)	0.38 (0.61)	0.46 (0.6)	1.37 (0.82)	0.03 (0.08)	0.08 (0.15)	0.17 (0.28)	0.72 (0.3)
Clustering 2: 7 Clusters																							
0	60	734.77 (314.73)	550.08 (248.42)	656.17 (315.65)	7659.47 (3964.76)	3825.55 (2067.64)	3833.92 (1989.11)	0.5 (0.07)	0.02 (0.01)	0.0 (0.0)	81.01 (27.95)	222.67 (131.42)	336.75 (190.56)	139.6 (81.55)	35.75 (25.12)	287.95 (240.95)	282.17 (144.46)	71.42 (50.31)	14.63 (12.34)	0.31 (0.11)	0.45 (0.08)	0.19 (0.09)	0.05 (0.04)
1	23049	66.21 (45.51)	57.11 (40.12)	58.97 (41.33)	1603.97 (1936.2)	798.91 (976.34)	805.07 (975.39)	0.5 (0.07)	0.04 (0.04)	0.0 (0.0)	105.38 (40.05)	12.94 (12.57)	27.75 (20.52)	17.49 (13.83)	8.03 (8.12)	17.47 (14.74)	25.97 (19.55)	11.36 (9.97)	4.18 (4.99)	0.2 (0.12)	0.42 (0.11)	0.26 (0.1)	0.12 (0.1)
2	502	317.77 (60.17)	263.39 (63.0)	291.42 (55.54)	6096.02 (3370.09)	3079.11 (1764.01)	3016.91 (1652.85)	0.5 (0.05)	0.03 (0.02)	0.0 (0.0)	104.68 (33.35)	66.2 (39.24)	137.2 (40.0)	80.9 (26.24)	33.48 (18.15)	94.99 (48.79)	130.46 (37.88)	49.89 (20.28)	16.08 (11.22)	0.2 (0.11)	0.43 (0.08)	0.26 (0.08)	0.11 (0.07)
3	268	382.76 (144.22)	345.82 (126.21)	353.34 (135.17)	20308.22 (9014.02)	10257.29 (4666.43)	10050.93 (4499.98)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	111.81 (28.31)	51.3 (32.83)	135.99 (65.19)	114.07 (48.84)	81.4 (25.39)	75.23 (39.96)	145.69 (67.76)	83.97 (34.92)	48.45 (17.76)	0.13 (0.05)	0.34 (0.06)	0.3 (0.05)	0.23 (0.08)
4	248431	3.91 (4.95)	3.29 (4.36)	3.03 (4.47)	56.35 (195.57)	28.28 (100.95)	28.06 (99.12)	0.4 (0.3)	0.04 (0.15)	0.0 (0.0)	79.47 (122.66)	1.14 (1.51)	1.62 (2.38)	0.84 (1.56)	0.31 (0.86)	1.05 (1.71)	1.32 (2.19)	0.52 (1.13)	0.14 (0.52)	0.39 (0.41)	0.39 (0.38)	0.17 (0.28)	0.05 (0.16)
5	387	158.66 (54.63)	146.12 (50.28)	143.63 (49.02)	14801.19 (4555.62)	7402.1 (2409.12)	7399.09 (2371.87)	0.5 (0.04)	0.04 (0.02)	0.0 (0.0)	111.99 (33.9)	21.25 (13.9)	51.85 (23.53)	42.46 (16.65)	43.1 (14.3)	28.77 (14.15)	52.97 (22.55)	32.89 (13.39)	28.98 (9.82)	0.13 (0.06)	0.32 (0.07)	0.26 (0.06)	0.29 (0.1)
6	2	2.5 (0.5)	2.0 (0.0)	2.5 (0.5)	2399.0 (145.0)	1086.5 (76.5)	1312.5 (68.5)	0.55 (0.0)	0.75 (0.25)	0.04 (0.01)	38.08 (35.42)	0.0 (0.0)	0.5 (0.5)	0.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.5 (0.5)	0.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.17 (0.17)	0.0 (0.0)	0.83 (0.17)
Clustering 3: 5 Clusters																							
0	889	248.51 (97.81)	212.34 (81.99)	227.08 (90.32)	9885.56 (5837.45)	4961.0 (2979.52)	4924.56 (2951.64)	0.5 (0.05)	0.03 (0.02)	0.0 (0.0)	107.86 (33.78)	46.63 (38.08)	100.05 (54.18)	64.17 (29.54)	37.66 (17.26)	66.16 (50.09)	96.73 (50.08)	42.49 (19.52)	21.7 (12.41)	0.17 (0.1)	0.38 (0.09)	0.26 (0.07)	0.19 (0.12)
1	60	734.77 (314.73)	550.08 (248.42)	656.17 (315.65)	7659.47 (3964.76)	3825.55 (2067.64)	3833.92 (1989.11)	0.5 (0.07)	0.02 (0.01)	0.0 (0.0)	81.01 (27.95)	222.67 (131.42)	336.75 (190.56)	139.6 (81.55)	35.75 (25.12)	287.95 (240.95)	282.17 (144.46)	71.42 (50.31)	14.63 (12.34)	0.31 (0.11)	0.45 (0.08)	0.19 (0.09)	0.05 (0.04)
2	248433	3.91 (4.95)	3.29 (4.36)	3.03 (4.47)	56.36 (195.68)	28.29 (101.0)	28.07 (99.18)	0.4 (0.3)	0.04 (0.15)	0.0 (0.0)	79.47 (122.65)	1.14 (1.51)	1.62 (2.38)	0.84 (1.56)	0.31 (0.86)	1.05 (1.71)	1.32 (2.19)	0.52 (1.13)	0.14 (0.52)	0.39 (0.41)	0.39 (0.38)	0.17 (0.28)	0.05 (0.16)
3	268	382.76 (144.22)	345.82 (126.21)	353.34 (135.17)	20308.22 (9014.02)	10257.29 (4666.43)	10050.93 (4499.98)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	111.81 (28.31)	51.3 (32.83)	135.99 (65.19)	114.07 (48.84)	81.4 (25.39)	75.23 (39.96)	145.69 (67.76)	83.97 (34.92)	48.45 (17.76)	0.13 (0.05)	0.34 (0.06)	0.3 (0.05)	0.23 (0.08)
4	23049	66.21 (45.51)	57.11 (40.12)	58.97 (41.33)	1603.97 (1936.2)	798.91 (976.34)	805.07 (975.39)	0.5 (0.07)	0.04 (0.04)	0.0 (0.0)	105.38 (40.05)	12.94 (12.57)	27.75 (20.52)	17.49 (13.83)	8.03 (8.12)	17.47 (14.74)	25.97 (19.55)	11.36 (9.97)	4.18 (4.99)	0.2 (0.12)	0.42 (0.11)	0.26 (0.1)	0.12 (0.1)
The entire data set																							

Appendix C

Results from the clustering algorithms from the individual months

Table C.1: k-means on data from the 1st month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	26772	3.13 (3.13)	2.68 (2.75)	2.53 (2.82)	20.47 (41.17)	10.62 (21.61)	9.85 (20.64)	0.43 (0.26)	0.02 (0.07)	0.0 (0.0)	32.76 (41.66)	0.54 (0.99)	2.19 (1.85)	0.3 (0.63)	0.1 (0.34)	1.08 (1.43)	1.23 (1.58)	0.17 (0.47)	0.04 (0.2)	0.1 (0.16)	0.83 (0.21)	0.05 (0.09)	0.02 (0.06)
1	8335	23.0 (8.87)	19.79 (8.06)	20.49 (8.04)	476.5 (438.86)	236.09 (221.95)	240.41 (226.05)	0.5 (0.07)	0.04 (0.04)	0.0 (0.0)	43.58 (17.78)	4.45 (4.06)	9.73 (5.21)	6.06 (3.39)	2.75 (2.12)	6.09 (4.29)	9.03 (4.8)	3.96 (2.56)	1.4 (1.49)	0.18 (0.12)	0.41 (0.13)	0.27 (0.12)	0.13 (0.1)
2	601	65.07 (31.36)	59.37 (28.07)	60.86 (29.76)	4410.25 (1813.83)	2197.94 (925.62)	2212.31 (939.84)	0.5 (0.04)	0.03 (0.03)	0.0 (0.0)	42.66 (12.21)	7.7 (6.52)	21.31 (12.55)	18.68 (10.59)	17.39 (7.92)	11.43 (7.8)	23.45 (13.79)	14.58 (8.27)	11.39 (5.35)	0.11 (0.07)	0.31 (0.09)	0.28 (0.07)	0.29 (0.11)
3	18562	1.44 (1.44)	1.42 (1.39)	0.07 (0.37)	2.08 (5.81)	1.93 (4.89)	0.16 (1.42)	0.01 (0.06)	0.0 (0.04)	0.0 (0.0)	34.23 (67.35)	1.29 (1.0)	0.13 (0.49)	0.02 (0.13)	0.0 (0.06)	0.04 (0.28)	0.02 (0.16)	0.0 (0.05)	0.0 (0.02)	0.96 (0.12)	0.04 (0.12)	0.0 (0.03)	0.0 (0.01)
4	2287	59.41 (19.92)	51.29 (17.49)	53.99 (18.31)	1260.37 (665.89)	630.37 (345.57)	629.99 (337.19)	0.5 (0.06)	0.03 (0.03)	0.0 (0.0)	45.3 (13.96)	10.67 (8.27)	24.79 (11.84)	16.41 (7.35)	7.54 (3.95)	15.3 (9.53)	24.12 (10.98)	10.78 (5.35)	3.78 (2.65)	0.17 (0.1)	0.41 (0.1)	0.28 (0.09)	0.14 (0.09)
5	273	173.1 (63.42)	138.74 (47.65)	157.83 (65.48)	2471.08 (1789.53)	1238.5 (917.8)	1232.58 (890.11)	0.5 (0.07)	0.02 (0.02)	0.0 (0.0)	43.25 (11.0)	40.82 (32.0)	75.81 (33.89)	42.55 (22.2)	13.93 (10.9)	55.62 (38.16)	70.93 (38.08)	25.02 (15.01)	6.26 (6.12)	0.23 (0.12)	0.44 (0.1)	0.25 (0.1)	0.08 (0.06)
6	8948	2.7 (2.56)	1.08 (1.83)	2.31 (2.17)	11.03 (32.41)	4.3 (14.53)	6.73 (19.54)	0.82 (0.21)	0.01 (0.05)	0.0 (0.0)	19.96 (30.89)	1.86 (1.45)	0.63 (1.0)	0.16 (0.44)	0.06 (0.24)	1.73 (1.52)	0.45 (0.84)	0.11 (0.36)	0.03 (0.16)	0.81 (0.23)	0.15 (0.2)	0.03 (0.08)	0.01 (0.05)
7	15202	4.1 (3.37)	3.61 (2.99)	3.65 (3.11)	53.43 (67.3)	26.62 (34.6)	26.81 (34.67)	0.49 (0.15)	0.04 (0.09)	0.0 (0.0)	36.31 (37.13)	0.7 (1.02)	1.22 (1.48)	1.95 (1.33)	0.24 (0.51)	0.81 (1.15)	1.79 (1.76)	0.97 (1.08)	0.09 (0.3)	0.13 (0.17)	0.22 (0.2)	0.62 (0.27)	0.04 (0.08)
8	1603	2.66 (1.33)	2.37 (1.29)	2.15 (1.44)	96.22 (262.87)	50.14 (142.84)	46.08 (127.88)	0.46 (0.21)	0.86 (0.23)	0.0 (0.0)	35.44 (37.48)	0.51 (0.71)	0.99 (0.85)	0.69 (0.78)	0.46 (0.82)	0.5 (0.7)	0.86 (0.87)	0.52 (0.73)	0.27 (0.66)	0.2 (0.28)	0.39 (0.33)	0.26 (0.29)	0.15 (0.24)
9	4056	3.23 (2.65)	2.94 (2.46)	2.94 (2.49)	247.55 (387.14)	124.64 (200.44)	122.91 (194.58)	0.5 (0.12)	0.03 (0.08)	0.0 (0.0)	27.96 (33.23)	0.41 (0.72)	0.72 (1.0)	0.47 (0.78)	1.62 (1.07)	0.48 (0.8)	0.65 (0.98)	0.94 (1.01)	0.87 (0.94)	0.1 (0.17)	0.16 (0.2)	0.1 (0.16)	0.64 (0.27)
Clustering 2: 7 Clusters																							
0	26428	2.76 (2.84)	2.28 (2.48)	2.23 (2.57)	13.07 (25.05)	6.78 (13.33)	6.29 (12.56)	0.45 (0.27)	0.02 (0.06)	0.0 (0.0)	31.0 (41.46)	0.53 (0.96)	2.01 (1.84)	0.18 (0.46)	0.04 (0.21)	1.05 (1.47)	1.07 (1.43)	0.09 (0.31)	0.01 (0.12)	0.12 (0.19)	0.84 (0.21)	0.03 (0.08)	0.01 (0.04)
1	442	147.96 (59.55)	120.37 (45.32)	134.9 (59.71)	2165.76 (1604.97)	1088.72 (824.83)	1077.04 (796.88)	0.5 (0.06)	0.03 (0.02)	0.0 (0.0)	44.57 (11.92)	33.1 (27.96)	64.92 (30.75)	37.55 (19.58)	12.4 (9.57)	45.62 (33.33)	61.5 (33.04)	22.36 (13.28)	5.42 (5.4)	0.22 (0.12)	0.44 (0.1)	0.26 (0.1)	0.09 (0.06)
2	26386	6.53 (5.92)	5.62 (5.17)	5.79 (5.32)	128.07 (205.35)	63.84 (106.39)	64.23 (102.81)	0.5 (0.13)	0.04 (0.08)	0.0 (0.0)	36.62 (33.56)	1.25 (1.81)	2.39 (2.94)	2.11 (1.86)	0.78 (1.07)	1.58 (2.16)	2.55 (2.76)	1.3 (1.41)	0.36 (0.7)	0.15 (0.17)	0.27 (0.21)	0.43 (0.31)	0.15 (0.25)
3	5707	37.81 (16.28)	32.55 (14.25)	34.07 (14.86)	799.15 (559.27)	396.59 (284.93)	402.56 (286.37)	0.5 (0.07)	0.04 (0.04)	0.0 (0.0)	44.46 (15.76)	7.06 (6.28)	15.95 (8.99)	10.17 (5.66)	4.64 (3.05)	9.95 (7.12)	15.12 (8.39)	6.64 (4.05)	2.36 (2.04)	0.18 (0.11)	0.41 (0.12)	0.27 (0.11)	0.14 (0.11)
4	25109	1.63 (1.77)	1.25 (1.6)	0.5 (1.12)	3.16 (11.82)	2.14 (7.68)	1.02 (5.24)	0.23 (0.39)	0.01 (0.04)	0.0 (0.0)	30.31 (60.75)	1.42 (1.21)	0.17 (0.57)	0.03 (0.2)	0.01 (0.09)	0.42 (0.94)	0.06 (0.29)	0.01 (0.12)	0.0 (0.06)	0.95 (0.13)	0.04 (0.12)	0.01 (0.05)	0.0 (0.02)
5	913	61.81 (27.05)	56.26 (24.34)	57.4 (25.55)	3648.99 (1809.33)	1819.12 (916.95)	1829.87 (933.73)	0.5 (0.05)	0.03 (0.03)	0.0 (0.0)	43.17 (12.62)	7.67 (6.14)	20.85 (11.28)	17.83 (9.33)	15.45 (7.04)	11.34 (7.21)	22.48 (11.93)	13.75 (7.31)	9.83 (4.95)	0.12 (0.07)	0.33 (0.09)	0.28 (0.08)	0.27 (0.11)
6	1654	2.71 (1.42)	2.43 (1.39)	2.21 (1.53)	109.77 (288.65)	57.17 (155.95)	52.59 (140.61)	0.46 (0.21)	0.84 (0.24)	0.0 (0.0)	35.28 (37.23)	0.5 (0.7)	1.0 (0.86)	0.69 (0.79)	0.52 (0.92)	0.5 (0.7)	0.87 (0.89)	0.53 (0.74)	0.31 (0.73)	0.2 (0.28)	0.39 (0.33)	0.26 (0.29)	0.16 (0.25)
Clustering 3: 5 Clusters																							
0	28567	6.82 (6.51)	5.87 (5.66)	6.04 (5.83)	130.67 (198.19)	65.26 (102.56)	65.4 (99.58)	0.5 (0.13)	0.08 (0.19)	0.0 (0.0)	36.84 (33.49)	1.32 (1.97)	2.56 (3.26)	2.14 (1.99)	0.8 (1.1)	1.68 (2.39)	2.66 (2.99)	1.33 (1.47)	0.38 (0.72)	0.15 (0.18)	0.28 (0.22)	0.42 (0.31)	0.15 (0.24)
1	5498	39.84 (16.96)	34.51 (14.79)	36.05 (15.48)	1051.09 (859.06)	523.1 (439.04)	527.99 (434.08)	0.5 (0.06)	0.04 (0.04)	0.0 (0.0)	44.02 (15.61)	7.14 (6.56)	16.36 (9.4)	10.8 (5.82)	5.53 (3.64)	10.11 (7.52)	15.75 (8.67)	7.21 (4.24)	2.98 (2.54)	0.17 (0.11)	0.4 (0.12)	0.27 (0.1)	0.16 (0.12)
2	25252	1.64 (1.77)	1.26 (1.61)	0.5 (1.12)	3.24 (12.4)	2.18 (7.87)	1.05 (5.65)	0.23 (0.39)	0.01 (0.06)	0.0 (0.0)	30.36 (60.65)	1.42 (1.22)	0.17 (0.57)	0.04 (0.2)	0.01 (0.09)	0.42 (0.94)	0.06 (0.3)	0.01 (0.12)	0.0 (0.06)	0.95 (0.13)	0.04 (0.12)	0.01 (0.05)	0.0 (0.02)
3	886	113.87 (56.36)	96.64 (42.31)	104.58 (54.13)	3276.66 (2186.71)	1633.02 (1102.79)	1643.64 (1112.49)	0.5 (0.05)	0.03 (0.02)	0.0 (0.0)	44.69 (11.3)	21.31 (23.18)	45.92 (29.57)	30.6 (16.6)	16.04 (9.13)	29.91 (28.39)	45.47 (29.31)	20.33 (10.69)	8.88 (6.44)	0.17 (0.1)	0.39 (0.1)	0.28 (0.08)	0.17 (0.11)
4	26436	2.73 (2.79)	2.25 (2.42)	2.2 (2.52)	12.81 (24.62)	6.63 (13.01)	6.18 (12.59)	0.45 (0.27)	0.02 (0.1)	0.0 (0.0)	30.91 (41.44)	0.52 (0.95)	1.99 (1.8)	0.17 (0.45)	0.04 (0.2)	1.04 (1.45)	1.06 (1.41)	0.09 (0.3)	0.01 (0.12)	0.12 (0.19)	0.84 (0.21)	0.03 (0.08)	0.01 (0.04)
The entire data set																							
0	86639	7.25 (16.34)	6.17 (13.85)	6.17 (15.12)	148.14 (522.07)	74.07 (262.25)	74.07 (263.05)	0.41 (0.29)	0.04 (0.13)	0.0 (0.0)	33.68 (44.72)	1.68 (4.06)	3.01 (7.21)	1.77 (4.66)	0.79 (2.5)	1.94 (5.27)	2.68 (7.04)	1.13 (3.14)	0.41 (1.5)	0.38 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)

Table C.2: k-means on data from the 2nd month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	7369	25.19 (9.78)	21.79 (8.84)	22.51 (8.94)	571.0 (564.06)	282.35 (284.1)	288.64 (291.02)	0.51 (0.07)	0.05 (0.05)	0.0 (0.0)	49.75 (18.73)	4.74 (4.17)	10.59 (5.71)	6.69 (3.73)	3.17 (2.44)	6.6 (4.62)	9.86 (5.23)	4.41 (2.8)	1.64 (1.71)	0.18 (0.12)	0.41 (0.13)	0.27 (0.11)	0.14 (0.11)
1	17391	1.45 (1.54)	1.42 (1.5)	0.08 (0.43)	2.29 (10.0)	2.09 (8.47)	0.2 (2.36)	0.01 (0.06)	0.0 (0.04)	0.0 (0.0)	38.03 (64.84)	1.29 (1.05)	0.13 (0.53)	0.02 (0.15)	0.0 (0.07)	0.05 (0.31)	0.03 (0.19)	0.0 (0.06)	0.0 (0.02)	0.96 (0.12)	0.03 (0.11)	0.0 (0.03)	0.0 (0.01)
2	8363	2.57 (2.61)	0.93 (1.68)	2.24 (2.29)	9.26 (27.44)	3.44 (12.82)	5.82 (15.26)	0.84 (0.2)	0.01 (0.04)	0.0 (0.0)	22.05 (36.35)	1.83 (1.52)	0.56 (0.98)	0.14 (0.42)	0.05 (0.22)	1.71 (1.61)	0.42 (0.84)	0.1 (0.34)	0.02 (0.14)	0.83 (0.22)	0.14 (0.2)	0.03 (0.08)	0.01 (0.04)
3	615	78.86 (36.3)	71.91 (33.09)	73.36 (34.4)	5155.66 (2315.06)	2587.69 (1221.69)	2567.96 (1150.66)	0.5 (0.05)	0.04 (0.03)	0.0 (0.0)	49.33 (13.75)	9.58 (7.28)	26.34 (15.02)	23.13 (12.8)	19.8 (8.57)	14.2 (9.01)	28.65 (16.19)	17.95 (9.87)	12.55 (5.75)	0.12 (0.07)	0.32 (0.09)	0.28 (0.08)	0.28 (0.11)
4	26952	3.24 (3.38)	2.79 (2.98)	2.63 (3.0)	22.18 (47.74)	11.43 (24.38)	10.75 (24.41)	0.44 (0.25)	0.02 (0.07)	0.0 (0.0)	38.31 (46.62)	0.58 (1.07)	2.23 (1.95)	0.32 (0.68)	0.11 (0.36)	1.12 (1.51)	1.27 (1.65)	0.19 (0.5)	0.04 (0.21)	0.11 (0.17)	0.83 (0.21)	0.05 (0.09)	0.02 (0.06)
5	15127	4.23 (3.68)	3.72 (3.25)	3.74 (3.37)	54.92 (73.23)	27.32 (37.71)	27.59 (37.56)	0.49 (0.16)	0.04 (0.09)	0.0 (0.0)	41.36 (41.26)	0.75 (1.1)	1.26 (1.59)	1.98 (1.38)	0.23 (0.52)	0.84 (1.22)	1.83 (1.9)	0.98 (1.1)	0.09 (0.31)	0.14 (0.18)	0.21 (0.2)	0.62 (0.28)	0.03 (0.07)
6	231	198.86 (66.06)	160.7 (49.39)	180.54 (67.59)	2964.48 (2088.74)	1479.35 (1051.28)	1485.13 (1065.16)	0.5 (0.06)	0.03 (0.02)	0.0 (0.0)	48.03 (13.55)	46.86 (38.31)	85.83 (33.89)	48.93 (19.83)	17.25 (12.27)	63.37 (48.45)	79.96 (31.33)	29.91 (15.51)	7.3 (7.22)	0.22 (0.12)	0.43 (0.08)	0.26 (0.1)	0.09 (0.06)
7	4288	3.25 (2.73)	2.96 (2.52)	2.88 (2.5)	226.22 (355.96)	115.49 (190.47)	110.73 (178.37)	0.5 (0.14)	0.03 (0.08)	0.0 (0.0)	33.21 (36.69)	0.42 (0.73)	0.76 (1.04)	0.5 (0.79)	1.58 (1.06)	0.46 (0.78)	0.68 (1.02)	0.93 (1.03)	0.8 (0.89)	0.1 (0.17)	0.17 (0.2)	0.11 (0.16)	0.62 (0.27)
8	1672	2.71 (1.38)	2.4 (1.31)	2.18 (1.52)	115.36 (351.25)	59.0 (181.47)	56.36 (176.48)	0.46 (0.2)	0.84 (0.24)	0.0 (0.0)	44.55 (45.23)	0.55 (0.71)	0.92 (0.85)	0.71 (0.84)	0.53 (0.82)	0.5 (0.69)	0.84 (0.87)	0.54 (0.78)	0.3 (0.63)	0.22 (0.29)	0.36 (0.33)	0.26 (0.28)	0.17 (0.25)
9	2105	67.92 (22.23)	58.51 (19.76)	61.47 (20.39)	1398.81 (760.11)	695.43 (388.72)	703.38 (388.33)	0.5 (0.06)	0.04 (0.04)	0.0 (0.0)	51.66 (14.98)	12.57 (9.68)	28.54 (13.0)	18.62 (8.24)	8.18 (4.42)	17.57 (10.82)	27.7 (12.26)	12.14 (6.01)	4.06 (2.93)	0.18 (0.1)	0.41 (0.1)	0.28 (0.09)	0.13 (0.09)
Clustering 2: 7 Clusters																							
0	23926	1.63 (1.8)	1.23 (1.62)	0.53 (1.17)	3.2 (12.43)	2.14 (7.72)	1.07 (5.6)	0.24 (0.4)	0.0 (0.04)	0.0 (0.0)	33.84 (59.51)	1.42 (1.21)	0.17 (0.6)	0.04 (0.2)	0.01 (0.09)	0.45 (0.98)	0.07 (0.31)	0.01 (0.12)	0.0 (0.05)	0.95 (0.13)	0.04 (0.11)	0.01 (0.05)	0.0 (0.02)
1	4874	44.93 (19.45)	38.81 (17.09)	40.52 (17.63)	1016.01 (738.4)	502.31 (370.92)	513.69 (381.54)	0.5 (0.06)	0.04 (0.04)	0.0 (0.0)	50.91 (16.44)	8.31 (7.47)	18.84 (10.53)	12.14 (6.69)	5.64 (3.63)	11.67 (8.32)	17.98 (9.86)	7.98 (4.8)	2.89 (2.42)	0.17 (0.11)	0.41 (0.11)	0.27 (0.1)	0.15 (0.11)
2	26610	6.99 (6.73)	6.04 (5.88)	6.17 (6.02)	139.33 (229.57)	69.69 (120.39)	69.64 (115.59)	0.5 (0.14)	0.04 (0.08)	0.0 (0.0)	41.97 (36.97)	1.34 (1.98)	2.6 (3.31)	2.21 (2.05)	0.83 (1.19)	1.68 (2.39)	2.72 (3.06)	1.37 (1.54)	0.39 (0.76)	0.15 (0.17)	0.27 (0.21)	0.43 (0.31)	0.15 (0.25)
3	343	174.86 (64.42)	142.58 (49.02)	158.66 (64.15)	2533.92 (1880.3)	1268.38 (947.0)	1265.55 (956.25)	0.5 (0.06)	0.03 (0.02)	0.0 (0.0)	48.97 (13.9)	40.02 (33.6)	76.29 (31.56)	43.76 (18.85)	14.8 (11.21)	54.51 (42.47)	71.48 (29.34)	26.5 (14.5)	6.17 (6.41)	0.22 (0.11)	0.44 (0.08)	0.26 (0.09)	0.09 (0.05)
4	1704	2.75 (1.47)	2.44 (1.41)	2.21 (1.59)	125.1 (368.89)	64.28 (192.37)	60.82 (184.4)	0.46 (0.2)	0.83 (0.25)	0.0 (0.0)	44.31 (44.89)	0.55 (0.71)	0.92 (0.86)	0.71 (0.85)	0.57 (0.89)	0.49 (0.69)	0.83 (0.87)	0.56 (0.81)	0.32 (0.68)	0.22 (0.29)	0.35 (0.32)	0.25 (0.29)	0.18 (0.25)
5	769	78.56 (33.22)	71.38 (30.28)	72.99 (31.43)	4629.57 (2330.14)	2321.51 (1221.31)	2308.05 (1156.62)	0.5 (0.05)	0.04 (0.03)	0.0 (0.0)	49.31 (13.36)	9.75 (7.11)	26.88 (14.16)	23.2 (11.95)	18.73 (8.03)	14.51 (8.46)	29.07 (15.22)	17.72 (9.2)	11.7 (5.49)	0.12 (0.07)	0.33 (0.09)	0.29 (0.08)	0.26 (0.11)
6	25887	2.71 (2.88)	2.25 (2.5)	2.2 (2.6)	13.05 (26.34)	6.75 (13.86)	6.3 (13.17)	0.45 (0.27)	0.02 (0.06)	0.0 (0.0)	36.03 (46.88)	0.51 (0.97)	1.99 (1.86)	0.17 (0.45)	0.04 (0.2)	1.04 (1.51)	1.06 (1.42)	0.09 (0.3)	0.01 (0.12)	0.12 (0.19)	0.85 (0.21)	0.03 (0.08)	0.01 (0.04)
Clustering 3: 5 Clusters																							
0	28475	7.15 (7.2)	6.18 (6.27)	6.31 (6.45)	144.23 (236.39)	72.23 (123.96)	72.0 (119.04)	0.5 (0.14)	0.08 (0.2)	0.0 (0.0)	42.36 (37.25)	1.36 (2.07)	2.69 (3.55)	2.23 (2.18)	0.86 (1.23)	1.72 (2.55)	2.78 (3.25)	1.4 (1.61)	0.41 (0.78)	0.15 (0.17)	0.27 (0.22)	0.42 (0.31)	0.15 (0.25)
1	781	131.5 (60.69)	112.54 (45.81)	120.68 (57.45)	4014.73 (2679.2)	2009.72 (1377.3)	2005.01 (1336.56)	0.5 (0.05)	0.03 (0.03)	0.0 (0.0)	49.5 (12.73)	24.08 (26.7)	52.13 (31.09)	35.92 (16.35)	19.36 (10.03)	33.66 (33.84)	52.07 (27.75)	24.38 (11.54)	10.56 (7.1)	0.16 (0.1)	0.38 (0.09)	0.28 (0.08)	0.18 (0.11)
2	24104	1.65 (1.84)	1.24 (1.66)	0.54 (1.19)	3.29 (12.64)	2.19 (7.83)	1.1 (5.72)	0.24 (0.4)	0.01 (0.07)	0.0 (0.0)	33.89 (59.47)	1.42 (1.24)	0.17 (0.6)	0.04 (0.22)	0.01 (0.1)	0.45 (0.99)	0.07 (0.32)	0.02 (0.13)	0.0 (0.05)	0.95 (0.13)	0.04 (0.11)	0.01 (0.05)	0.0 (0.02)
3	4711	47.18 (20.05)	40.91 (17.57)	42.66 (18.18)	1284.63 (1108.09)	637.32 (564.34)	647.31 (559.05)	0.5 (0.06)	0.04 (0.04)	0.0 (0.0)	50.67 (16.28)	8.5 (7.72)	19.43 (10.99)	12.79 (6.82)	6.46 (4.18)	11.95 (8.72)	18.7 (10.15)	8.53 (4.97)	3.48 (2.91)	0.17 (0.1)	0.4 (0.11)	0.27 (0.1)	0.16 (0.12)
4	26042	2.74 (2.93)	2.27 (2.53)	2.22 (2.64)	13.16 (27.38)	6.79 (14.28)	6.38 (13.79)	0.45 (0.27)	0.02 (0.09)	0.0 (0.0)	35.98 (46.72)	0.52 (0.99)	2.0 (1.88)	0.18 (0.46)	0.04 (0.2)	1.05 (1.54)	1.07 (1.43)	0.09 (0.31)	0.01 (0.12)	0.12 (0.19)	0.85 (0.21)	0.03 (0.08)	0.01 (0.04)
The entire data set																							
0	84113	7.6 (18.05)	6.49 (15.4)	6.49 (16.61)	163.07 (614.17)	81.54 (310.03)	81.54 (308.01)	0.41 (0.29)	0.04 (0.14)	0.0 (0.0)	38.49 (46.88)	1.73 (4.46)	3.15 (7.83)	1.87 (5.11)	0.85 (2.79)	2.02 (5.83)	2.82 (7.54)	1.21 (3.51)	0.44 (1.65)	0.37 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.17)

Table C.3: *k*-means on data from the 3rd month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	1436	91.05 (28.17)	77.74 (24.96)	82.82 (26.13)	1767.0 (923.35)	879.95 (466.32)	887.05 (474.96)	0.5 (0.05)	0.04 (0.04)	0.0 (0.0)	58.22 (18.02)	17.26 (12.45)	38.42 (15.9)	24.92 (10.28)	10.45 (5.6)	24.56 (14.08)	37.12 (14.9)	16.11 (7.65)	5.03 (3.66)	0.18 (0.1)	0.42 (0.09)	0.28 (0.08)	0.13 (0.08)
1	25311	3.32 (3.87)	2.86 (3.4)	2.68 (3.38)	22.27 (52.44)	11.55 (27.23)	10.72 (26.57)	0.43 (0.25)	0.02 (0.07)	0.0 (0.0)	44.81 (62.54)	0.6 (1.21)	2.31 (2.28)	0.3 (0.66)	0.11 (0.38)	1.17 (1.71)	1.3 (1.81)	0.17 (0.48)	0.04 (0.23)	0.1 (0.17)	0.84 (0.21)	0.04 (0.09)	0.01 (0.05)
2	5851	33.32 (12.75)	29.06 (11.57)	29.75 (11.66)	836.19 (754.61)	415.22 (380.34)	420.97 (385.74)	0.5 (0.07)	0.05 (0.05)	0.0 (0.0)	58.29 (22.54)	6.13 (5.23)	13.93 (7.24)	8.78 (4.7)	4.48 (3.12)	8.52 (5.78)	12.95 (6.72)	5.9 (3.61)	2.38 (2.2)	0.18 (0.11)	0.41 (0.12)	0.27 (0.11)	0.15 (0.11)
3	1781	2.78 (1.55)	2.49 (1.54)	2.26 (1.58)	107.49 (387.31)	54.59 (200.21)	52.89 (192.91)	0.46 (0.2)	0.85 (0.23)	0.0 (0.0)	53.85 (64.93)	0.53 (0.72)	1.05 (0.97)	0.73 (0.88)	0.46 (0.81)	0.54 (0.74)	0.93 (0.97)	0.52 (0.75)	0.26 (0.62)	0.2 (0.28)	0.39 (0.33)	0.26 (0.29)	0.14 (0.23)
4	17153	1.49 (1.71)	1.46 (1.65)	0.09 (0.46)	2.37 (10.62)	2.15 (9.26)	0.22 (2.23)	0.01 (0.06)	0.01 (0.04)	0.0 (0.0)	47.09 (93.31)	1.32 (1.17)	0.14 (0.56)	0.02 (0.16)	0.0 (0.07)	0.06 (0.33)	0.03 (0.2)	0.0 (0.06)	0.0 (0.02)	0.96 (0.12)	0.04 (0.11)	0.0 (0.04)	0.0 (0.01)
5	17758	5.66 (5.18)	4.95 (4.58)	5.0 (4.67)	83.09 (115.32)	41.21 (59.6)	41.87 (58.69)	0.5 (0.15)	0.04 (0.09)	0.0 (0.0)	50.59 (53.02)	1.05 (1.5)	1.9 (2.36)	2.27 (1.73)	0.45 (0.8)	1.23 (1.7)	2.36 (2.47)	1.22 (1.35)	0.19 (0.48)	0.14 (0.17)	0.24 (0.2)	0.56 (0.29)	0.05 (0.09)
6	7923	2.61 (2.83)	0.94 (1.81)	2.27 (2.46)	9.95 (31.2)	3.74 (14.09)	6.21 (17.95)	0.85 (0.2)	0.01 (0.05)	0.0 (0.0)	25.29 (45.97)	1.86 (1.67)	0.58 (1.06)	0.13 (0.39)	0.05 (0.24)	1.75 (1.76)	0.41 (0.86)	0.09 (0.32)	0.02 (0.16)	0.83 (0.22)	0.14 (0.2)	0.02 (0.07)	0.01 (0.05)
7	137	258.26 (102.38)	199.58 (80.54)	241.7 (99.36)	3775.12 (2413.16)	1840.5 (1236.02)	1934.62 (1208.35)	0.52 (0.06)	0.03 (0.02)	0.0 (0.0)	51.42 (15.16)	62.4 (45.21)	110.76 (59.29)	62.9 (35.84)	22.2 (16.28)	87.55 (71.72)	105.76 (49.66)	38.71 (22.8)	9.69 (8.91)	0.24 (0.12)	0.43 (0.1)	0.25 (0.11)	0.09 (0.06)
8	3526	3.23 (3.03)	2.97 (2.8)	2.89 (2.79)	277.9 (481.27)	141.52 (253.05)	136.38 (240.71)	0.49 (0.13)	0.03 (0.08)	0.0 (0.0)	38.79 (49.32)	0.38 (0.71)	0.72 (1.09)	0.44 (0.82)	1.68 (1.22)	0.44 (0.78)	0.62 (1.05)	0.96 (1.08)	0.87 (1.0)	0.09 (0.16)	0.15 (0.2)	0.09 (0.15)	0.67 (0.27)
9	477	95.3 (45.8)	87.17 (41.8)	87.95 (43.21)	6336.5 (2591.65)	3181.38 (1372.73)	3155.12 (1281.47)	0.5 (0.04)	0.04 (0.03)	0.0 (0.0)	57.07 (16.32)	11.94 (9.82)	31.57 (18.35)	27.31 (15.72)	24.48 (10.33)	17.01 (11.08)	33.68 (19.93)	21.66 (12.84)	15.61 (6.13)	0.12 (0.07)	0.32 (0.08)	0.28 (0.07)	0.29 (0.1)
Clustering 2: 7 Clusters																							
0	31938	6.29 (6.69)	5.2 (5.98)	5.52 (5.94)	118.55 (231.16)	58.98 (119.36)	59.57 (116.61)	0.58 (0.21)	0.04 (0.08)	0.0 (0.0)	45.25 (50.45)	1.49 (2.0)	2.27 (3.25)	1.84 (2.09)	0.69 (1.15)	1.74 (2.36)	2.32 (3.02)	1.14 (1.5)	0.33 (0.73)	0.29 (0.34)	0.23 (0.21)	0.36 (0.33)	0.12 (0.23)
1	4648	48.5 (22.66)	41.83 (19.36)	43.62 (20.8)	1071.91 (771.54)	533.26 (392.04)	538.65 (392.96)	0.5 (0.06)	0.05 (0.05)	0.0 (0.0)	58.77 (21.01)	9.14 (8.26)	20.59 (12.14)	12.89 (7.3)	5.89 (3.68)	12.9 (9.97)	19.29 (11.11)	8.46 (5.1)	2.96 (2.46)	0.18 (0.11)	0.41 (0.11)	0.27 (0.1)	0.14 (0.11)
2	1873	2.85 (1.67)	2.56 (1.67)	2.34 (1.69)	115.25 (378.92)	58.96 (199.0)	56.29 (185.71)	0.46 (0.2)	0.83 (0.24)	0.0 (0.0)	53.77 (64.37)	0.53 (0.73)	1.07 (1.01)	0.74 (0.89)	0.52 (0.9)	0.55 (0.76)	0.94 (0.99)	0.55 (0.78)	0.3 (0.68)	0.2 (0.28)	0.39 (0.33)	0.26 (0.29)	0.16 (0.25)
3	17752	1.53 (1.79)	1.46 (1.71)	0.14 (0.53)	2.25 (7.95)	2.04 (7.53)	0.21 (0.98)	0.03 (0.11)	0.01 (0.04)	0.0 (0.0)	46.62 (92.04)	1.36 (1.24)	0.15 (0.61)	0.02 (0.14)	0.0 (0.05)	0.1 (0.44)	0.03 (0.19)	0.0 (0.04)	0.0 (0.0)	0.96 (0.13)	0.04 (0.12)	0.0 (0.04)	0.0 (0.02)
4	280	211.13 (86.49)	170.0 (66.87)	195.92 (84.48)	3576.06 (2561.89)	1775.37 (1314.23)	1800.69 (1274.14)	0.51 (0.06)	0.03 (0.03)	0.0 (0.0)	54.27 (16.63)	46.5 (37.31)	88.99 (47.52)	54.62 (28.26)	21.02 (15.68)	65.01 (55.87)	86.75 (40.72)	34.77 (19.66)	9.39 (8.88)	0.21 (0.12)	0.42 (0.09)	0.26 (0.09)	0.1 (0.07)
5	24021	2.6 (2.84)	2.22 (2.59)	2.1 (2.56)	13.03 (30.36)	6.92 (16.97)	6.11 (14.71)	0.44 (0.27)	0.02 (0.07)	0.0 (0.0)	42.58 (63.51)	0.43 (0.84)	1.97 (1.9)	0.17 (0.48)	0.04 (0.21)	0.97 (1.4)	1.03 (1.46)	0.08 (0.31)	0.01 (0.12)	0.11 (0.18)	0.86 (0.2)	0.03 (0.08)	0.01 (0.04)
6	841	87.02 (33.75)	78.71 (30.36)	80.06 (31.42)	4710.15 (2535.05)	2351.49 (1318.73)	2358.65 (1258.18)	0.5 (0.05)	0.04 (0.04)	0.0 (0.0)	57.1 (16.65)	11.78 (9.12)	30.57 (15.33)	25.04 (11.88)	19.63 (7.82)	16.85 (9.93)	32.06 (15.75)	18.94 (9.18)	12.21 (5.47)	0.13 (0.07)	0.34 (0.08)	0.28 (0.07)	0.25 (0.11)
Clustering 3: 5 Clusters																							
0	25435	2.88 (3.24)	2.4 (2.83)	2.34 (2.86)	14.48 (30.91)	7.5 (16.36)	6.98 (15.67)	0.45 (0.27)	0.03 (0.12)	0.0 (0.0)	42.8 (62.18)	0.55 (1.08)	2.08 (2.03)	0.21 (0.53)	0.05 (0.22)	1.09 (1.6)	1.13 (1.57)	0.1 (0.33)	0.02 (0.13)	0.12 (0.19)	0.84 (0.21)	0.03 (0.08)	0.01 (0.04)
1	23555	1.69 (2.04)	1.29 (1.86)	0.55 (1.23)	3.42 (12.45)	2.27 (7.61)	1.15 (5.86)	0.24 (0.39)	0.01 (0.07)	0.0 (0.0)	41.48 (84.48)	1.46 (1.37)	0.18 (0.67)	0.04 (0.22)	0.01 (0.1)	0.45 (1.01)	0.07 (0.34)	0.02 (0.13)	0.0 (0.06)	0.95 (0.13)	0.04 (0.12)	0.01 (0.05)	0.0 (0.03)
2	798	142.06 (75.66)	120.73 (57.56)	131.38 (72.36)	4614.23 (2896.74)	2301.1 (1497.69)	2313.13 (1434.52)	0.51 (0.05)	0.04 (0.03)	0.0 (0.0)	56.8 (15.9)	25.81 (27.65)	55.46 (39.25)	38.9 (22.26)	21.88 (11.39)	36.45 (40.02)	55.93 (35.21)	26.87 (14.84)	12.14 (7.5)	0.16 (0.1)	0.37 (0.09)	0.28 (0.08)	0.18 (0.1)
3	27159	7.36 (7.64)	6.39 (6.69)	6.49 (6.83)	153.88 (264.45)	76.87 (136.75)	77.01 (133.31)	0.5 (0.14)	0.09 (0.2)	0.0 (0.0)	50.21 (50.55)	1.4 (2.16)	2.75 (3.72)	2.31 (2.29)	0.9 (1.29)	1.76 (2.63)	2.86 (3.43)	1.44 (1.66)	0.43 (0.83)	0.15 (0.17)	0.27 (0.21)	0.43 (0.31)	0.15 (0.25)
4	4406	51.59 (22.83)	44.72 (19.73)	46.51 (20.87)	1359.36 (1133.53)	676.69 (575.64)	682.67 (572.82)	0.5 (0.06)	0.05 (0.04)	0.0 (0.0)	58.49 (20.7)	9.43 (8.59)	21.45 (12.27)	13.79 (7.43)	6.92 (4.45)	13.31 (10.09)	20.29 (11.21)	9.24 (5.36)	3.66 (3.1)	0.17 (0.1)	0.4 (0.11)	0.27 (0.1)	0.16 (0.12)
The entire data set																							
0	81353	8.04 (20.18)	6.86 (17.09)	6.86 (18.67)	175.77 (679.4)	87.89 (342.49)	87.89 (340.55)	0.41 (0.29)	0.04 (0.14)	0.0 (0.0)	45.88 (64.64)	1.82 (4.83)	3.33 (8.79)	1.98 (5.72)	0.91 (3.11)	2.14 (6.64)	2.98 (8.42)	1.28 (3.93)	0.47 (1.82)	0.37 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.17)

Table C.4: k-means on data from the 4th month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	23649	3.27 (3.34)	2.73 (2.91)	2.69 (2.98)	22.05 (49.37)	11.21 (25.63)	10.84 (24.94)	0.46 (0.26)	0.02 (0.07)	0.0 (0.0)	37.27 (63.12)	0.65 (1.14)	2.19 (1.89)	0.32 (0.68)	0.11 (0.36)	1.19 (1.55)	1.27 (1.62)	0.18 (0.49)	0.04 (0.22)	0.13 (0.19)	0.81 (0.22)	0.05 (0.09)	0.02 (0.05)
1	218	103.68 (55.34)	94.3 (48.48)	96.3 (52.26)	6484.11 (2626.68)	3266.89 (1357.45)	3217.22 (1324.39)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	48.25 (13.37)	13.31 (12.19)	34.57 (22.41)	29.6 (18.06)	26.2 (10.61)	18.75 (13.62)	37.77 (24.51)	22.85 (14.71)	16.93 (6.6)	0.12 (0.08)	0.32 (0.08)	0.28 (0.06)	0.28 (0.1)
2	3666	3.71 (3.34)	3.36 (3.04)	3.36 (3.13)	243.59 (362.03)	122.69 (187.68)	120.9 (183.99)	0.5 (0.13)	0.03 (0.08)	0.0 (0.0)	30.5 (37.38)	0.48 (0.83)	0.9 (1.24)	0.63 (0.98)	1.7 (1.21)	0.59 (0.97)	0.82 (1.18)	1.08 (1.14)	0.88 (1.01)	0.1 (0.16)	0.17 (0.2)	0.12 (0.16)	0.61 (0.28)
3	5501	27.07 (12.1)	23.15 (10.51)	24.1 (10.98)	445.86 (323.22)	221.8 (164.59)	224.06 (164.87)	0.5 (0.07)	0.05 (0.05)	0.0 (0.0)	49.4 (24.34)	5.52 (5.13)	11.79 (6.87)	7.06 (4.08)	2.7 (1.99)	7.63 (5.8)	10.84 (6.26)	4.37 (2.84)	1.26 (1.32)	0.19 (0.12)	0.43 (0.13)	0.27 (0.11)	0.11 (0.09)
4	12351	4.04 (3.55)	3.55 (3.16)	3.61 (3.27)	54.32 (75.83)	26.78 (37.7)	27.54 (39.54)	0.5 (0.16)	0.03 (0.08)	0.0 (0.0)	39.58 (48.81)	0.69 (1.06)	1.16 (1.49)	1.96 (1.4)	0.23 (0.53)	0.8 (1.17)	1.75 (1.82)	0.96 (1.11)	0.09 (0.32)	0.13 (0.18)	0.2 (0.2)	0.64 (0.28)	0.03 (0.07)
5	1549	47.55 (18.39)	42.82 (16.44)	43.82 (17.18)	1977.69 (1036.43)	982.72 (528.47)	994.97 (531.75)	0.5 (0.05)	0.04 (0.04)	0.0 (0.0)	47.06 (19.73)	6.63 (4.87)	17.34 (9.39)	13.92 (6.88)	9.67 (3.98)	9.89 (6.28)	17.87 (9.23)	10.31 (5.23)	5.76 (2.87)	0.13 (0.08)	0.35 (0.1)	0.29 (0.09)	0.23 (0.12)
6	1370	2.81 (1.46)	2.51 (1.42)	2.31 (1.61)	93.33 (255.62)	46.34 (132.5)	46.99 (126.45)	0.47 (0.2)	0.83 (0.25)	0.0 (0.0)	46.96 (70.87)	0.55 (0.72)	0.99 (0.91)	0.78 (0.88)	0.49 (0.82)	0.53 (0.73)	0.89 (0.97)	0.6 (0.82)	0.28 (0.64)	0.21 (0.29)	0.37 (0.32)	0.27 (0.29)	0.15 (0.23)
7	459	122.85 (46.16)	100.97 (38.52)	113.44 (43.96)	1733.96 (918.99)	863.35 (474.3)	870.61 (463.96)	0.51 (0.06)	0.03 (0.03)	0.0 (0.0)	46.33 (14.45)	25.34 (20.37)	53.5 (23.4)	33.75 (17.14)	10.27 (6.49)	35.67 (22.85)	53.51 (26.25)	19.97 (11.53)	4.28 (3.62)	0.2 (0.11)	0.44 (0.09)	0.28 (0.1)	0.09 (0.05)
8	19716	1.68 (1.83)	1.26 (1.6)	0.57 (1.27)	3.64 (13.08)	2.31 (7.48)	1.33 (6.17)	0.24 (0.4)	0.01 (0.04)	0.0 (0.0)	33.15 (79.06)	1.44 (1.23)	0.17 (0.57)	0.05 (0.24)	0.01 (0.12)	0.46 (1.01)	0.08 (0.34)	0.02 (0.17)	0.0 (0.06)	0.95 (0.13)	0.04 (0.11)	0.01 (0.05)	0.0 (0.02)
9	3	880.33 (174.27)	505.0 (195.67)	848.67 (138.0)	2761.33 (1033.72)	1296.0 (714.49)	1465.33 (333.31)	0.56 (0.08)	0.01 (0.0)	0.0 (0.0)	25.58 (4.08)	378.33 (42.03)	450.67 (156.42)	47.0 (14.76)	4.33 (4.78)	625.33 (143.79)	210.33 (14.06)	10.67 (5.44)	2.33 (2.05)	0.44 (0.09)	0.5 (0.1)	0.05 (0.01)	0.0 (0.0)
Clustering 2: 7 Clusters																							
0	21574	2.76 (2.94)	2.31 (2.57)	2.24 (2.63)	13.64 (27.32)	7.06 (14.32)	6.58 (13.74)	0.45 (0.26)	0.02 (0.06)	0.0 (0.0)	36.08 (64.28)	0.52 (0.98)	2.01 (1.89)	0.19 (0.47)	0.04 (0.21)	1.05 (1.5)	1.08 (1.46)	0.09 (0.31)	0.02 (0.13)	0.12 (0.19)	0.84 (0.21)	0.03 (0.08)	0.01 (0.04)
1	3967	40.96 (17.77)	35.46 (15.21)	37.19 (16.39)	1066.61 (870.51)	529.87 (439.31)	536.73 (443.08)	0.5 (0.06)	0.04 (0.04)	0.0 (0.0)	48.22 (20.91)	7.4 (6.8)	16.82 (9.84)	11.18 (6.03)	5.56 (3.6)	10.54 (8.12)	16.29 (9.16)	7.34 (4.34)	3.02 (2.54)	0.17 (0.1)	0.4 (0.12)	0.28 (0.1)	0.16 (0.12)
2	708	110.4 (52.65)	95.45 (42.96)	102.37 (49.51)	3645.05 (2530.21)	1826.48 (1294.35)	1818.57 (1260.58)	0.5 (0.05)	0.04 (0.03)	0.0 (0.0)	46.79 (14.02)	18.39 (18.65)	42.9 (25.96)	31.6 (17.38)	17.51 (9.61)	26.25 (21.66)	44.72 (27.65)	21.46 (12.07)	9.94 (6.8)	0.15 (0.09)	0.37 (0.09)	0.29 (0.08)	0.19 (0.11)
3	19745	1.67 (1.9)	1.25 (1.67)	0.55 (1.26)	3.28 (10.83)	2.17 (6.75)	1.11 (5.03)	0.24 (0.4)	0.01 (0.04)	0.0 (0.0)	33.06 (79.08)	1.44 (1.3)	0.18 (0.61)	0.04 (0.21)	0.01 (0.1)	0.46 (1.04)	0.07 (0.32)	0.02 (0.13)	0.0 (0.05)	0.95 (0.13)	0.04 (0.12)	0.01 (0.05)	0.0 (0.02)
4	1431	2.87 (1.55)	2.56 (1.52)	2.36 (1.67)	105.49 (293.82)	52.88 (155.06)	52.61 (143.78)	0.47 (0.2)	0.81 (0.26)	0.0 (0.0)	46.5 (69.74)	0.54 (0.72)	0.99 (0.91)	0.78 (0.89)	0.56 (0.95)	0.53 (0.72)	0.89 (0.97)	0.62 (0.84)	0.32 (0.73)	0.21 (0.29)	0.36 (0.32)	0.27 (0.29)	0.16 (0.25)
5	21054	6.75 (6.65)	5.84 (5.81)	5.98 (5.92)	127.7 (200.18)	63.62 (102.63)	64.08 (101.88)	0.5 (0.14)	0.04 (0.08)	0.0 (0.0)	40.59 (44.62)	1.3 (2.0)	2.5 (3.29)	2.17 (1.99)	0.78 (1.11)	1.66 (2.45)	2.63 (2.99)	1.33 (1.48)	0.36 (0.7)	0.14 (0.17)	0.26 (0.21)	0.45 (0.32)	0.15 (0.25)
6	3	880.33 (174.27)	505.0 (195.67)	848.67 (138.0)	2761.33 (1033.72)	1296.0 (714.49)	1465.33 (333.31)	0.56 (0.08)	0.01 (0.0)	0.0 (0.0)	25.58 (4.08)	378.33 (42.03)	450.67 (156.42)	47.0 (14.76)	4.33 (4.78)	625.33 (143.79)	210.33 (14.06)	10.67 (5.44)	2.33 (2.05)	0.44 (0.09)	0.5 (0.1)	0.05 (0.01)	0.0 (0.0)
Clustering 3: 5 Clusters																							
0	38377	4.56 (5.23)	3.87 (4.56)	3.9 (4.66)	47.65 (92.91)	23.79 (47.06)	23.85 (47.22)	0.47 (0.22)	0.02 (0.06)	0.0 (0.0)	38.6 (56.8)	0.92 (1.62)	2.29 (2.62)	1.11 (1.64)	0.25 (0.63)	1.37 (2.02)	1.83 (2.38)	0.6 (1.1)	0.1 (0.36)	0.13 (0.18)	0.59 (0.35)	0.24 (0.32)	0.03 (0.07)
1	4406	39.19 (17.48)	33.96 (15.04)	35.56 (16.1)	1000.92 (849.54)	497.38 (428.91)	503.54 (431.82)	0.5 (0.06)	0.04 (0.04)	0.0 (0.0)	48.31 (21.13)	7.07 (6.52)	16.08 (9.58)	10.76 (5.89)	5.28 (3.56)	10.09 (7.78)	15.57 (8.93)	7.06 (4.25)	2.85 (2.49)	0.17 (0.1)	0.4 (0.12)	0.28 (0.1)	0.16 (0.11)
2	19592	1.6 (1.73)	1.22 (1.57)	0.5 (1.12)	3.21 (12.25)	2.13 (7.2)	1.08 (5.69)	0.24 (0.4)	0.0 (0.04)	0.0 (0.0)	32.9 (79.22)	1.4 (1.24)	0.14 (0.51)	0.04 (0.22)	0.01 (0.1)	0.42 (0.96)	0.06 (0.27)	0.02 (0.14)	0.0 (0.05)	0.95 (0.13)	0.04 (0.11)	0.01 (0.06)	0.0 (0.02)
3	714	113.8 (73.11)	97.18 (51.86)	105.68 (69.51)	3620.8 (2530.49)	1813.62 (1294.94)	1807.18 (1260.22)	0.5 (0.05)	0.04 (0.03)	0.0 (0.0)	46.7 (14.04)	20.05 (29.95)	44.72 (38.28)	31.65 (17.37)	17.38 (9.66)	28.93 (45.36)	45.52 (29.51)	21.38 (12.06)	9.85 (6.82)	0.15 (0.1)	0.37 (0.09)	0.29 (0.08)	0.19 (0.11)
4	5393	4.02 (3.78)	3.64 (3.49)	3.56 (3.52)	209.92 (346.58)	105.35 (178.51)	104.57 (176.19)	0.49 (0.15)	0.28 (0.36)	0.0 (0.0)	37.9 (52.16)	0.59 (0.94)	1.11 (1.45)	0.93 (1.33)	1.4 (1.33)	0.67 (1.04)	1.07 (1.49)	1.1 (1.31)	0.72 (0.97)	0.13 (0.21)	0.24 (0.26)	0.18 (0.23)	0.45 (0.35)
The entire data set																							
0	68482	7.04 (17.05)	6.0 (14.23)	6.0 (15.87)	146.3 (558.59)	73.15 (281.31)	73.15 (280.09)	0.41 (0.29)	0.04 (0.13)	0.0 (0.0)	37.62 (62.16)	1.62 (4.46)	2.91 (7.57)	1.73 (4.76)	0.77 (2.58)	1.89 (6.39)	2.6 (7.07)	1.1 (3.25)	0.4 (1.57)	0.37 (0.4)	0.39 (0.37)	0.18 (0.27)	0.06 (0.17)

Table C.5: *k*-means on data from the 5th month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	8031	2.58 (2.76)	0.94 (1.77)	2.23 (2.4)	9.75 (29.01)	3.72 (13.67)	6.03 (15.87)	0.85 (0.2)	0.01 (0.05)	0.0 (0.0)	25.77 (48.08)	1.84 (1.61)	0.55 (0.99)	0.14 (0.42)	0.05 (0.24)	1.71 (1.67)	0.4 (0.86)	0.1 (0.35)	0.02 (0.16)	0.84 (0.21)	0.13 (0.19)	0.02 (0.07)	0.01 (0.04)
1	6724	29.7 (11.81)	25.71 (10.63)	26.38 (10.82)	617.57 (577.2)	307.25 (294.08)	310.33 (294.79)	0.5 (0.07)	0.05 (0.05)	0.0 (0.0)	59.99 (23.47)	5.79 (4.98)	12.63 (6.78)	7.79 (4.37)	3.49 (2.61)	7.9 (5.56)	11.68 (6.23)	5.06 (3.23)	1.74 (1.78)	0.19 (0.12)	0.42 (0.12)	0.26 (0.11)	0.13 (0.1)
2	414	108.09 (51.12)	98.52 (45.95)	100.22 (48.23)	7027.98 (2997.09)	3509.58 (1570.31)	3518.4 (1493.27)	0.5 (0.04)	0.04 (0.03)	0.0 (0.0)	59.11 (15.67)	13.59 (10.79)	36.08 (20.03)	31.35 (17.47)	27.07 (12.13)	19.91 (12.6)	38.66 (21.53)	24.25 (13.69)	17.41 (7.91)	0.12 (0.07)	0.32 (0.08)	0.28 (0.07)	0.27 (0.1)
3	27121	3.37 (3.69)	2.91 (3.25)	2.73 (3.26)	23.03 (52.15)	11.89 (26.97)	11.14 (26.49)	0.44 (0.25)	0.02 (0.07)	0.0 (0.0)	49.84 (70.26)	0.62 (1.19)	2.3 (2.12)	0.33 (0.7)	0.12 (0.39)	1.17 (1.64)	1.32 (1.76)	0.19 (0.52)	0.04 (0.22)	0.11 (0.17)	0.83 (0.21)	0.05 (0.09)	0.02 (0.06)
4	18452	1.47 (1.61)	1.45 (1.56)	0.09 (0.46)	2.2 (6.82)	1.99 (5.51)	0.21 (1.69)	0.01 (0.06)	0.01 (0.04)	0.0 (0.0)	51.58 (92.06)	1.32 (1.12)	0.13 (0.53)	0.02 (0.14)	0.0 (0.06)	0.05 (0.32)	0.03 (0.2)	0.0 (0.07)	0.0 (0.01)	0.96 (0.12)	0.04 (0.11)	0.0 (0.03)	0.0 (0.01)
5	1782	2.62 (1.23)	2.35 (1.19)	2.11 (1.35)	91.45 (352.05)	45.49 (180.53)	45.96 (176.01)	0.46 (0.2)	0.87 (0.21)	0.0 (0.0)	62.51 (71.58)	0.56 (0.72)	0.99 (0.88)	0.68 (0.78)	0.38 (0.69)	0.52 (0.72)	0.88 (0.89)	0.51 (0.73)	0.2 (0.52)	0.22 (0.29)	0.38 (0.32)	0.26 (0.29)	0.13 (0.23)
6	16540	4.8 (4.2)	4.21 (3.71)	4.25 (3.85)	67.7 (93.11)	33.46 (46.42)	34.24 (48.26)	0.49 (0.15)	0.04 (0.09)	0.0 (0.0)	52.78 (54.82)	0.87 (1.25)	1.48 (1.82)	2.11 (1.55)	0.34 (0.67)	0.99 (1.38)	2.02 (2.08)	1.1 (1.24)	0.14 (0.4)	0.14 (0.18)	0.22 (0.2)	0.59 (0.28)	0.05 (0.09)
7	219	231.53 (95.47)	179.22 (76.18)	216.46 (89.56)	3092.41 (1879.17)	1533.31 (982.11)	1559.1 (925.92)	0.51 (0.06)	0.02 (0.02)	0.0 (0.0)	50.77 (16.01)	57.3 (48.1)	100.82 (47.93)	55.04 (27.85)	18.37 (12.86)	79.92 (63.48)	95.96 (46.07)	33.02 (19.23)	7.57 (6.74)	0.24 (0.14)	0.44 (0.1)	0.24 (0.09)	0.08 (0.05)
8	2030	76.49 (26.95)	66.58 (23.04)	69.61 (25.28)	1983.84 (1147.2)	991.5 (595.83)	992.34 (578.94)	0.5 (0.06)	0.04 (0.03)	0.0 (0.0)	60.0 (18.06)	13.39 (10.79)	31.23 (15.37)	21.17 (9.34)	10.7 (5.39)	19.25 (13.05)	30.35 (14.07)	14.38 (6.88)	5.63 (3.73)	0.16 (0.09)	0.4 (0.1)	0.28 (0.09)	0.16 (0.1)
9	3655	3.27 (2.99)	2.98 (2.77)	2.94 (2.78)	272.67 (533.87)	137.81 (278.22)	134.85 (267.65)	0.5 (0.13)	0.04 (0.09)	0.0 (0.0)	40.73 (50.12)	0.41 (0.73)	0.72 (1.08)	0.46 (0.84)	1.67 (1.18)	0.46 (0.8)	0.64 (1.05)	1.0 (1.08)	0.84 (0.99)	0.1 (0.17)	0.15 (0.19)	0.09 (0.16)	0.66 (0.27)
Clustering 2: 7 Clusters																							
0	26782	3.11 (3.66)	2.6 (3.17)	2.53 (3.23)	16.51 (37.8)	8.49 (19.16)	8.02 (19.73)	0.45 (0.26)	0.02 (0.06)	0.0 (0.0)	47.71 (70.11)	0.61 (1.21)	2.2 (2.23)	0.24 (0.6)	0.06 (0.25)	1.17 (1.76)	1.22 (1.72)	0.12 (0.38)	0.02 (0.15)	0.12 (0.19)	0.83 (0.21)	0.04 (0.08)	0.01 (0.04)
1	294	218.52 (87.91)	177.52 (66.73)	203.87 (83.0)	5058.94 (3819.52)	2533.57 (1960.86)	2525.37 (1889.32)	0.51 (0.05)	0.03 (0.02)	0.0 (0.0)	53.36 (14.73)	46.8 (44.66)	89.87 (45.44)	55.69 (24.83)	26.16 (16.46)	66.17 (58.9)	88.14 (42.31)	36.2 (18.07)	13.36 (11.04)	0.2 (0.12)	0.41 (0.09)	0.26 (0.08)	0.13 (0.09)
2	1414	87.01 (32.17)	76.76 (27.21)	80.04 (30.36)	3453.47 (2219.01)	1723.72 (1137.32)	1729.74 (1118.88)	0.5 (0.05)	0.04 (0.03)	0.0 (0.0)	59.81 (17.63)	13.6 (11.35)	33.2 (17.59)	24.6 (10.95)	15.63 (7.18)	19.71 (13.95)	33.61 (16.78)	17.64 (8.01)	9.09 (5.43)	0.15 (0.08)	0.37 (0.1)	0.28 (0.08)	0.21 (0.11)
3	24848	1.66 (1.94)	1.28 (1.74)	0.51 (1.23)	3.15 (11.39)	2.14 (6.91)	1.01 (5.17)	0.23 (0.39)	0.01 (0.04)	0.0 (0.0)	45.11 (84.41)	1.44 (1.31)	0.18 (0.62)	0.04 (0.21)	0.01 (0.09)	0.43 (0.99)	0.07 (0.34)	0.01 (0.13)	0.0 (0.05)	0.95 (0.13)	0.04 (0.12)	0.01 (0.05)	0.0 (0.02)
4	5373	40.39 (17.67)	34.93 (15.28)	36.1 (16.1)	879.91 (730.83)	438.26 (371.76)	441.65 (372.76)	0.5 (0.06)	0.04 (0.04)	0.0 (0.0)	59.78 (21.65)	7.77 (7.23)	17.15 (9.85)	10.68 (5.88)	4.78 (3.23)	10.8 (8.48)	15.98 (8.68)	6.92 (4.2)	2.41 (2.17)	0.18 (0.11)	0.41 (0.12)	0.27 (0.11)	0.14 (0.11)
5	24078	6.27 (5.92)	5.43 (5.19)	5.55 (5.31)	123.67 (205.13)	61.51 (107.12)	62.16 (103.45)	0.5 (0.14)	0.04 (0.08)	0.0 (0.0)	51.82 (51.04)	1.2 (1.79)	2.19 (2.78)	2.12 (1.91)	0.77 (1.09)	1.45 (2.04)	2.43 (2.73)	1.32 (1.45)	0.35 (0.7)	0.15 (0.18)	0.25 (0.21)	0.45 (0.32)	0.15 (0.25)
6	2179	2.73 (1.41)	2.43 (1.34)	2.2 (1.48)	88.53 (337.57)	44.12 (173.64)	44.41 (168.28)	0.47 (0.2)	0.8 (0.24)	0.0 (0.0)	62.74 (72.55)	0.6 (0.74)	1.03 (0.92)	0.7 (0.81)	0.39 (0.74)	0.57 (0.78)	0.92 (0.95)	0.51 (0.75)	0.2 (0.54)	0.23 (0.28)	0.38 (0.31)	0.26 (0.29)	0.13 (0.23)
Clustering 3: 5 Clusters																							
0	25053	1.66 (1.94)	1.29 (1.77)	0.51 (1.17)	3.13 (10.36)	2.13 (6.51)	0.99 (4.61)	0.23 (0.39)	0.01 (0.07)	0.0 (0.0)	45.33 (84.53)	1.44 (1.33)	0.18 (0.62)	0.04 (0.21)	0.01 (0.09)	0.42 (0.96)	0.07 (0.33)	0.01 (0.12)	0.0 (0.04)	0.95 (0.13)	0.04 (0.12)	0.01 (0.05)	0.0 (0.02)
1	695	158.49 (79.68)	132.91 (60.5)	147.55 (75.02)	4909.24 (3271.02)	2446.0 (1681.17)	2463.24 (1624.84)	0.51 (0.05)	0.03 (0.02)	0.0 (0.0)	57.04 (15.84)	29.96 (33.89)	62.63 (40.25)	42.31 (21.35)	23.58 (12.54)	42.61 (44.84)	62.78 (37.5)	29.1 (14.6)	13.06 (8.78)	0.17 (0.1)	0.38 (0.1)	0.28 (0.08)	0.18 (0.1)
2	26148	2.78 (3.03)	2.33 (2.65)	2.24 (2.7)	13.61 (29.33)	7.02 (14.95)	6.58 (15.52)	0.45 (0.27)	0.02 (0.11)	0.0 (0.0)	47.29 (71.14)	0.53 (1.02)	2.03 (1.94)	0.19 (0.48)	0.04 (0.21)	1.05 (1.54)	1.09 (1.48)	0.09 (0.31)	0.02 (0.13)	0.12 (0.19)	0.84 (0.21)	0.03 (0.08)	0.01 (0.04)
3	4131	55.63 (23.78)	48.52 (20.6)	50.32 (22.0)	1565.99 (1331.17)	781.89 (680.59)	784.1 (671.91)	0.5 (0.06)	0.04 (0.04)	0.0 (0.0)	59.86 (19.66)	9.88 (8.91)	22.78 (12.83)	15.21 (7.91)	7.76 (4.96)	14.1 (10.78)	21.87 (11.63)	10.2 (5.72)	4.15 (3.46)	0.17 (0.1)	0.39 (0.11)	0.28 (0.1)	0.16 (0.12)
4	28941	8.0 (8.44)	6.91 (7.39)	7.05 (7.5)	154.6 (263.32)	76.97 (135.55)	77.63 (132.81)	0.5 (0.14)	0.09 (0.2)	0.0 (0.0)	53.39 (50.27)	1.58 (2.5)	3.06 (4.13)	2.43 (2.45)	0.93 (1.34)	1.99 (2.96)	3.11 (3.75)	1.53 (1.77)	0.43 (0.84)	0.16 (0.18)	0.28 (0.21)	0.42 (0.31)	0.15 (0.24)
The entire data set																							
0	84968	8.07 (20.58)	6.9 (17.38)	6.9 (19.07)	174.06 (701.32)	87.03 (353.69)	87.03 (351.63)	0.4 (0.29)	0.04 (0.14)	0.0 (0.0)	49.48 (67.55)	1.85 (5.14)	3.34 (8.89)	1.98 (5.73)	0.9 (3.17)	2.16 (6.9)	2.99 (8.59)	1.29 (3.95)	0.46 (1.88)	0.38 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)

Table C.6: GMM on data from the 1st month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	15417	1.24 (0.61)	1.11 (0.66)	0.93 (0.81)	4.11 (2.81)	2.34 (1.76)	1.78 (1.86)	0.4 (0.31)	0.0 (0.0)	0.0 (0.0)	27.52 (46.59)	0.0 (0.0)	1.24 (0.61)	0.0 (0.0)	0.0 (0.0)	0.41 (0.61)	0.51 (0.66)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)
1	5433	18.84 (6.95)	16.28 (6.36)	17.1 (6.65)	301.71 (148.64)	148.28 (75.97)	153.43 (76.61)	0.51 (0.06)	0.04 (0.04)	0.0 (0.0)	43.13 (18.24)	3.36 (2.55)	8.06 (4.34)	5.19 (3.01)	2.23 (1.43)	4.94 (3.33)	7.64 (4.04)	3.41 (2.32)	1.1 (0.92)	0.18 (0.12)	0.42 (0.14)	0.28 (0.13)	0.13 (0.08)
2	23846	2.58 (1.73)	1.86 (1.65)	2.11 (1.59)	14.08 (13.49)	6.96 (7.18)	7.13 (7.2)	0.59 (0.27)	0.0 (0.0)	0.0 (0.0)	29.46 (37.65)	0.98 (0.95)	0.93 (1.16)	0.67 (0.77)	0.0 (0.0)	0.99 (1.06)	0.86 (1.06)	0.26 (0.51)	0.0 (0.0)	0.41 (0.37)	0.26 (0.27)	0.32 (0.38)	0.0 (0.0)
3	17084	1.15 (0.44)	1.15 (0.44)	0.0 (0.0)	1.2 (0.6)	1.2 (0.6)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	33.27 (68.96)	1.11 (0.41)	0.04 (0.18)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.98 (0.09)	0.02 (0.09)	0.0 (0.0)	0.0 (0.0)
4	3023	6.6 (3.77)	5.97 (3.54)	5.89 (3.67)	539.65 (728.87)	271.46 (369.81)	268.19 (369.23)	0.5 (0.12)	0.22 (0.27)	0.0 (0.0)	35.89 (29.07)	0.95 (1.2)	1.92 (1.67)	1.55 (1.48)	2.18 (1.58)	1.17 (1.25)	1.85 (1.68)	1.42 (1.35)	1.45 (1.26)	0.13 (0.17)	0.27 (0.2)	0.22 (0.19)	0.37 (0.21)
5	4734	3.14 (1.91)	2.76 (1.71)	2.86 (1.82)	111.63 (95.55)	54.8 (47.67)	56.83 (50.09)	0.51 (0.11)	0.0 (0.0)	0.0 (0.0)	30.33 (33.42)	0.5 (0.75)	0.94 (1.12)	0.56 (0.82)	1.14 (0.4)	0.63 (0.9)	0.84 (1.05)	0.87 (0.79)	0.52 (0.57)	0.12 (0.19)	0.23 (0.24)	0.13 (0.19)	0.51 (0.31)
6	1528	90.27 (53.71)	74.71 (44.06)	79.62 (53.65)	1931.1 (2077.15)	966.53 (1040.98)	964.57 (1060.25)	0.48 (0.13)	0.03 (0.02)	0.0 (0.0)	44.27 (13.96)	20.63 (18.79)	37.8 (26.31)	22.2 (17.05)	9.63 (9.51)	25.85 (23.77)	34.97 (27.39)	13.84 (11.58)	4.96 (6.05)	0.25 (0.16)	0.41 (0.11)	0.23 (0.11)	0.11 (0.09)
7	4118	39.21 (14.04)	34.66 (12.94)	35.93 (13.53)	1214.48 (970.17)	605.8 (494.88)	608.67 (485.55)	0.5 (0.06)	0.04 (0.03)	0.0 (0.0)	44.21 (14.72)	6.08 (3.97)	15.55 (7.58)	11.13 (5.59)	6.45 (4.25)	9.09 (5.26)	15.53 (7.51)	7.67 (4.37)	3.65 (2.98)	0.15 (0.09)	0.39 (0.11)	0.28 (0.09)	0.18 (0.11)
8	4375	3.62 (1.26)	3.11 (1.31)	2.79 (1.47)	24.65 (20.32)	12.61 (10.54)	12.04 (10.75)	0.46 (0.2)	0.41 (0.33)	0.0 (0.0)	42.49 (40.24)	0.99 (1.04)	1.64 (1.11)	0.89 (0.93)	0.1 (0.3)	0.99 (1.0)	1.33 (1.1)	0.47 (0.69)	0.0 (0.0)	0.27 (0.28)	0.46 (0.29)	0.25 (0.26)	0.02 (0.08)
9	7081	10.68 (3.94)	8.94 (3.8)	9.0 (3.95)	78.59 (43.85)	39.94 (23.27)	38.65 (22.76)	0.49 (0.12)	0.05 (0.05)	0.0 (0.0)	42.44 (22.95)	2.65 (2.16)	5.05 (2.76)	2.57 (1.85)	0.41 (0.61)	3.33 (2.42)	4.22 (2.56)	1.45 (1.33)	0.0 (0.0)	0.25 (0.18)	0.47 (0.18)	0.25 (0.17)	0.04 (0.07)
Clustering 2: 7 Clusters																							
0	22552	1.81 (1.2)	1.42 (1.02)	1.27 (1.18)	4.87 (3.38)	2.68 (2.08)	2.19 (2.19)	0.42 (0.3)	0.0 (0.0)	0.0 (0.0)	28.7 (43.36)	0.48 (0.84)	1.33 (0.75)	0.0 (0.0)	0.0 (0.0)	0.71 (0.96)	0.56 (0.72)	0.0 (0.0)	0.0 (0.0)	0.16 (0.25)	0.84 (0.25)	0.0 (0.0)	0.0 (0.0)
1	9297	19.14 (9.19)	16.54 (8.25)	17.31 (8.75)	310.07 (221.49)	153.45 (111.57)	156.62 (112.9)	0.51 (0.07)	0.05 (0.05)	0.0 (0.0)	42.96 (18.75)	3.46 (2.74)	8.16 (4.96)	5.27 (3.56)	2.24 (1.73)	5.01 (3.61)	7.75 (4.83)	3.53 (2.56)	1.02 (1.17)	0.18 (0.12)	0.42 (0.14)	0.27 (0.13)	0.13 (0.08)
2	21213	1.13 (0.41)	0.89 (0.58)	0.24 (0.48)	1.13 (0.41)	0.89 (0.58)	0.24 (0.48)	0.21 (0.4)	0.0 (0.0)	0.0 (0.0)	29.59 (64.35)	1.13 (0.41)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.24 (0.48)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
3	19637	4.57 (3.7)	3.83 (3.18)	3.8 (3.31)	29.81 (22.98)	15.01 (11.93)	14.8 (12.27)	0.49 (0.17)	0.05 (0.1)	0.0 (0.0)	37.27 (36.47)	1.14 (1.57)	1.97 (2.28)	1.45 (1.14)	0.0 (0.0)	1.35 (1.8)	1.88 (1.9)	0.57 (0.76)	0.0 (0.0)	0.2 (0.23)	0.33 (0.27)	0.48 (0.33)	0.0 (0.0)
4	5117	3.66 (2.45)	3.17 (2.13)	3.34 (2.33)	104.19 (72.84)	51.25 (36.74)	52.95 (38.45)	0.51 (0.11)	0.0 (0.0)	0.0 (0.0)	30.94 (31.26)	0.63 (0.92)	1.18 (1.37)	0.71 (1.0)	1.15 (0.41)	0.81 (1.12)	1.07 (1.32)	0.98 (0.89)	0.48 (0.56)	0.13 (0.19)	0.24 (0.24)	0.14 (0.19)	0.48 (0.31)
5	4399	4.74 (3.01)	4.27 (2.83)	4.13 (2.99)	321.87 (564.7)	162.49 (288.46)	159.38 (282.46)	0.48 (0.16)	0.4 (0.38)	0.0 (0.0)	36.15 (34.41)	0.74 (0.95)	1.47 (1.29)	1.17 (1.18)	1.36 (1.42)	0.86 (1.02)	1.39 (1.31)	1.06 (1.14)	0.82 (1.12)	0.16 (0.22)	0.32 (0.26)	0.24 (0.24)	0.27 (0.25)
6	4424	57.94 (41.45)	49.62 (33.83)	51.89 (39.54)	1646.45 (1518.63)	822.75 (765.81)	823.7 (771.03)	0.49 (0.1)	0.03 (0.03)	0.0 (0.0)	43.86 (14.55)	11.21 (13.45)	23.46 (19.86)	15.14 (12.36)	8.13 (6.81)	14.92 (16.74)	22.43 (19.73)	10.04 (8.33)	4.5 (4.43)	0.19 (0.14)	0.39 (0.12)	0.26 (0.1)	0.17 (0.12)
Clustering 3: 5 Clusters																							
0	43618	1.47 (0.94)	1.16 (0.88)	0.75 (0.99)	3.0 (2.92)	1.8 (1.74)	1.21 (1.77)	0.32 (0.37)	0.0 (0.0)	0.0 (0.0)	29.14 (54.61)	0.8 (0.75)	0.67 (0.82)	0.0 (0.0)	0.0 (0.0)	0.47 (0.77)	0.28 (0.56)	0.0 (0.0)	0.0 (0.0)	0.57 (0.46)	0.43 (0.46)	0.0 (0.0)	0.0 (0.0)
1	7959	3.31 (1.87)	2.95 (1.78)	2.8 (1.88)	139.94 (241.96)	70.07 (123.92)	69.87 (121.19)	0.48 (0.17)	0.28 (0.34)	0.0 (0.0)	35.3 (38.08)	0.62 (0.86)	1.11 (1.09)	0.73 (0.92)	0.85 (0.9)	0.66 (0.85)	0.97 (1.05)	0.64 (0.83)	0.53 (0.73)	0.17 (0.24)	0.3 (0.29)	0.19 (0.24)	0.33 (0.36)
2	7351	44.68 (36.73)	38.42 (30.28)	40.28 (34.46)	1269.15 (1309.65)	633.47 (660.99)	635.68 (662.81)	0.5 (0.09)	0.04 (0.04)	0.0 (0.0)	43.12 (15.98)	8.32 (11.2)	18.04 (17.23)	11.89 (10.75)	6.44 (5.82)	11.37 (13.97)	17.37 (16.96)	7.98 (7.22)	3.57 (3.74)	0.17 (0.12)	0.38 (0.13)	0.26 (0.11)	0.18 (0.13)
3	14900	3.1 (2.11)	2.61 (1.79)	2.73 (2.0)	28.79 (20.74)	14.37 (10.9)	14.42 (11.06)	0.5 (0.16)	0.0 (0.0)	0.0 (0.0)	33.63 (37.26)	0.63 (0.94)	1.12 (1.4)	1.22 (0.74)	0.13 (0.36)	0.79 (1.15)	1.31 (1.25)	0.63 (0.7)	0.0 (0.0)	0.15 (0.21)	0.26 (0.26)	0.55 (0.34)	0.04 (0.12)
4	12811	12.75 (6.64)	10.85 (6.04)	11.12 (6.43)	142.97 (112.84)	71.09 (56.66)	71.89 (58.18)	0.5 (0.11)	0.06 (0.06)	0.0 (0.0)	42.76 (22.38)	2.76 (2.28)	5.73 (3.75)	3.27 (2.59)	0.99 (1.14)	3.66 (2.85)	5.1 (3.6)	1.97 (1.87)	0.39 (0.66)	0.22 (0.17)	0.45 (0.17)	0.25 (0.16)	0.08 (0.09)
The entire data set																							
0	86639	7.25 (16.34)	6.17 (13.85)	6.17 (15.12)	148.14 (522.07)	74.07 (262.25)	74.07 (263.05)	0.41 (0.29)	0.04 (0.13)	0.0 (0.0)	33.68 (44.72)	1.68 (4.06)	3.01 (7.21)	1.77 (4.66)	0.79 (2.5)	1.94 (5.27)	2.68 (7.04)	1.13 (3.14)	0.41 (1.5)	0.38 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)

Table C.7: GMM on data from the 2nd month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	4365	3.14 (2.13)	2.72 (1.87)	2.8 (1.97)	95.56 (74.62)	47.21 (38.44)	48.35 (39.16)	0.5 (0.13)	0.0 (0.0)	0.0 (0.0)	34.72 (36.42)	0.55 (0.84)	1.01 (1.24)	0.58 (0.88)	1.0 (0.0)	0.69 (1.0)	0.88 (1.17)	0.81 (0.75)	0.42 (0.49)	0.13 (0.19)	0.24 (0.24)	0.13 (0.19)	0.5 (0.33)
1	4377	31.9 (11.56)	27.41 (10.64)	28.78 (11.15)	446.53 (258.1)	221.11 (129.71)	225.42 (132.77)	0.5 (0.06)	0.04 (0.04)	0.0 (0.0)	51.04 (17.42)	6.0 (3.86)	14.11 (6.46)	8.77 (4.81)	3.03 (2.26)	8.74 (4.99)	13.26 (6.36)	5.48 (3.54)	1.3 (1.31)	0.19 (0.11)	0.44 (0.12)	0.27 (0.11)	0.1 (0.07)
2	15976	1.13 (0.39)	1.13 (0.39)	0.0 (0.0)	1.18 (0.58)	1.18 (0.58)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	36.55 (65.99)	1.1 (0.36)	0.03 (0.18)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.98 (0.09)	0.02 (0.09)	0.0 (0.0)	0.0 (0.0)
3	15507	1.23 (0.59)	1.1 (0.64)	0.94 (0.77)	4.13 (2.71)	2.32 (1.72)	1.81 (1.82)	0.41 (0.3)	0.0 (0.0)	0.0 (0.0)	32.5 (52.59)	0.0 (0.0)	1.23 (0.59)	0.0 (0.0)	0.0 (0.0)	0.41 (0.61)	0.52 (0.65)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)
4	2110	92.15 (49.95)	78.29 (41.0)	82.62 (48.23)	2044.77 (1804.18)	1025.12 (918.72)	1019.65 (901.73)	0.49 (0.09)	0.03 (0.03)	0.0 (0.0)	50.51 (14.41)	18.58 (18.78)	38.1 (23.82)	24.24 (15.02)	11.23 (9.04)	24.63 (23.51)	36.44 (22.82)	15.78 (10.91)	5.77 (5.71)	0.2 (0.13)	0.41 (0.1)	0.26 (0.1)	0.13 (0.08)
5	22286	2.31 (1.39)	1.64 (1.39)	1.9 (1.28)	12.82 (11.76)	6.32 (6.39)	6.49 (6.28)	0.6 (0.27)	0.0 (0.0)	0.0 (0.0)	33.52 (43.51)	0.91 (0.85)	0.78 (0.97)	0.62 (0.69)	0.0 (0.0)	0.91 (0.95)	0.74 (0.9)	0.25 (0.48)	0.0 (0.0)	0.42 (0.38)	0.24 (0.26)	0.34 (0.39)	0.0 (0.0)
6	9501	7.57 (4.81)	6.28 (4.24)	6.09 (4.51)	41.86 (32.66)	21.23 (16.88)	20.63 (17.26)	0.47 (0.17)	0.19 (0.28)	0.0 (0.0)	49.23 (35.6)	2.13 (2.17)	3.68 (2.92)	1.76 (1.73)	0.0 (0.0)	2.42 (2.49)	2.97 (2.59)	0.7 (0.96)	0.0 (0.0)	0.28 (0.24)	0.48 (0.24)	0.24 (0.22)	0.0 (0.0)
7	4871	11.19 (3.76)	9.73 (3.57)	10.11 (3.71)	181.67 (91.99)	89.04 (46.36)	92.63 (48.58)	0.51 (0.07)	0.06 (0.06)	0.0 (0.0)	47.84 (23.78)	1.99 (1.59)	4.51 (2.33)	3.03 (1.98)	1.66 (0.87)	2.78 (1.9)	4.3 (2.34)	2.39 (1.58)	0.64 (0.73)	0.18 (0.13)	0.4 (0.16)	0.27 (0.15)	0.16 (0.08)
8	2571	25.58 (14.97)	23.33 (14.0)	23.21 (14.45)	1751.92 (1757.77)	873.14 (894.79)	878.78 (888.72)	0.5 (0.1)	0.06 (0.05)	0.0 (0.0)	46.51 (18.35)	3.31 (2.71)	8.45 (5.62)	6.9 (5.02)	6.92 (4.88)	4.64 (3.39)	8.61 (6.18)	5.53 (4.28)	4.44 (3.45)	0.14 (0.1)	0.33 (0.12)	0.26 (0.11)	0.28 (0.12)
9	2549	4.25 (2.01)	3.89 (1.93)	3.71 (2.04)	343.27 (514.63)	174.71 (271.24)	168.57 (252.58)	0.49 (0.14)	0.29 (0.32)	0.0 (0.0)	39.56 (34.82)	0.57 (0.8)	1.12 (1.06)	0.91 (0.99)	1.65 (0.96)	0.61 (0.8)	1.07 (1.07)	1.08 (1.0)	0.95 (0.9)	0.12 (0.17)	0.25 (0.22)	0.2 (0.21)	0.43 (0.24)
Clustering 2: 7 Clusters																							
0	17625	1.21 (0.45)	1.17 (0.42)	0.13 (0.44)	1.38 (0.86)	1.24 (0.64)	0.13 (0.44)	0.04 (0.14)	0.0 (0.0)	0.0 (0.0)	36.73 (64.48)	1.09 (0.4)	0.12 (0.39)	0.0 (0.0)	0.0 (0.0)	0.13 (0.44)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.94 (0.2)	0.06 (0.2)	0.0 (0.0)	0.0 (0.0)
1	4138	66.11 (45.7)	56.89 (37.8)	59.29 (43.16)	1944.17 (1848.03)	971.85 (940.99)	972.32 (928.31)	0.5 (0.09)	0.04 (0.03)	0.0 (0.0)	50.26 (15.33)	12.58 (15.08)	26.63 (21.34)	17.5 (13.38)	9.4 (7.71)	16.81 (19.02)	25.57 (20.46)	11.76 (9.41)	5.14 (4.95)	0.18 (0.12)	0.39 (0.11)	0.26 (0.1)	0.16 (0.11)
2	35316	1.79 (1.18)	1.35 (1.12)	1.45 (1.14)	8.89 (9.15)	4.49 (4.92)	4.4 (4.96)	0.52 (0.3)	0.0 (0.0)	0.0 (0.0)	32.72 (47.77)	0.49 (0.75)	0.95 (0.86)	0.35 (0.56)	0.0 (0.0)	0.66 (0.83)	0.66 (0.79)	0.14 (0.36)	0.0 (0.0)	0.24 (0.36)	0.56 (0.43)	0.2 (0.35)	0.0 (0.0)
3	3778	6.82 (4.24)	6.17 (3.98)	6.03 (4.14)	479.5 (723.76)	241.31 (370.63)	238.19 (363.64)	0.49 (0.13)	0.24 (0.28)	0.0 (0.0)	43.45 (31.97)	0.97 (1.15)	2.08 (1.77)	1.63 (1.61)	2.12 (1.67)	1.19 (1.28)	2.0 (1.8)	1.6 (1.46)	1.24 (1.36)	0.14 (0.16)	0.29 (0.2)	0.23 (0.19)	0.34 (0.2)
4	7806	20.7 (10.58)	17.92 (9.47)	18.79 (10.01)	321.49 (223.21)	158.71 (112.22)	162.78 (114.1)	0.51 (0.06)	0.05 (0.04)	0.0 (0.0)	49.23 (20.0)	3.68 (2.87)	8.89 (5.59)	5.75 (3.94)	2.38 (1.79)	5.44 (3.98)	8.47 (5.44)	3.85 (2.75)	1.03 (1.17)	0.18 (0.12)	0.42 (0.13)	0.27 (0.13)	0.12 (0.07)
5	5135	3.58 (2.46)	3.12 (2.19)	3.22 (2.3)	114.41 (96.67)	56.53 (49.39)	57.88 (50.01)	0.51 (0.12)	0.0 (0.0)	0.0 (0.0)	34.89 (35.1)	0.59 (0.88)	1.15 (1.37)	0.68 (0.98)	1.15 (0.4)	0.77 (1.07)	1.02 (1.3)	0.92 (0.86)	0.51 (0.59)	0.13 (0.19)	0.24 (0.24)	0.14 (0.19)	0.49 (0.32)
6	10315	7.33 (4.67)	6.07 (4.14)	5.88 (4.33)	41.15 (31.18)	20.86 (16.2)	20.29 (16.52)	0.47 (0.18)	0.18 (0.28)	0.0 (0.0)	48.4 (35.3)	2.09 (2.14)	3.49 (2.86)	1.75 (1.67)	0.0 (0.0)	2.33 (2.41)	2.85 (2.51)	0.7 (0.94)	0.0 (0.0)	0.28 (0.24)	0.46 (0.24)	0.26 (0.23)	0.0 (0.0)
Clustering 3: 5 Clusters																							
0	54005	1.68 (1.19)	1.36 (1.08)	1.07 (1.25)	6.85 (9.08)	3.65 (4.75)	3.19 (4.91)	0.37 (0.34)	0.0 (0.0)	0.0 (0.0)	34.17 (53.57)	0.71 (0.76)	0.71 (0.89)	0.25 (0.53)	0.0 (0.0)	0.51 (0.81)	0.47 (0.77)	0.1 (0.32)	0.0 (0.0)	0.47 (0.45)	0.4 (0.43)	0.14 (0.3)	0.0 (0.0)
1	11136	12.18 (6.13)	10.24 (5.59)	10.54 (5.93)	113.25 (83.07)	56.29 (41.93)	56.95 (42.92)	0.5 (0.11)	0.04 (0.05)	0.0 (0.0)	48.02 (24.23)	2.8 (2.34)	5.57 (3.56)	3.07 (2.44)	0.73 (0.91)	3.64 (2.87)	4.9 (3.37)	1.78 (1.71)	0.23 (0.47)	0.23 (0.17)	0.45 (0.18)	0.25 (0.16)	0.06 (0.08)
2	4701	10.08 (6.8)	9.14 (6.36)	9.07 (6.45)	645.25 (828.86)	323.38 (426.13)	321.86 (413.4)	0.49 (0.12)	0.12 (0.2)	0.0 (0.0)	41.48 (27.53)	1.36 (1.46)	3.29 (2.83)	2.61 (2.4)	2.82 (2.2)	1.87 (1.83)	3.23 (2.86)	2.15 (2.13)	1.83 (1.56)	0.13 (0.13)	0.29 (0.19)	0.23 (0.17)	0.35 (0.23)
3	8185	3.07 (1.36)	2.69 (1.31)	2.55 (1.4)	52.5 (40.49)	26.16 (21.04)	26.34 (21.36)	0.48 (0.17)	0.24 (0.33)	0.0 (0.0)	43.29 (43.05)	0.66 (0.86)	1.12 (1.04)	0.69 (0.87)	0.6 (0.57)	0.69 (0.85)	0.96 (1.0)	0.7 (0.75)	0.21 (0.41)	0.19 (0.25)	0.33 (0.29)	0.2 (0.25)	0.28 (0.33)
4	6086	55.98 (40.74)	48.18 (33.97)	50.4 (38.15)	1416.77 (1636.51)	706.48 (830.24)	710.3 (821.81)	0.5 (0.08)	0.04 (0.03)	0.0 (0.0)	50.62 (15.76)	10.51 (12.93)	23.0 (18.62)	15.05 (11.84)	7.41 (7.03)	14.41 (16.24)	22.11 (17.86)	9.97 (8.4)	3.92 (4.45)	0.18 (0.12)	0.41 (0.11)	0.27 (0.1)	0.14 (0.1)
The entire data set																							
0	84113	7.6 (18.05)	6.49 (15.4)	6.49 (16.61)	163.07 (614.17)	81.54 (310.03)	81.54 (308.01)	0.41 (0.29)	0.04 (0.14)	0.0 (0.0)	38.49 (46.88)	1.73 (4.46)	3.15 (7.83)	1.87 (5.11)	0.85 (2.79)	2.02 (5.83)	2.82 (7.54)	1.21 (3.51)	0.44 (1.65)	0.37 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.17)

Table C.8: GMM on data from the 3rd month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	17922	4.34 (3.46)	3.66 (3.02)	3.58 (3.05)	28.42 (21.31)	14.38 (11.14)	14.04 (11.45)	0.48 (0.17)	0.04 (0.09)	0.0 (0.0)	48.98 (46.57)	1.1 (1.53)	1.85 (2.14)	1.4 (1.07)	0.0 (0.0)	1.25 (1.69)	1.78 (1.79)	0.55 (0.71)	0.0 (0.0)	0.19 (0.23)	0.32 (0.27)	0.49 (0.34)	0.0 (0.0)
1	24965	1.62 (1.05)	1.13 (1.01)	1.23 (0.99)	4.13 (3.18)	2.14 (2.07)	1.99 (1.9)	0.53 (0.34)	0.0 (0.0)	0.0 (0.0)	34.7 (55.43)	0.57 (0.81)	1.06 (0.79)	0.0 (0.0)	0.0 (0.0)	0.77 (0.85)	0.46 (0.65)	0.0 (0.0)	0.0 (0.0)	0.3 (0.4)	0.7 (0.4)	0.0 (0.0)	0.0 (0.0)
2	3867	44.51 (16.72)	39.12 (15.15)	40.55 (15.94)	1011.33 (661.94)	502.69 (334.05)	508.64 (335.21)	0.5 (0.05)	0.05 (0.04)	0.0 (0.0)	59.24 (20.29)	7.35 (4.85)	18.52 (8.81)	12.51 (6.35)	6.13 (3.99)	10.96 (6.29)	17.93 (8.63)	8.52 (4.98)	3.14 (2.62)	0.16 (0.09)	0.41 (0.11)	0.28 (0.09)	0.15 (0.09)
3	3883	8.5 (2.43)	7.43 (2.43)	7.64 (2.53)	149.96 (79.07)	73.36 (40.29)	76.6 (41.35)	0.51 (0.07)	0.09 (0.1)	0.0 (0.0)	57.3 (36.75)	1.51 (1.29)	3.28 (1.8)	2.28 (1.57)	1.44 (0.79)	2.04 (1.49)	3.08 (1.79)	1.94 (1.31)	0.58 (0.69)	0.18 (0.15)	0.38 (0.18)	0.26 (0.17)	0.18 (0.09)
4	1666	106.83 (65.78)	90.11 (52.98)	95.92 (64.28)	3160.67 (2675.19)	1580.99 (1365.7)	1579.69 (1340.08)	0.49 (0.1)	0.04 (0.03)	0.0 (0.0)	57.22 (17.89)	21.89 (21.87)	43.1 (32.15)	27.38 (20.36)	14.46 (11.46)	28.87 (30.9)	40.95 (30.73)	18.18 (14.15)	7.93 (7.23)	0.21 (0.14)	0.39 (0.1)	0.25 (0.1)	0.15 (0.12)
5	4533	20.3 (5.82)	17.28 (5.75)	17.84 (5.87)	249.05 (140.31)	123.57 (71.02)	125.48 (73.06)	0.5 (0.08)	0.04 (0.04)	0.0 (0.0)	56.97 (24.53)	4.32 (3.15)	9.01 (3.98)	5.24 (2.93)	1.72 (1.4)	5.78 (3.57)	8.16 (3.81)	3.19 (2.18)	0.71 (0.89)	0.21 (0.14)	0.44 (0.14)	0.26 (0.13)	0.09 (0.07)
6	2116	10.97 (7.12)	10.07 (6.65)	9.33 (6.87)	1024.6 (1185.02)	516.06 (600.69)	508.54 (599.89)	0.48 (0.15)	0.17 (0.22)	0.0 (0.0)	46.62 (32.07)	1.71 (2.1)	3.38 (2.92)	2.46 (2.39)	3.41 (2.6)	1.85 (1.95)	3.01 (2.81)	2.14 (2.15)	2.32 (1.97)	0.14 (0.16)	0.28 (0.18)	0.21 (0.16)	0.37 (0.22)
7	15614	1.13 (0.4)	1.13 (0.4)	0.0 (0.0)	1.18 (0.56)	1.18 (0.56)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	45.06 (94.86)	1.1 (0.37)	0.03 (0.17)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.98 (0.09)	0.02 (0.09)	0.0 (0.0)	0.0 (0.0)
8	2506	2.65 (0.88)	2.35 (0.96)	2.12 (1.14)	29.77 (29.78)	14.9 (14.92)	14.86 (15.94)	0.46 (0.2)	0.66 (0.33)	0.0 (0.0)	84.02 (146.82)	0.58 (0.71)	1.09 (0.88)	0.73 (0.84)	0.25 (0.47)	0.59 (0.72)	0.95 (0.88)	0.51 (0.7)	0.07 (0.25)	0.23 (0.29)	0.42 (0.33)	0.27 (0.3)	0.09 (0.17)
9	4281	2.96 (1.84)	2.62 (1.65)	2.67 (1.77)	125.75 (136.17)	62.64 (69.17)	63.11 (69.87)	0.5 (0.13)	0.0 (0.0)	0.0 (0.0)	39.77 (44.62)	0.45 (0.72)	0.87 (1.08)	0.5 (0.76)	1.13 (0.4)	0.57 (0.83)	0.77 (1.04)	0.81 (0.76)	0.52 (0.6)	0.12 (0.18)	0.22 (0.24)	0.12 (0.19)	0.54 (0.32)
Clustering 2: 7 Clusters																							
0	25176	1.63 (1.02)	1.15 (1.0)	1.25 (0.98)	4.41 (3.59)	2.28 (2.24)	2.13 (2.08)	0.53 (0.34)	0.0 (0.0)	0.0 (0.0)	35.46 (58.56)	0.54 (0.77)	1.07 (0.77)	0.03 (0.16)	0.0 (0.0)	0.76 (0.83)	0.5 (0.68)	0.0 (0.0)	0.0 (0.0)	0.29 (0.39)	0.7 (0.39)	0.01 (0.05)	0.0 (0.0)
1	8741	17.48 (9.5)	15.04 (8.54)	15.67 (8.88)	249.55 (170.39)	123.09 (85.64)	126.47 (87.49)	0.51 (0.07)	0.05 (0.05)	0.0 (0.0)	55.98 (27.1)	3.35 (2.87)	7.57 (5.11)	4.67 (3.49)	1.89 (1.48)	4.72 (3.68)	7.05 (4.93)	3.1 (2.43)	0.79 (0.94)	0.19 (0.14)	0.42 (0.15)	0.26 (0.14)	0.12 (0.08)
2	3064	11.69 (8.84)	10.72 (8.29)	10.02 (8.31)	987.61 (1275.35)	497.83 (647.96)	489.77 (641.26)	0.48 (0.14)	0.13 (0.19)	0.0 (0.0)	46.41 (32.41)	1.83 (2.33)	3.67 (3.39)	2.77 (2.78)	3.42 (2.93)	2.04 (2.17)	3.35 (3.32)	2.34 (2.44)	2.3 (2.14)	0.14 (0.16)	0.27 (0.18)	0.21 (0.17)	0.37 (0.25)
3	17947	4.42 (3.53)	3.72 (3.09)	3.62 (3.12)	28.82 (21.8)	14.58 (11.39)	14.23 (11.71)	0.48 (0.18)	0.05 (0.11)	0.0 (0.0)	49.85 (49.22)	1.13 (1.56)	1.89 (2.18)	1.4 (1.1)	0.0 (0.0)	1.28 (1.72)	1.78 (1.83)	0.56 (0.73)	0.0 (0.0)	0.19 (0.23)	0.32 (0.28)	0.49 (0.35)	0.0 (0.0)
4	4239	72.78 (51.1)	62.36 (41.73)	65.64 (48.71)	1885.49 (1951.81)	940.45 (990.65)	945.04 (978.12)	0.5 (0.07)	0.04 (0.04)	0.0 (0.0)	58.53 (19.03)	13.85 (15.68)	29.77 (23.87)	19.35 (15.25)	9.81 (8.78)	19.04 (21.54)	28.42 (22.83)	12.97 (10.68)	5.22 (5.48)	0.19 (0.12)	0.4 (0.11)	0.27 (0.1)	0.14 (0.1)
5	6244	2.88 (1.47)	2.58 (1.4)	2.51 (1.49)	72.46 (55.35)	35.9 (28.25)	36.56 (29.38)	0.49 (0.15)	0.27 (0.38)	0.0 (0.0)	55.64 (91.06)	0.49 (0.69)	0.93 (0.95)	0.61 (0.81)	0.86 (0.57)	0.56 (0.75)	0.83 (0.93)	0.79 (0.79)	0.33 (0.5)	0.15 (0.23)	0.28 (0.28)	0.18 (0.24)	0.38 (0.34)
6	15942	1.15 (0.42)	1.13 (0.4)	0.02 (0.14)	1.19 (0.56)	1.17 (0.55)	0.02 (0.14)	0.01 (0.07)	0.0 (0.0)	0.0 (0.0)	45.03 (94.38)	1.12 (0.39)	0.03 (0.17)	0.0 (0.0)	0.0 (0.0)	0.02 (0.14)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.99 (0.08)	0.01 (0.08)	0.0 (0.0)	0.0 (0.0)
Clustering 3: 5 Clusters																							
0	10344	14.9 (7.81)	12.96 (7.19)	13.12 (7.48)	374.94 (462.59)	186.61 (233.64)	188.33 (233.66)	0.5 (0.1)	0.07 (0.13)	0.0 (0.0)	52.86 (29.57)	2.88 (2.56)	6.12 (4.24)	3.86 (3.05)	2.04 (1.85)	3.82 (3.11)	5.61 (4.1)	2.58 (2.29)	1.12 (1.27)	0.19 (0.16)	0.38 (0.18)	0.24 (0.16)	0.19 (0.22)
1	5322	64.61 (48.61)	55.37 (39.98)	58.17 (46.18)	1743.08 (1945.89)	871.06 (987.41)	872.02 (975.51)	0.5 (0.08)	0.04 (0.04)	0.0 (0.0)	58.11 (20.08)	12.34 (14.43)	26.57 (22.42)	17.09 (14.48)	8.61 (8.35)	16.92 (19.81)	25.29 (21.5)	11.38 (10.17)	4.58 (5.21)	0.19 (0.12)	0.41 (0.11)	0.26 (0.1)	0.14 (0.11)
2	35751	2.02 (1.62)	1.55 (1.47)	1.63 (1.51)	10.45 (11.94)	5.29 (6.28)	5.16 (6.44)	0.52 (0.3)	0.0 (0.0)	0.0 (0.0)	38.49 (62.33)	0.56 (0.88)	1.04 (1.04)	0.42 (0.68)	0.0 (0.0)	0.72 (0.95)	0.75 (0.96)	0.17 (0.42)	0.0 (0.0)	0.24 (0.35)	0.55 (0.42)	0.21 (0.35)	0.0 (0.0)
3	12908	4.87 (2.75)	4.23 (2.53)	4.09 (2.64)	57.96 (42.63)	28.96 (21.78)	29.0 (22.65)	0.48 (0.15)	0.2 (0.29)	0.0 (0.0)	56.7 (54.94)	1.11 (1.24)	2.07 (1.78)	1.15 (1.26)	0.54 (0.6)	1.3 (1.39)	1.76 (1.65)	0.86 (0.93)	0.17 (0.38)	0.21 (0.23)	0.38 (0.27)	0.21 (0.22)	0.19 (0.29)
4	17028	1.21 (0.46)	1.16 (0.43)	0.11 (0.41)	1.34 (0.79)	1.22 (0.61)	0.11 (0.41)	0.04 (0.13)	0.0 (0.0)	0.0 (0.0)	45.13 (92.64)	1.1 (0.41)	0.1 (0.35)	0.0 (0.0)	0.0 (0.0)	0.11 (0.41)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.95 (0.17)	0.05 (0.17)	0.0 (0.0)	0.0 (0.0)
The entire data set																							
0	81353	8.04 (20.18)	6.86 (17.09)	6.86 (18.67)	175.77 (679.4)	87.89 (342.49)	87.89 (340.55)	0.41 (0.29)	0.04 (0.14)	0.0 (0.0)	45.88 (64.64)	1.82 (4.83)	3.33 (8.79)	1.98 (5.72)	0.91 (3.11)	2.14 (6.64)	2.98 (8.42)	1.28 (3.93)	0.47 (1.82)	0.37 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.17)

Table C.9: GMM on data from the 4th month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. de-gree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neigh-bors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 mes-sage out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 mes-sage (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	21584	1.64 (1.09)	1.15 (1.04)	1.25 (1.02)	4.18 (3.23)	2.15 (2.06)	2.03 (1.94)	0.53 (0.34)	0.0 (0.0)	0.0 (0.0)	27.04 (45.82)	0.57 (0.82)	1.08 (0.82)	0.0 (0.0)	0.0 (0.0)	0.77 (0.87)	0.47 (0.66)	0.0 (0.0)	0.0 (0.0)	0.3 (0.4)	0.7 (0.4)	0.0 (0.0)	0.0 (0.0)
1	1640	80.72 (46.21)	69.08 (39.08)	72.85 (44.91)	2409.53 (2134.85)	1207.22 (1084.99)	1202.3 (1070.62)	0.49 (0.09)	0.03 (0.03)	0.0 (0.0)	47.05 (15.7)	15.49 (14.8)	32.17 (21.84)	21.64 (15.63)	11.41 (9.01)	20.5 (17.73)	31.72 (23.55)	14.21 (11.03)	6.42 (5.96)	0.19 (0.14)	0.39 (0.11)	0.26 (0.1)	0.16 (0.12)
2	2482	3.16 (1.29)	2.82 (1.29)	2.65 (1.45)	53.51 (60.38)	26.34 (29.8)	27.17 (31.82)	0.48 (0.18)	0.55 (0.36)	0.0 (0.0)	82.64 (155.42)	0.63 (0.77)	1.14 (0.94)	0.88 (0.91)	0.51 (0.64)	0.68 (0.79)	1.03 (0.95)	0.75 (0.84)	0.19 (0.42)	0.21 (0.27)	0.38 (0.3)	0.27 (0.28)	0.14 (0.19)
3	15347	4.22 (3.34)	3.58 (2.92)	3.52 (2.97)	28.15 (20.67)	14.3 (10.89)	13.85 (11.05)	0.48 (0.17)	0.04 (0.08)	0.0 (0.0)	40.48 (44.3)	1.03 (1.45)	1.8 (2.1)	1.39 (1.03)	0.0 (0.0)	1.22 (1.64)	1.76 (1.75)	0.54 (0.71)	0.0 (0.0)	0.18 (0.23)	0.31 (0.28)	0.5 (0.34)	0.0 (0.0)
4	2207	11.7 (6.87)	10.66 (6.57)	10.59 (6.54)	887.67 (916.3)	442.54 (463.4)	445.13 (464.67)	0.5 (0.1)	0.11 (0.12)	0.0 (0.0)	43.68 (34.69)	1.61 (1.56)	3.65 (2.75)	2.91 (2.39)	3.52 (2.6)	2.11 (1.78)	3.61 (2.88)	2.56 (2.17)	2.32 (1.93)	0.14 (0.13)	0.3 (0.17)	0.23 (0.15)	0.33 (0.19)
5	3	880.33 (174.27)	505.0 (195.67)	848.67 (138.0)	2761.33 (1033.72)	1296.0 (714.49)	1465.33 (333.31)	0.56 (0.08)	0.01 (0.0)	0.0 (0.0)	25.58 (4.08)	378.33 (42.03)	450.67 (156.42)	47.0 (14.76)	4.33 (4.78)	625.33 (143.79)	210.33 (14.06)	10.67 (5.44)	2.33 (2.05)	0.44 (0.09)	0.5 (0.1)	0.05 (0.01)	0.0 (0.0)
6	4520	14.35 (5.12)	12.22 (4.92)	12.52 (4.95)	167.77 (84.23)	83.07 (42.91)	84.7 (44.67)	0.5 (0.09)	0.05 (0.05)	0.0 (0.0)	48.48 (29.23)	3.09 (2.6)	6.27 (3.45)	3.65 (2.39)	1.33 (0.98)	4.03 (2.9)	5.61 (3.17)	2.43 (1.75)	0.46 (0.64)	0.21 (0.15)	0.43 (0.15)	0.26 (0.15)	0.11 (0.08)
7	4137	3.4 (2.26)	2.98 (2.02)	3.1 (2.16)	115.91 (100.94)	57.7 (51.9)	58.2 (52.14)	0.5 (0.12)	0.0 (0.0)	0.0 (0.0)	30.21 (32.68)	0.53 (0.81)	1.08 (1.29)	0.64 (0.94)	1.15 (0.43)	0.71 (0.99)	0.97 (1.24)	0.91 (0.84)	0.5 (0.58)	0.12 (0.18)	0.23 (0.24)	0.14 (0.19)	0.51 (0.32)
8	13119	1.13 (0.41)	1.13 (0.41)	0.0 (0.0)	1.17 (0.53)	1.17 (0.53)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	36.74 (90.5)	1.1 (0.38)	0.03 (0.17)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.99 (0.08)	0.01 (0.08)	0.0 (0.0)	0.0 (0.0)
9	3443	34.6 (11.67)	30.05 (10.82)	31.57 (11.22)	636.51 (393.17)	316.04 (198.04)	320.48 (200.37)	0.5 (0.05)	0.04 (0.04)	0.0 (0.0)	48.45 (20.5)	6.07 (3.83)	14.61 (6.71)	9.77 (4.86)	4.16 (2.92)	9.07 (5.01)	14.12 (6.55)	6.31 (3.83)	2.07 (1.88)	0.18 (0.1)	0.42 (0.12)	0.28 (0.1)	0.13 (0.08)
Clustering 2: 7 Clusters																							
0	17290	1.74 (1.17)	1.4 (1.02)	1.23 (1.12)	4.88 (3.26)	2.67 (2.04)	2.21 (2.13)	0.43 (0.3)	0.0 (0.0)	0.0 (0.0)	29.77 (47.91)	0.42 (0.82)	1.32 (0.73)	0.0 (0.0)	0.0 (0.0)	0.64 (0.92)	0.59 (0.69)	0.0 (0.0)	0.0 (0.0)	0.13 (0.24)	0.87 (0.24)	0.0 (0.0)	0.0 (0.0)
1	7276	19.55 (10.0)	16.91 (8.97)	17.63 (9.5)	316.02 (234.98)	156.13 (117.68)	159.89 (120.24)	0.51 (0.07)	0.05 (0.05)	0.0 (0.0)	48.73 (26.67)	3.63 (2.92)	8.3 (5.36)	5.34 (3.77)	2.28 (1.84)	5.22 (3.92)	7.83 (5.18)	3.55 (2.64)	1.03 (1.21)	0.19 (0.13)	0.41 (0.14)	0.27 (0.13)	0.13 (0.09)
2	15952	4.39 (3.59)	3.7 (3.13)	3.65 (3.17)	28.82 (21.98)	14.63 (11.6)	14.19 (11.64)	0.48 (0.17)	0.04 (0.1)	0.0 (0.0)	41.45 (46.98)	1.1 (1.57)	1.88 (2.21)	1.41 (1.1)	0.0 (0.0)	1.29 (1.75)	1.8 (1.84)	0.55 (0.73)	0.0 (0.0)	0.19 (0.23)	0.32 (0.28)	0.49 (0.34)	0.0 (0.0)
3	17424	1.16 (0.4)	0.89 (0.56)	0.32 (0.55)	1.22 (0.59)	0.91 (0.59)	0.32 (0.55)	0.24 (0.41)	0.0 (0.0)	0.0 (0.0)	32.54 (84.16)	1.11 (0.35)	0.05 (0.22)	0.0 (0.0)	0.0 (0.0)	0.32 (0.55)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.98 (0.11)	0.02 (0.11)	0.0 (0.0)	0.0 (0.0)
4	3111	4.14 (2.41)	3.73 (2.28)	3.58 (2.46)	231.23 (433.96)	116.28 (220.08)	114.96 (218.37)	0.48 (0.17)	0.45 (0.38)	0.0 (0.0)	65.34 (127.44)	0.67 (0.85)	1.34 (1.2)	1.05 (1.1)	1.07 (1.12)	0.77 (0.91)	1.23 (1.2)	0.98 (1.06)	0.59 (0.88)	0.17 (0.24)	0.33 (0.28)	0.25 (0.25)	0.24 (0.25)
5	4188	3.59 (2.48)	3.12 (2.18)	3.29 (2.35)	111.75 (89.97)	55.25 (45.98)	56.51 (46.58)	0.51 (0.12)	0.0 (0.0)	0.0 (0.0)	29.94 (30.39)	0.57 (0.86)	1.17 (1.4)	0.69 (0.99)	1.16 (0.44)	0.78 (1.09)	1.06 (1.33)	0.95 (0.86)	0.49 (0.58)	0.12 (0.18)	0.24 (0.24)	0.14 (0.19)	0.5 (0.32)
6	3241	59.07 (48.78)	50.74 (38.3)	53.33 (46.57)	1840.96 (1744.32)	920.95 (885.05)	920.01 (876.0)	0.49 (0.09)	0.04 (0.04)	0.0 (0.0)	46.41 (19.22)	11.17 (16.45)	23.54 (23.56)	15.7 (13.43)	8.67 (7.43)	15.07 (24.07)	22.86 (20.85)	10.48 (9.31)	4.92 (4.84)	0.18 (0.13)	0.38 (0.12)	0.26 (0.11)	0.18 (0.13)
Clustering 3: 5 Clusters																							
0	26924	2.13 (1.65)	1.77 (1.44)	1.71 (1.56)	11.89 (12.23)	6.11 (6.45)	5.78 (6.62)	0.46 (0.26)	0.0 (0.0)	0.0 (0.0)	33.53 (59.71)	0.45 (0.85)	1.19 (1.03)	0.49 (0.72)	0.0 (0.0)	0.65 (0.98)	0.86 (0.97)	0.19 (0.45)	0.0 (0.0)	0.12 (0.22)	0.63 (0.4)	0.25 (0.37)	0.0 (0.0)
1	7059	3.03 (1.47)	2.69 (1.4)	2.59 (1.49)	86.9 (115.24)	43.04 (57.59)	43.86 (59.6)	0.49 (0.16)	0.25 (0.33)	0.0 (0.0)	44.37 (62.16)	0.58 (0.8)	1.01 (1.02)	0.67 (0.87)	0.76 (0.7)	0.64 (0.82)	0.88 (0.96)	0.72 (0.8)	0.35 (0.56)	0.17 (0.24)	0.3 (0.29)	0.19 (0.24)	0.34 (0.35)
2	5510	44.79 (42.03)	38.6 (33.63)	40.63 (39.64)	1365.26 (1477.76)	681.8 (749.09)	683.46 (741.3)	0.5 (0.08)	0.05 (0.05)	0.0 (0.0)	46.53 (23.02)	8.23 (13.29)	17.87 (19.83)	12.04 (11.63)	6.65 (6.34)	11.38 (19.25)	17.4 (17.83)	8.09 (7.99)	3.76 (4.07)	0.17 (0.13)	0.37 (0.14)	0.26 (0.12)	0.2 (0.16)
3	18468	1.22 (0.49)	0.95 (0.62)	0.35 (0.61)	1.38 (0.88)	1.02 (0.79)	0.35 (0.61)	0.24 (0.4)	0.0 (0.0)	0.0 (0.0)	32.57 (82.55)	1.11 (0.45)	0.11 (0.37)	0.0 (0.0)	0.0 (0.0)	0.35 (0.61)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.94 (0.19)	0.06 (0.19)	0.0 (0.0)	0.0 (0.0)
4	10521	12.7 (6.92)	10.85 (6.31)	11.06 (6.65)	146.11 (119.93)	72.76 (60.33)	73.35 (61.64)	0.49 (0.11)	0.05 (0.06)	0.0 (0.0)	47.78 (30.53)	2.77 (2.36)	5.66 (3.86)	3.24 (2.68)	1.03 (1.18)	3.65 (2.93)	5.03 (3.69)	1.98 (1.89)	0.4 (0.68)	0.23 (0.18)	0.44 (0.18)	0.25 (0.16)	0.08 (0.09)
The entire data set																							
0	68482	7.04 (17.05)	6.0 (14.23)	6.0 (15.87)	146.3 (558.59)	73.15 (281.31)	73.15 (280.09)	0.41 (0.29)	0.04 (0.13)	0.0 (0.0)	37.62 (62.16)	1.62 (4.46)	2.91 (7.57)	1.73 (4.76)	0.77 (2.58)	1.89 (6.39)	2.6 (7.07)	1.1 (3.25)	0.4 (1.57)	0.37 (0.4)	0.39 (0.37)	0.18 (0.27)	0.06 (0.17)

Table C.10: GMM on data from the 5th month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	16903	1.14 (0.42)	1.14 (0.42)	0.0 (0.0)	1.19 (0.58)	1.19 (0.58)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	49.88 (93.99)	1.11 (0.39)	0.03 (0.18)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.98 (0.09)	0.02 (0.09)	0.0 (0.0)	0.0 (0.0)
1	4825	36.79 (14.46)	31.93 (13.33)	33.18 (13.9)	592.39 (389.98)	295.0 (195.99)	297.4 (198.67)	0.5 (0.06)	0.04 (0.04)	0.0 (0.0)	60.46 (21.45)	6.75 (4.41)	15.89 (7.68)	10.21 (5.69)	3.94 (2.98)	9.75 (5.6)	15.13 (7.62)	6.51 (4.25)	1.78 (1.8)	0.19 (0.11)	0.43 (0.11)	0.28 (0.11)	0.11 (0.07)
2	9808	8.35 (4.98)	6.86 (4.46)	6.77 (4.67)	45.89 (32.91)	23.16 (17.08)	22.73 (17.49)	0.48 (0.16)	0.05 (0.08)	0.0 (0.0)	58.33 (38.88)	2.4 (2.33)	4.0 (3.08)	1.95 (1.72)	0.0 (0.0)	2.75 (2.58)	3.26 (2.73)	0.76 (0.99)	0.0 (0.0)	0.29 (0.22)	0.47 (0.21)	0.25 (0.17)	0.0 (0.0)
3	2053	18.46 (10.0)	16.93 (9.51)	16.55 (9.56)	1404.39 (1391.31)	698.56 (702.82)	705.84 (707.38)	0.5 (0.1)	0.08 (0.08)	0.0 (0.0)	55.42 (28.83)	2.56 (2.32)	5.88 (3.89)	4.76 (3.4)	5.27 (3.6)	3.22 (2.42)	5.77 (4.05)	4.09 (3.08)	3.47 (2.65)	0.14 (0.12)	0.31 (0.14)	0.25 (0.13)	0.3 (0.15)
4	26005	1.62 (1.05)	1.14 (1.01)	1.22 (0.99)	4.18 (3.21)	2.17 (2.06)	2.01 (1.93)	0.53 (0.34)	0.0 (0.0)	0.0 (0.0)	38.35 (62.08)	0.56 (0.8)	1.07 (0.8)	0.0 (0.0)	0.0 (0.0)	0.75 (0.84)	0.47 (0.67)	0.0 (0.0)	0.0 (0.0)	0.3 (0.4)	0.7 (0.4)	0.0 (0.0)	0.0 (0.0)
5	2177	101.43 (63.1)	86.13 (50.37)	91.74 (61.02)	2926.66 (2579.16)	1466.39 (1316.42)	1460.27 (1292.64)	0.5 (0.09)	0.04 (0.03)	0.0 (0.0)	58.47 (17.69)	20.06 (22.62)	40.81 (29.93)	26.42 (18.13)	14.14 (10.91)	27.09 (29.92)	39.13 (29.1)	17.75 (12.82)	7.77 (7.07)	0.19 (0.13)	0.39 (0.11)	0.26 (0.1)	0.15 (0.1)
6	4398	3.2 (2.0)	2.83 (1.81)	2.9 (1.88)	94.58 (65.58)	46.54 (33.06)	48.03 (34.92)	0.5 (0.12)	0.0 (0.0)	0.0 (0.0)	44.24 (49.74)	0.53 (0.79)	0.97 (1.16)	0.58 (0.83)	1.12 (0.37)	0.68 (0.94)	0.85 (1.09)	0.93 (0.79)	0.44 (0.56)	0.13 (0.19)	0.23 (0.24)	0.13 (0.19)	0.51 (0.32)
7	2128	3.81 (1.65)	3.45 (1.59)	3.31 (1.66)	261.71 (401.84)	130.48 (202.88)	131.23 (203.46)	0.5 (0.14)	0.37 (0.34)	0.0 (0.0)	52.83 (50.78)	0.57 (0.78)	1.08 (1.0)	0.85 (0.88)	1.32 (0.79)	0.61 (0.78)	1.01 (0.99)	0.97 (0.89)	0.72 (0.75)	0.13 (0.18)	0.26 (0.23)	0.22 (0.24)	0.39 (0.25)
8	5151	11.75 (4.11)	10.19 (3.87)	10.57 (4.05)	177.44 (88.0)	87.67 (44.79)	89.77 (46.19)	0.51 (0.07)	0.06 (0.07)	0.0 (0.0)	59.66 (32.17)	2.16 (1.73)	4.85 (2.62)	3.15 (2.05)	1.59 (0.85)	3.04 (2.09)	4.55 (2.58)	2.39 (1.61)	0.59 (0.7)	0.18 (0.14)	0.41 (0.16)	0.27 (0.15)	0.14 (0.07)
9	11520	2.04 (0.98)	1.84 (0.94)	1.72 (0.99)	18.31 (10.83)	9.31 (6.01)	9.0 (6.33)	0.48 (0.19)	0.13 (0.29)	0.0 (0.0)	55.98 (88.9)	0.37 (0.6)	0.6 (0.76)	1.07 (0.58)	0.0 (0.0)	0.39 (0.62)	0.91 (0.83)	0.42 (0.56)	0.0 (0.0)	0.14 (0.23)	0.23 (0.28)	0.63 (0.34)	0.0 (0.0)
Clustering 2: 7 Clusters																							
0	25642	1.29 (0.51)	1.01 (0.64)	0.46 (0.7)	1.81 (1.73)	1.21 (1.05)	0.61 (1.07)	0.26 (0.39)	0.0 (0.0)	0.0 (0.0)	44.83 (85.38)	1.06 (0.47)	0.23 (0.54)	0.0 (0.0)	0.0 (0.0)	0.36 (0.6)	0.1 (0.33)	0.0 (0.0)	0.0 (0.0)	0.89 (0.27)	0.11 (0.27)	0.0 (0.0)	0.0 (0.0)
1	7483	27.4 (13.37)	23.89 (12.1)	24.8 (12.57)	512.44 (368.05)	254.17 (184.95)	258.26 (187.26)	0.5 (0.06)	0.05 (0.05)	0.0 (0.0)	59.63 (23.99)	4.86 (3.79)	11.54 (6.98)	7.58 (4.91)	3.41 (2.63)	7.08 (4.94)	11.02 (6.75)	5.02 (3.61)	1.68 (1.66)	0.18 (0.11)	0.41 (0.13)	0.28 (0.12)	0.14 (0.1)
2	29527	2.1 (1.62)	1.75 (1.4)	1.66 (1.55)	12.06 (12.26)	6.21 (6.42)	5.85 (6.74)	0.45 (0.27)	0.0 (0.0)	0.0 (0.0)	44.63 (69.75)	0.43 (0.83)	1.16 (1.01)	0.51 (0.72)	0.0 (0.0)	0.62 (0.96)	0.83 (0.98)	0.2 (0.46)	0.0 (0.0)	0.11 (0.21)	0.63 (0.4)	0.26 (0.37)	0.0 (0.0)
3	3191	81.66 (60.88)	69.74 (49.26)	73.6 (58.2)	2618.64 (2370.31)	1310.41 (1207.21)	1308.23 (1191.45)	0.5 (0.1)	0.04 (0.04)	0.0 (0.0)	58.09 (19.6)	15.93 (19.94)	32.7 (28.17)	21.24 (17.35)	11.8 (9.99)	21.37 (26.48)	31.26 (27.42)	14.37 (12.08)	6.61 (6.39)	0.19 (0.13)	0.38 (0.11)	0.25 (0.1)	0.17 (0.13)
4	3734	3.84 (2.48)	3.47 (2.31)	3.32 (2.48)	293.59 (549.91)	146.22 (279.38)	147.38 (275.42)	0.48 (0.18)	0.45 (0.42)	0.0 (0.0)	52.13 (57.09)	0.62 (0.81)	1.09 (1.06)	0.96 (1.11)	1.17 (1.21)	0.62 (0.83)	1.07 (1.13)	0.83 (1.07)	0.8 (0.93)	0.17 (0.24)	0.29 (0.27)	0.24 (0.26)	0.31 (0.3)
5	8592	10.29 (3.68)	8.65 (3.6)	8.61 (3.71)	87.1 (53.41)	43.68 (27.47)	43.42 (28.02)	0.49 (0.12)	0.06 (0.07)	0.0 (0.0)	60.67 (34.5)	2.62 (2.17)	4.7 (2.61)	2.43 (1.79)	0.54 (0.71)	3.12 (2.29)	3.96 (2.44)	1.39 (1.3)	0.15 (0.36)	0.25 (0.19)	0.45 (0.18)	0.24 (0.17)	0.06 (0.08)
6	6799	3.31 (1.45)	2.9 (1.41)	2.71 (1.45)	51.73 (36.93)	25.85 (18.98)	25.87 (19.75)	0.48 (0.16)	0.13 (0.17)	0.0 (0.0)	57.31 (61.18)	0.79 (1.0)	1.24 (1.13)	0.63 (0.82)	0.65 (0.58)	0.81 (0.94)	1.0 (1.03)	0.7 (0.74)	0.19 (0.4)	0.21 (0.25)	0.33 (0.28)	0.17 (0.21)	0.3 (0.34)
Clustering 3: 5 Clusters																							
0	7451	53.19 (48.3)	45.86 (39.72)	48.06 (45.4)	1565.61 (1840.8)	781.95 (934.22)	783.66 (923.4)	0.5 (0.08)	0.05 (0.05)	0.0 (0.0)	58.38 (22.56)	9.92 (14.42)	21.43 (21.76)	14.2 (13.58)	7.64 (7.77)	13.63 (19.07)	20.63 (21.04)	9.61 (9.44)	4.19 (4.87)	0.17 (0.12)	0.38 (0.13)	0.26 (0.11)	0.18 (0.14)
1	37627	2.0 (1.52)	1.55 (1.4)	1.6 (1.43)	10.14 (11.44)	5.15 (6.03)	4.99 (6.18)	0.51 (0.3)	0.0 (0.0)	0.0 (0.0)	42.45 (68.68)	0.55 (0.84)	1.05 (1.0)	0.4 (0.66)	0.0 (0.0)	0.72 (0.92)	0.73 (0.93)	0.16 (0.41)	0.0 (0.0)	0.24 (0.35)	0.56 (0.42)	0.2 (0.34)	0.0 (0.0)
2	14082	12.06 (6.82)	10.29 (6.14)	10.41 (6.52)	132.46 (109.53)	66.03 (54.93)	66.43 (56.45)	0.49 (0.11)	0.06 (0.07)	0.0 (0.0)	60.11 (33.77)	2.71 (2.41)	5.39 (3.83)	3.03 (2.53)	0.93 (1.11)	3.49 (2.94)	4.71 (3.6)	1.86 (1.83)	0.35 (0.63)	0.23 (0.18)	0.44 (0.18)	0.25 (0.16)	0.08 (0.09)
3	17636	1.18 (0.46)	1.14 (0.42)	0.06 (0.31)	1.24 (0.64)	1.18 (0.57)	0.06 (0.31)	0.02 (0.12)	0.0 (0.0)	0.0 (0.0)	49.61 (93.01)	1.13 (0.42)	0.05 (0.22)	0.0 (0.0)	0.0 (0.0)	0.06 (0.31)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.97 (0.11)	0.03 (0.11)	0.0 (0.0)	0.0 (0.0)
4	8172	2.9 (1.42)	2.59 (1.36)	2.44 (1.47)	104.69 (203.94)	51.89 (102.48)	52.79 (103.9)	0.48 (0.17)	0.28 (0.36)	0.0 (0.0)	55.13 (64.02)	0.57 (0.79)	0.93 (0.95)	0.6 (0.81)	0.79 (0.74)	0.57 (0.76)	0.82 (0.92)	0.67 (0.76)	0.37 (0.6)	0.18 (0.25)	0.29 (0.29)	0.18 (0.24)	0.35 (0.35)
The entire data set																							
0	84968	8.07 (20.58)	6.9 (17.38)	6.9 (19.07)	174.06 (701.32)	87.03 (353.69)	87.03 (351.63)	0.4 (0.29)	0.04 (0.14)	0.0 (0.0)	49.48 (67.55)	1.85 (5.14)	3.34 (8.89)	1.98 (5.73)	0.9 (3.17)	2.16 (6.9)	2.99 (8.59)	1.29 (3.95)	0.46 (1.88)	0.38 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)

Table C.11: New GMM clustering on data from the 4th month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{CC_{ind}}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
10 Clusters																							
0	2591	15.73 (6.22)	13.24 (5.98)	13.12 (5.62)	122.78 (52.25)	62.88 (28.43)	59.91 (27.53)	0.49 (0.11)	0.05 (0.05)	0.0 (0.0)	48.01 (26.89)	4.04 (3.4)	7.23 (4.11)	3.6 (2.43)	0.86 (0.73)	4.86 (3.67)	5.98 (3.43)	2.27 (1.62)	0.0 (0.0)	0.24 (0.17)	0.45 (0.16)	0.24 (0.15)	0.07 (0.06)
1	1420	44.82 (8.8)	38.57 (8.99)	40.46 (9.3)	637.04 (301.4)	318.0 (153.82)	319.04 (154.36)	0.5 (0.06)	0.04 (0.04)	0.0 (0.0)	49.46 (19.0)	8.51 (4.5)	19.78 (6.41)	12.26 (4.93)	4.27 (2.76)	12.38 (5.34)	18.68 (6.36)	7.52 (4.1)	1.87 (1.64)	0.19 (0.1)	0.44 (0.11)	0.27 (0.1)	0.1 (0.06)
2	909	7.02 (2.09)	6.33 (2.19)	6.34 (2.18)	439.51 (407.58)	218.2 (211.67)	221.32 (201.39)	0.51 (0.08)	0.16 (0.14)	0.0 (0.0)	42.74 (37.23)	1.06 (1.1)	2.14 (1.47)	1.68 (1.33)	2.14 (1.17)	1.34 (1.14)	2.01 (1.43)	1.6 (1.3)	1.39 (0.97)	0.15 (0.16)	0.3 (0.19)	0.23 (0.17)	0.32 (0.17)
3	482	117.96 (61.57)	94.18 (55.17)	105.69 (61.23)	1690.88 (1711.45)	839.23 (856.81)	851.65 (877.65)	0.51 (0.12)	0.03 (0.03)	0.0 (0.0)	43.54 (15.07)	29.04 (19.87)	50.26 (28.62)	29.38 (22.96)	9.28 (10.62)	36.03 (23.95)	48.32 (33.41)	17.24 (16.1)	4.1 (5.74)	0.28 (0.17)	0.42 (0.11)	0.23 (0.12)	0.07 (0.06)
4	1126	67.72 (23.0)	60.93 (21.23)	62.59 (22.39)	2769.72 (2152.34)	1387.22 (1100.8)	1382.51 (1070.21)	0.5 (0.05)	0.04 (0.03)	0.0 (0.0)	48.57 (15.86)	9.39 (5.71)	25.37 (11.25)	19.74 (8.67)	13.22 (7.49)	14.39 (7.47)	26.16 (11.49)	14.15 (6.93)	7.89 (5.5)	0.14 (0.07)	0.37 (0.1)	0.29 (0.08)	0.2 (0.1)
5	544	12.7 (8.76)	11.45 (8.4)	10.41 (8.15)	1802.56 (1428.22)	912.92 (717.64)	889.63 (744.95)	0.47 (0.18)	0.05 (0.06)	0.0 (0.0)	34.67 (30.83)	2.26 (2.6)	3.72 (3.21)	2.5 (2.34)	4.23 (3.44)	2.13 (2.07)	3.11 (2.97)	2.13 (2.27)	3.04 (2.72)	0.15 (0.16)	0.25 (0.18)	0.17 (0.14)	0.43 (0.28)
6	1885	11.61 (2.69)	10.16 (2.79)	10.56 (2.85)	250.89 (119.52)	121.74 (61.48)	129.15 (61.8)	0.52 (0.07)	0.07 (0.07)	0.0 (0.0)	48.5 (32.84)	1.96 (1.52)	4.46 (2.04)	3.15 (1.75)	2.04 (1.1)	2.76 (1.76)	4.29 (1.96)	2.33 (1.63)	1.18 (0.73)	0.17 (0.13)	0.38 (0.15)	0.27 (0.14)	0.18 (0.09)
7	1330	24.79 (5.32)	21.27 (5.26)	22.66 (5.57)	303.09 (93.9)	149.32 (49.08)	153.76 (50.34)	0.51 (0.06)	0.03 (0.02)	0.0 (0.0)	49.49 (23.38)	4.5 (2.65)	11.02 (4.11)	7.19 (3.26)	2.08 (1.21)	6.79 (3.38)	10.77 (4.19)	4.21 (2.41)	0.9 (0.75)	0.18 (0.1)	0.44 (0.12)	0.29 (0.12)	0.09 (0.05)
8	3	880.33 (174.27)	505.0 (195.67)	848.67 (138.0)	2761.33 (1033.72)	1296.0 (714.49)	1465.33 (333.31)	0.56 (0.08)	0.01 (0.0)	0.0 (0.0)	25.58 (4.08)	378.33 (42.03)	450.67 (156.42)	47.0 (14.76)	4.33 (4.78)	625.33 (143.79)	210.33 (14.06)	10.67 (5.44)	2.33 (2.05)	0.44 (0.09)	0.5 (0.1)	0.05 (0.01)	0.0 (0.0)
9	1523	23.98 (6.31)	21.69 (5.99)	22.2 (6.27)	950.72 (324.52)	470.18 (267.5)	480.54 (264.2)	0.51 (0.05)	0.06 (0.06)	0.0 (0.0)	48.73 (25.09)	3.17 (2.07)	8.4 (3.48)	6.94 (3.09)	5.46 (2.52)	4.81 (2.58)	8.72 (3.57)	5.39 (2.73)	3.28 (1.87)	0.13 (0.08)	0.35 (0.11)	0.29 (0.1)	0.23 (0.09)
The entire data set																							
0	11810	28.97 (29.56)	25.02 (25.33)	26.09 (27.6)	750.26 (1164.92)	374.27 (588.56)	375.99 (584.01)	0.5 (0.09)	0.06 (0.07)	0.0 (0.0)	47.37 (26.68)	5.41 (7.53)	11.81 (13.02)	7.8 (9.1)	3.97 (5.08)	7.43 (9.43)	11.34 (13.36)	5.22 (6.25)	2.1 (3.23)	0.18 (0.14)	0.39 (0.15)	0.26 (0.13)	0.16 (0.14)

Table C.12: BIRCH on data from the 1st month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	525	89.25 (20.65)	74.75 (20.43)	80.25 (20.22)	1407.74 (589.86)	708.95 (308.7)	698.8 (297.93)	0.5 (0.07)	0.03 (0.03)	0.0 (0.0)	45.79 (13.42)	18.79 (13.52)	38.06 (14.17)	23.54 (9.76)	8.86 (4.47)	25.95 (15.63)	35.36 (12.92)	15.06 (7.38)	3.87 (2.58)	0.21 (0.13)	0.42 (0.09)	0.27 (0.11)	0.11 (0.06)
1	44	246.7 (71.7)	176.02 (70.14)	230.07 (69.75)	1808.39 (1247.13)	889.32 (661.62)	919.07 (609.73)	0.52 (0.07)	0.01 (0.01)	0.0 (0.0)	38.03 (10.27)	76.23 (35.66)	117.61 (40.88)	44.43 (35.77)	8.43 (7.35)	109.27 (42.02)	97.61 (59.08)	20.48 (17.04)	2.7 (3.16)	0.32 (0.14)	0.48 (0.11)	0.17 (0.1)	0.03 (0.03)
2	75	148.36 (64.05)	131.99 (54.77)	141.87 (62.01)	7775.28 (2155.99)	3917.76 (1069.05)	3857.52 (1163.07)	0.5 (0.03)	0.02 (0.02)	0.0 (0.0)	41.75 (9.94)	18.67 (14.89)	49.96 (31.08)	45.29 (22.73)	34.44 (9.26)	29.45 (21.23)	57.2 (33.53)	34.15 (15.43)	21.07 (6.51)	0.12 (0.06)	0.32 (0.08)	0.3 (0.06)	0.26 (0.1)
3	80889	4.15 (5.67)	3.5 (4.97)	3.35 (5.29)	54.45 (126.08)	27.31 (64.15)	27.14 (64.08)	0.4 (0.3)	0.04 (0.14)	0.0 (0.0)	33.0 (45.99)	1.11 (1.46)	1.76 (2.75)	0.95 (1.87)	0.34 (0.86)	1.14 (1.91)	1.47 (2.63)	0.59 (1.3)	0.15 (0.49)	0.39 (0.41)	0.39 (0.38)	0.17 (0.27)	0.06 (0.16)
4	21	2.38 (0.58)	2.33 (0.56)	2.19 (0.79)	793.19 (797.51)	457.57 (504.82)	335.62 (324.73)	0.45 (0.12)	0.87 (0.21)	0.02 (0.01)	16.63 (14.22)	0.05 (0.21)	0.19 (0.39)	0.24 (0.43)	1.9 (0.61)	0.14 (0.35)	0.1 (0.29)	0.38 (0.49)	1.57 (0.66)	0.02 (0.07)	0.07 (0.15)	0.1 (0.19)	0.81 (0.21)
5	512	73.85 (26.65)	66.53 (22.64)	68.81 (25.23)	3806.88 (1411.26)	1894.22 (727.02)	1912.66 (735.73)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	44.65 (11.1)	9.66 (7.61)	25.32 (11.6)	21.31 (9.19)	17.56 (5.6)	14.06 (8.62)	27.4 (12.25)	16.45 (7.52)	10.9 (3.81)	0.12 (0.07)	0.34 (0.07)	0.29 (0.07)	0.25 (0.09)
6	131	157.16 (26.11)	132.88 (24.05)	138.32 (36.08)	2136.91 (878.33)	1076.06 (465.0)	1060.85 (443.73)	0.49 (0.07)	0.02 (0.02)	0.0 (0.0)	44.54 (10.79)	33.47 (19.1)	69.29 (17.7)	42.63 (15.21)	11.77 (6.25)	41.15 (17.54)	68.08 (23.44)	23.85 (10.89)	5.24 (3.85)	0.21 (0.11)	0.44 (0.08)	0.27 (0.09)	0.08 (0.04)
7	1949	25.03 (11.68)	22.62 (10.79)	23.13 (10.93)	1639.71 (949.88)	814.8 (490.84)	824.92 (487.72)	0.51 (0.06)	0.04 (0.05)	0.0 (0.0)	38.58 (18.65)	3.28 (2.88)	8.3 (5.1)	6.86 (4.18)	6.6 (3.35)	4.77 (3.34)	8.68 (5.37)	5.23 (3.33)	4.45 (2.29)	0.12 (0.1)	0.3 (0.13)	0.25 (0.12)	0.33 (0.21)
8	2492	46.49 (15.72)	39.53 (14.59)	40.68 (16.7)	635.09 (438.7)	318.17 (218.97)	316.92 (226.44)	0.48 (0.1)	0.03 (0.03)	0.0 (0.0)	46.23 (15.74)	10.17 (6.3)	20.56 (9.44)	11.68 (6.65)	4.08 (3.4)	13.2 (7.28)	18.61 (9.57)	7.1 (4.86)	1.76 (1.93)	0.24 (0.17)	0.43 (0.12)	0.24 (0.12)	0.09 (0.07)
9	1	684.0 (0.0)	291.0 (0.0)	682.0 (0.0)	4715.0 (0.0)	1776.0 (0.0)	2939.0 (0.0)	0.62 (0.0)	0.0 (0.0)	0.0 (0.0)	20.75 (0.0)	320.0 (0.0)	307.0 (0.0)	38.0 (0.0)	19.0 (0.0)	338.0 (0.0)	309.0 (0.0)	20.0 (0.0)	15.0 (0.0)	0.47 (0.0)	0.45 (0.0)	0.06 (0.0)	0.03 (0.0)
Clustering 2: 7 Clusters																							
0	82838	4.64 (6.68)	3.95 (5.94)	3.81 (6.26)	91.75 (307.39)	45.84 (154.7)	45.91 (155.66)	0.4 (0.3)	0.04 (0.14)	0.0 (0.0)	33.13 (45.55)	1.16 (1.54)	1.91 (3.0)	1.09 (2.15)	0.49 (1.38)	1.23 (2.03)	1.64 (2.94)	0.7 (1.55)	0.25 (0.89)	0.38 (0.41)	0.39 (0.38)	0.17 (0.27)	0.06 (0.17)
1	3017	53.93 (23.26)	45.66 (20.66)	47.57 (22.95)	769.54 (552.56)	386.17 (279.53)	383.37 (280.64)	0.49 (0.1)	0.03 (0.03)	0.0 (0.0)	46.16 (15.36)	11.67 (8.68)	23.61 (12.35)	13.74 (8.57)	4.92 (4.04)	15.42 (10.47)	21.53 (12.04)	8.48 (6.17)	2.13 (2.21)	0.23 (0.16)	0.43 (0.12)	0.25 (0.12)	0.09 (0.07)
2	75	148.36 (64.05)	131.99 (54.77)	141.87 (62.01)	7775.28 (2155.99)	3917.76 (1069.05)	3857.52 (1163.07)	0.5 (0.03)	0.02 (0.02)	0.0 (0.0)	41.75 (9.94)	18.67 (14.89)	49.96 (31.08)	45.29 (22.73)	34.44 (9.26)	29.45 (21.23)	57.2 (33.53)	34.15 (15.43)	21.07 (6.51)	0.12 (0.06)	0.32 (0.08)	0.3 (0.06)	0.26 (0.1)
3	45	256.42 (95.82)	178.58 (71.4)	240.11 (95.89)	1872.98 (1305.51)	909.02 (667.15)	963.96 (672.43)	0.52 (0.07)	0.01 (0.01)	0.0 (0.0)	37.65 (10.47)	81.64 (50.34)	121.82 (49.12)	44.29 (35.38)	8.67 (7.43)	114.36 (53.51)	102.31 (66.21)	20.47 (16.85)	2.98 (3.61)	0.32 (0.14)	0.48 (0.11)	0.16 (0.1)	0.03 (0.03)
4	21	2.38 (0.58)	2.33 (0.56)	2.19 (0.79)	793.19 (797.51)	457.57 (504.82)	335.62 (324.73)	0.45 (0.12)	0.87 (0.21)	0.02 (0.01)	16.63 (14.22)	0.05 (0.21)	0.19 (0.39)	0.24 (0.43)	1.9 (0.61)	0.14 (0.35)	0.1 (0.29)	0.38 (0.49)	1.57 (0.66)	0.02 (0.07)	0.07 (0.15)	0.1 (0.19)	0.81 (0.21)
5	512	73.85 (26.65)	66.53 (22.64)	68.81 (25.23)	3806.88 (1411.26)	1894.22 (727.02)	1912.66 (735.73)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	44.65 (11.1)	9.66 (7.61)	25.32 (11.6)	21.31 (9.19)	17.56 (5.6)	14.06 (8.62)	27.4 (12.25)	16.45 (7.52)	10.9 (3.81)	0.12 (0.07)	0.34 (0.07)	0.29 (0.07)	0.25 (0.09)
6	131	157.16 (26.11)	132.88 (24.05)	138.32 (36.08)	2136.91 (878.33)	1076.06 (465.0)	1060.85 (443.73)	0.49 (0.07)	0.02 (0.02)	0.0 (0.0)	44.54 (10.79)	33.47 (19.1)	69.29 (17.7)	42.63 (15.21)	11.77 (6.25)	41.15 (17.54)	68.08 (23.44)	23.85 (10.89)	5.24 (3.85)	0.21 (0.11)	0.44 (0.08)	0.27 (0.09)	0.08 (0.04)
Clustering 3: 5 Clusters																							
0	587	83.37 (41.98)	74.89 (36.16)	78.14 (40.51)	4313.91 (2021.42)	2152.76 (1031.2)	2161.15 (1032.72)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	44.28 (11.0)	10.81 (9.37)	28.47 (17.56)	24.37 (14.27)	19.72 (8.37)	16.02 (12.2)	31.21 (19.33)	18.71 (10.71)	12.2 (5.44)	0.12 (0.07)	0.33 (0.08)	0.29 (0.07)	0.26 (0.09)
1	82838	4.64 (6.68)	3.95 (5.94)	3.81 (6.26)	91.75 (307.39)	45.84 (154.7)	45.91 (155.66)	0.4 (0.3)	0.04 (0.14)	0.0 (0.0)	33.13 (45.55)	1.16 (1.54)	1.91 (3.0)	1.09 (2.15)	0.49 (1.38)	1.23 (2.03)	1.64 (2.94)	0.7 (1.55)	0.25 (0.89)	0.38 (0.41)	0.39 (0.38)	0.17 (0.27)	0.06 (0.17)
2	176	182.54 (68.78)	144.56 (46.16)	164.35 (72.74)	2069.43 (1011.55)	1033.35 (529.2)	1036.07 (513.76)	0.5 (0.07)	0.02 (0.02)	0.0 (0.0)	42.78 (11.12)	45.79 (36.89)	82.72 (37.09)	43.05 (22.2)	10.98 (6.71)	59.87 (44.51)	76.83 (41.87)	22.99 (12.77)	4.66 (3.91)	0.24 (0.13)	0.45 (0.09)	0.24 (0.1)	0.07 (0.04)
3	3017	53.93 (23.26)	45.66 (20.66)	47.57 (22.95)	769.54 (552.56)	386.17 (279.53)	383.37 (280.64)	0.49 (0.1)	0.03 (0.03)	0.0 (0.0)	46.16 (15.36)	11.67 (8.68)	23.61 (12.35)	13.74 (8.57)	4.92 (4.04)	15.42 (10.47)	21.53 (12.04)	8.48 (6.17)	2.13 (2.21)	0.23 (0.16)	0.43 (0.12)	0.25 (0.12)	0.09 (0.07)
4	21	2.38 (0.58)	2.33 (0.56)	2.19 (0.79)	793.19 (797.51)	457.57 (504.82)	335.62 (324.73)	0.45 (0.12)	0.87 (0.21)	0.02 (0.01)	16.63 (14.22)	0.05 (0.21)	0.19 (0.39)	0.24 (0.43)	1.9 (0.61)	0.14 (0.35)	0.1 (0.29)	0.38 (0.49)	1.57 (0.66)	0.02 (0.07)	0.07 (0.15)	0.1 (0.19)	0.81 (0.21)
The entire data set																							
0	86639	7.25 (16.34)	6.17 (13.85)	6.17 (15.12)	148.14 (522.07)	74.07 (262.25)	74.07 (263.05)	0.41 (0.29)	0.04 (0.13)	0.0 (0.0)	33.68 (44.72)	1.68 (4.06)	3.01 (7.21)	1.77 (4.66)	0.79 (2.5)	1.94 (5.27)	2.68 (7.04)	1.13 (3.14)	0.41 (1.5)	0.38 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)

Table C.13: BIRCH on data from the 2nd month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	4563	46.08 (18.91)	39.45 (17.44)	41.41 (17.84)	839.09 (637.19)	415.34 (318.27)	423.75 (327.88)	0.5 (0.08)	0.04 (0.04)	0.0 (0.0)	51.11 (16.52)	8.98 (6.19)	19.66 (9.92)	12.35 (7.13)	5.1 (4.0)	12.42 (7.28)	18.59 (9.99)	7.94 (5.27)	2.47 (2.48)	0.2 (0.13)	0.42 (0.11)	0.26 (0.11)	0.11 (0.09)
1	616	44.87 (23.8)	41.26 (21.94)	41.37 (22.31)	4115.69 (1720.49)	2065.29 (896.05)	2050.41 (885.45)	0.5 (0.05)	0.04 (0.05)	0.0 (0.0)	44.32 (18.86)	5.44 (4.8)	14.59 (9.34)	12.36 (7.97)	12.49 (6.38)	7.89 (5.55)	15.25 (9.89)	9.9 (6.41)	8.33 (4.33)	0.11 (0.09)	0.3 (0.12)	0.25 (0.11)	0.34 (0.19)
2	364	128.74 (31.98)	113.8 (27.8)	119.96 (31.34)	3265.71 (1145.21)	1633.57 (611.51)	1632.14 (573.16)	0.5 (0.04)	0.03 (0.03)	0.0 (0.0)	49.67 (13.0)	18.61 (10.11)	49.83 (18.56)	40.49 (13.04)	19.81 (6.72)	27.52 (12.05)	54.32 (19.71)	28.27 (9.72)	9.85 (4.63)	0.14 (0.06)	0.38 (0.08)	0.32 (0.08)	0.16 (0.06)
3	78155	3.88 (5.03)	3.29 (4.49)	3.1 (4.67)	63.12 (190.66)	31.67 (96.91)	31.45 (97.3)	0.4 (0.3)	0.04 (0.14)	0.0 (0.0)	37.62 (48.3)	1.04 (1.33)	1.62 (2.41)	0.87 (1.66)	0.36 (0.98)	1.04 (1.66)	1.34 (2.28)	0.54 (1.19)	0.18 (0.62)	0.38 (0.41)	0.39 (0.38)	0.17 (0.28)	0.06 (0.17)
4	167	128.57 (33.66)	103.6 (40.63)	104.86 (32.92)	990.99 (544.52)	519.95 (296.18)	471.05 (266.13)	0.47 (0.11)	0.03 (0.03)	0.0 (0.0)	51.37 (14.84)	39.66 (23.1)	60.22 (18.01)	23.74 (12.6)	4.95 (4.07)	45.04 (15.87)	47.0 (19.74)	11.34 (7.79)	1.48 (1.71)	0.31 (0.15)	0.47 (0.11)	0.18 (0.08)	0.04 (0.03)
5	4	526.5 (38.36)	289.75 (48.5)	519.25 (38.13)	2538.25 (1369.67)	1201.75 (811.02)	1336.5 (560.0)	0.55 (0.05)	0.01 (0.0)	0.0 (0.0)	32.9 (6.15)	232.25 (44.71)	231.75 (23.77)	54.25 (21.21)	8.25 (7.46)	332.25 (28.44)	166.25 (11.12)	18.25 (10.57)	2.5 (3.28)	0.44 (0.07)	0.44 (0.05)	0.1 (0.04)	0.02 (0.01)
6	83	108.63 (31.76)	99.94 (29.69)	103.23 (30.42)	8572.57 (3063.0)	4306.61 (1706.58)	4265.95 (1445.39)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	48.77 (10.08)	10.99 (6.2)	32.78 (14.45)	31.99 (11.56)	32.87 (8.69)	17.41 (8.49)	37.45 (14.25)	26.17 (9.44)	22.2 (5.85)	0.1 (0.05)	0.29 (0.07)	0.29 (0.05)	0.32 (0.09)
7	83	2.72 (1.06)	2.54 (1.06)	2.23 (1.24)	443.27 (566.65)	238.84 (310.88)	204.42 (289.32)	0.45 (0.19)	0.78 (0.29)	0.01 (0.0)	32.32 (28.11)	0.16 (0.4)	0.28 (0.59)	0.71 (0.7)	1.58 (0.89)	0.13 (0.34)	0.3 (0.48)	0.7 (0.85)	1.1 (0.83)	0.04 (0.11)	0.09 (0.2)	0.29 (0.28)	0.58 (0.28)
8	32	256.19 (54.03)	221.03 (46.67)	236.03 (49.34)	6901.88 (2520.76)	3459.09 (1297.56)	3442.78 (1304.6)	0.5 (0.04)	0.02 (0.01)	0.0 (0.0)	46.78 (8.77)	41.88 (26.5)	102.69 (33.69)	74.66 (17.09)	36.97 (13.14)	58.12 (22.67)	109.97 (36.5)	49.19 (10.83)	18.75 (9.66)	0.16 (0.07)	0.39 (0.07)	0.3 (0.06)	0.15 (0.06)
9	46	226.74 (47.76)	158.54 (51.65)	210.2 (48.26)	1861.33 (1275.18)	898.39 (602.25)	962.93 (692.36)	0.52 (0.06)	0.02 (0.02)	0.0 (0.0)	42.61 (13.57)	75.2 (27.98)	105.89 (28.98)	36.76 (19.25)	8.89 (7.37)	107.2 (35.52)	81.04 (33.2)	18.43 (12.05)	3.52 (4.21)	0.34 (0.12)	0.47 (0.08)	0.16 (0.07)	0.04 (0.03)
Clustering 2: 7 Clusters																							
0	78238	3.88 (5.02)	3.29 (4.49)	3.1 (4.66)	63.52 (191.85)	31.89 (97.62)	31.63 (97.87)	0.4 (0.3)	0.04 (0.14)	0.0 (0.0)	37.61 (48.28)	1.04 (1.33)	1.62 (2.41)	0.87 (1.66)	0.36 (0.98)	1.04 (1.66)	1.34 (2.28)	0.54 (1.19)	0.18 (0.62)	0.38 (0.41)	0.39 (0.38)	0.17 (0.28)	0.06 (0.17)
1	115	149.69 (76.9)	133.63 (64.71)	140.18 (69.91)	8107.68 (3016.62)	4070.78 (1647.65)	4036.9 (1455.17)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	48.21 (9.78)	19.58 (20.36)	52.23 (38.05)	43.86 (23.31)	34.01 (10.29)	28.74 (22.98)	57.63 (39.67)	32.57 (14.26)	21.24 (7.28)	0.11 (0.06)	0.32 (0.08)	0.29 (0.05)	0.27 (0.11)
2	213	149.77 (54.89)	115.46 (48.8)	127.61 (56.84)	1178.95 (843.74)	601.68 (413.95)	577.28 (447.24)	0.48 (0.11)	0.03 (0.03)	0.0 (0.0)	49.48 (15.01)	47.33 (28.31)	70.08 (28.09)	26.55 (15.27)	5.8 (5.23)	58.46 (33.53)	54.35 (27.2)	12.87 (9.35)	1.92 (2.61)	0.32 (0.15)	0.47 (0.1)	0.17 (0.08)	0.04 (0.03)
3	4563	46.08 (18.91)	39.45 (17.44)	41.41 (17.84)	839.09 (637.19)	415.34 (318.27)	423.75 (327.88)	0.5 (0.08)	0.04 (0.04)	0.0 (0.0)	51.11 (16.52)	8.98 (6.19)	19.66 (9.92)	12.35 (7.13)	5.1 (4.0)	12.42 (7.28)	18.59 (9.99)	7.94 (5.27)	2.47 (2.48)	0.2 (0.13)	0.42 (0.11)	0.26 (0.11)	0.11 (0.09)
4	616	44.87 (23.8)	41.26 (21.94)	41.37 (22.31)	4115.69 (1720.49)	2065.29 (896.05)	2050.41 (885.45)	0.5 (0.05)	0.04 (0.05)	0.0 (0.0)	44.32 (18.86)	5.44 (4.8)	14.59 (9.34)	12.36 (7.97)	12.49 (6.38)	7.89 (5.55)	15.25 (9.89)	9.9 (6.41)	8.33 (4.33)	0.11 (0.09)	0.3 (0.12)	0.25 (0.11)	0.34 (0.19)
5	4	526.5 (38.36)	289.75 (48.5)	519.25 (38.13)	2538.25 (1369.67)	1201.75 (811.02)	1336.5 (560.0)	0.55 (0.05)	0.01 (0.0)	0.0 (0.0)	32.9 (6.15)	232.25 (44.71)	231.75 (23.77)	54.25 (21.21)	8.25 (7.46)	332.25 (28.44)	166.25 (11.12)	18.25 (10.57)	2.5 (3.28)	0.44 (0.07)	0.44 (0.05)	0.1 (0.04)	0.02 (0.01)
6	364	128.74 (31.98)	113.8 (27.8)	119.96 (31.34)	3265.71 (1145.21)	1633.57 (611.51)	1632.14 (573.16)	0.5 (0.04)	0.03 (0.03)	0.0 (0.0)	49.67 (13.0)	18.61 (10.11)	49.83 (18.56)	40.49 (13.04)	19.81 (6.72)	27.52 (12.05)	54.32 (19.71)	28.27 (9.72)	9.85 (4.63)	0.14 (0.06)	0.38 (0.08)	0.32 (0.08)	0.16 (0.06)
Clustering 3: 5 Clusters																							
0	217	156.71 (74.51)	118.68 (54.14)	134.83 (77.29)	1204.01 (875.67)	612.74 (432.25)	591.27 (461.02)	0.48 (0.11)	0.03 (0.03)	0.0 (0.0)	49.18 (15.06)	50.74 (37.97)	73.06 (35.46)	27.06 (15.84)	5.85 (5.29)	63.51 (49.74)	56.41 (30.9)	12.97 (9.4)	1.93 (2.62)	0.32 (0.15)	0.47 (0.1)	0.17 (0.08)	0.04 (0.03)
1	78238	3.88 (5.02)	3.29 (4.49)	3.1 (4.66)	63.52 (191.85)	31.89 (97.62)	31.63 (97.87)	0.4 (0.3)	0.04 (0.14)	0.0 (0.0)	37.61 (48.28)	1.04 (1.33)	1.62 (2.41)	0.87 (1.66)	0.36 (0.98)	1.04 (1.66)	1.34 (2.28)	0.54 (1.19)	0.18 (0.62)	0.38 (0.41)	0.39 (0.38)	0.17 (0.28)	0.06 (0.17)
2	479	133.77 (47.72)	118.56 (40.8)	124.82 (44.66)	4428.19 (2731.07)	2218.7 (1421.14)	2209.48 (1346.5)	0.5 (0.04)	0.03 (0.03)	0.0 (0.0)	49.32 (12.32)	18.84 (13.32)	50.41 (24.71)	41.3 (16.18)	23.22 (9.83)	27.82 (15.41)	55.12 (25.98)	29.3 (11.14)	12.58 (7.26)	0.13 (0.06)	0.37 (0.08)	0.31 (0.08)	0.19 (0.09)
3	4563	46.08 (18.91)	39.45 (17.44)	41.41 (17.84)	839.09 (637.19)	415.34 (318.27)	423.75 (327.88)	0.5 (0.08)	0.04 (0.04)	0.0 (0.0)	51.11 (16.52)	8.98 (6.19)	19.66 (9.92)	12.35 (7.13)	5.1 (4.0)	12.42 (7.28)	18.59 (9.99)	7.94 (5.27)	2.47 (2.48)	0.2 (0.13)	0.42 (0.11)	0.26 (0.11)	0.11 (0.09)
4	616	44.87 (23.8)	41.26 (21.94)	41.37 (22.31)	4115.69 (1720.49)	2065.29 (896.05)	2050.41 (885.45)	0.5 (0.05)	0.04 (0.05)	0.0 (0.0)	44.32 (18.86)	5.44 (4.8)	14.59 (9.34)	12.36 (7.97)	12.49 (6.38)	7.89 (5.55)	15.25 (9.89)	9.9 (6.41)	8.33 (4.33)	0.11 (0.09)	0.3 (0.12)	0.25 (0.11)	0.34 (0.19)
The entire data set																							
0	84113	7.6 (18.05)	6.49 (15.4)	6.49 (16.61)	163.07 (614.17)	81.54 (310.03)	81.54 (308.01)	0.41 (0.29)	0.04 (0.14)	0.0 (0.0)	38.49 (46.88)	1.73 (4.46)	3.15 (7.83)	1.87 (5.11)	0.85 (2.79)	2.02 (5.83)	2.82 (7.54)	1.21 (3.51)	0.44 (1.65)	0.37 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.17)

Table C.14: BIRCH on data from the 3rd month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	276	104.42 (52.0)	95.8 (47.58)	96.62 (49.62)	7697.96 (2577.23)	3878.68 (1384.29)	3819.28 (1274.79)	0.5 (0.04)	0.04 (0.03)	0.0 (0.0)	55.84 (15.96)	12.75 (9.53)	34.34 (20.03)	29.85 (18.43)	27.48 (12.0)	18.33 (11.74)	37.0 (22.31)	23.57 (14.99)	17.71 (6.82)	0.12 (0.07)	0.32 (0.07)	0.27 (0.07)	0.29 (0.11)
1	295	143.9 (31.63)	123.91 (29.6)	134.12 (29.82)	2459.59 (1172.83)	1220.68 (588.86)	1238.9 (606.23)	0.5 (0.04)	0.04 (0.04)	0.0 (0.0)	58.86 (17.3)	23.35 (9.85)	61.06 (16.17)	43.59 (13.24)	15.89 (8.32)	35.62 (11.04)	63.41 (16.81)	28.58 (10.98)	6.51 (4.76)	0.16 (0.06)	0.43 (0.08)	0.3 (0.06)	0.11 (0.05)
2	73190	3.4 (4.12)	2.86 (3.66)	2.63 (3.8)	40.0 (119.62)	20.34 (63.26)	19.66 (59.51)	0.4 (0.31)	0.04 (0.15)	0.0 (0.0)	44.67 (67.63)	0.99 (1.29)	1.43 (2.12)	0.73 (1.4)	0.24 (0.62)	0.94 (1.51)	1.16 (1.99)	0.43 (0.95)	0.1 (0.37)	0.39 (0.42)	0.39 (0.39)	0.17 (0.28)	0.05 (0.16)
3	67	234.24 (63.38)	159.06 (61.24)	218.24 (63.15)	2226.82 (1248.39)	1064.15 (625.78)	1162.67 (647.28)	0.53 (0.08)	0.02 (0.02)	0.0 (0.0)	47.71 (13.21)	79.42 (35.01)	105.6 (40.63)	37.81 (20.4)	11.42 (7.97)	105.97 (48.4)	86.22 (31.13)	21.37 (14.23)	4.67 (3.95)	0.34 (0.12)	0.45 (0.1)	0.16 (0.08)	0.05 (0.03)
4	32	312.41 (62.95)	261.75 (59.19)	290.88 (62.82)	6096.62 (2844.73)	3012.72 (1514.85)	3083.91 (1392.02)	0.51 (0.04)	0.02 (0.02)	0.0 (0.0)	47.33 (13.7)	55.88 (27.7)	122.94 (35.93)	98.47 (42.54)	35.12 (20.31)	73.06 (31.62)	141.78 (54.77)	59.97 (24.57)	16.06 (11.81)	0.18 (0.08)	0.4 (0.1)	0.31 (0.1)	0.11 (0.07)
5	2	902.5 (125.5)	607.0 (83.0)	863.5 (137.5)	4736.5 (2944.5)	2239.5 (1342.5)	2497.0 (1602.0)	0.52 (0.02)	0.01 (0.0)	0.0 (0.0)	31.16 (2.6)	292.0 (32.0)	498.0 (8.0)	94.0 (84.0)	18.5 (17.5)	537.0 (79.0)	278.0 (171.0)	40.0 (37.0)	8.5 (8.5)	0.32 (0.01)	0.56 (0.09)	0.09 (0.08)	0.02 (0.02)
6	2884	48.2 (25.48)	39.91 (22.15)	41.71 (23.34)	494.82 (388.73)	250.22 (200.05)	244.6 (195.19)	0.49 (0.09)	0.04 (0.04)	0.0 (0.0)	59.67 (22.15)	12.0 (9.84)	22.35 (12.99)	10.99 (7.49)	2.85 (2.68)	15.53 (11.17)	19.1 (11.9)	6.04 (4.81)	1.04 (1.35)	0.25 (0.14)	0.46 (0.11)	0.23 (0.11)	0.06 (0.05)
7	4058	30.1 (15.6)	27.07 (14.22)	27.63 (14.49)	1266.12 (950.46)	626.63 (486.03)	639.49 (481.15)	0.51 (0.06)	0.05 (0.05)	0.0 (0.0)	54.83 (23.73)	4.17 (3.42)	11.16 (7.05)	8.6 (5.81)	6.16 (3.26)	6.29 (4.37)	11.15 (7.34)	6.58 (4.35)	3.61 (2.28)	0.13 (0.09)	0.35 (0.12)	0.27 (0.11)	0.24 (0.14)
8	537	84.37 (20.85)	75.78 (18.57)	78.15 (18.37)	2957.13 (987.94)	1461.84 (501.49)	1495.29 (524.12)	0.51 (0.04)	0.05 (0.04)	0.0 (0.0)	57.06 (15.83)	11.33 (7.77)	30.5 (11.73)	25.35 (7.38)	17.18 (4.54)	16.95 (8.89)	32.27 (10.77)	18.95 (5.81)	9.98 (3.7)	0.13 (0.07)	0.35 (0.08)	0.3 (0.07)	0.21 (0.07)
9	12	2.42 (0.49)	2.33 (0.62)	2.0 (0.71)	1263.67 (1364.91)	730.5 (829.63)	533.17 (601.07)	0.46 (0.15)	0.82 (0.27)	0.02 (0.01)	46.86 (50.74)	0.17 (0.37)	0.08 (0.28)	0.17 (0.37)	2.0 (0.58)	0.0 (0.0)	0.17 (0.37)	0.25 (0.43)	1.58 (0.76)	0.06 (0.12)	0.04 (0.14)	0.07 (0.16)	0.83 (0.2)
Clustering 2: 7 Clusters																							
0	77248	4.8 (8.02)	4.13 (7.25)	3.94 (7.47)	104.41 (368.56)	52.19 (185.73)	52.22 (186.11)	0.4 (0.3)	0.04 (0.15)	0.0 (0.0)	45.2 (66.1)	1.16 (1.64)	1.95 (3.41)	1.15 (2.59)	0.55 (1.63)	1.22 (2.14)	1.68 (3.4)	0.76 (1.93)	0.29 (1.01)	0.38 (0.41)	0.39 (0.38)	0.17 (0.28)	0.06 (0.17)
1	99	259.51 (73.05)	192.25 (77.31)	241.72 (71.61)	3477.67 (2635.6)	1693.99 (1355.49)	1783.68 (1310.47)	0.53 (0.07)	0.02 (0.02)	0.0 (0.0)	47.59 (13.37)	71.81 (34.62)	111.2 (40.0)	57.41 (40.88)	19.08 (17.3)	95.33 (46.32)	104.18 (47.97)	33.85 (25.65)	8.35 (9.16)	0.29 (0.13)	0.43 (0.1)	0.21 (0.11)	0.07 (0.06)
2	832	105.47 (38.03)	92.85 (32.61)	97.99 (35.36)	2780.72 (1083.66)	1376.34 (546.42)	1404.38 (568.03)	0.5 (0.04)	0.05 (0.04)	0.0 (0.0)	57.7 (16.39)	15.59 (10.32)	41.34 (19.88)	31.82 (13.17)	16.72 (6.18)	23.57 (13.19)	43.31 (19.92)	22.36 (9.26)	8.75 (4.43)	0.14 (0.07)	0.38 (0.08)	0.3 (0.07)	0.18 (0.08)
3	276	104.42 (52.0)	95.8 (47.58)	96.62 (49.62)	7697.96 (2577.23)	3878.68 (1384.29)	3819.28 (1274.79)	0.5 (0.04)	0.04 (0.03)	0.0 (0.0)	55.84 (15.96)	12.75 (9.53)	34.34 (20.03)	29.85 (18.43)	27.48 (12.0)	18.33 (11.74)	37.0 (22.31)	23.57 (14.99)	17.71 (6.82)	0.12 (0.07)	0.32 (0.07)	0.27 (0.07)	0.29 (0.11)
4	12	2.42 (0.49)	2.33 (0.62)	2.0 (0.71)	1263.67 (1364.91)	730.5 (829.63)	533.17 (601.07)	0.46 (0.15)	0.82 (0.27)	0.02 (0.01)	46.86 (50.74)	0.17 (0.37)	0.08 (0.28)	0.17 (0.37)	2.0 (0.58)	0.0 (0.0)	0.17 (0.37)	0.25 (0.43)	1.58 (0.76)	0.06 (0.12)	0.04 (0.14)	0.07 (0.16)	0.83 (0.2)
5	2	902.5 (125.5)	607.0 (83.0)	863.5 (137.5)	4736.5 (2944.5)	2239.5 (1342.5)	2497.0 (1602.0)	0.52 (0.02)	0.01 (0.0)	0.0 (0.0)	31.16 (2.6)	292.0 (32.0)	498.0 (8.0)	94.0 (84.0)	18.5 (17.5)	537.0 (79.0)	278.0 (171.0)	40.0 (37.0)	8.5 (8.5)	0.32 (0.01)	0.56 (0.09)	0.09 (0.08)	0.02 (0.02)
6	2884	48.2 (25.48)	39.91 (22.15)	41.71 (23.34)	494.82 (388.73)	250.22 (200.05)	244.6 (195.19)	0.49 (0.09)	0.04 (0.04)	0.0 (0.0)	59.67 (22.15)	12.0 (9.84)	22.35 (12.99)	10.99 (7.49)	2.85 (2.68)	15.53 (11.17)	19.1 (11.9)	6.04 (4.81)	1.04 (1.35)	0.25 (0.14)	0.46 (0.11)	0.23 (0.11)	0.06 (0.05)
Clustering 3: 5 Clusters																							
0	101	272.24 (116.48)	200.47 (96.61)	254.03 (113.6)	3502.59 (2647.88)	1704.79 (1357.36)	1797.8 (1320.61)	0.53 (0.07)	0.02 (0.02)	0.0 (0.0)	47.26 (13.44)	76.17 (46.22)	118.86 (66.89)	58.14 (42.48)	19.07 (17.3)	104.08 (77.54)	107.62 (58.49)	33.97 (25.94)	8.36 (9.15)	0.29 (0.13)	0.43 (0.1)	0.21 (0.11)	0.07 (0.06)
1	3716	61.02 (37.38)	51.76 (33.25)	54.31 (35.4)	1006.62 (1134.99)	502.35 (564.15)	504.27 (579.26)	0.5 (0.08)	0.04 (0.04)	0.0 (0.0)	59.23 (21.02)	12.81 (10.06)	26.6 (16.79)	15.66 (12.56)	5.96 (6.89)	17.33 (12.13)	24.52 (17.34)	9.69 (9.13)	2.77 (4.02)	0.23 (0.14)	0.44 (0.11)	0.25 (0.11)	0.09 (0.07)
2	77248	4.8 (8.02)	4.13 (7.25)	3.94 (7.47)	104.41 (368.56)	52.19 (185.73)	52.22 (186.11)	0.4 (0.3)	0.04 (0.15)	0.0 (0.0)	45.2 (66.1)	1.16 (1.64)	1.95 (3.41)	1.15 (2.59)	0.55 (1.63)	1.22 (2.14)	1.68 (3.4)	0.76 (1.93)	0.29 (1.01)	0.38 (0.41)	0.39 (0.38)	0.17 (0.28)	0.06 (0.17)
3	276	104.42 (52.0)	95.8 (47.58)	96.62 (49.62)	7697.96 (2577.23)	3878.68 (1384.29)	3819.28 (1274.79)	0.5 (0.04)	0.04 (0.03)	0.0 (0.0)	55.84 (15.96)	12.75 (9.53)	34.34 (20.03)	29.85 (18.43)	27.48 (12.0)	18.33 (11.74)	37.0 (22.31)	23.57 (14.99)	17.71 (6.82)	0.12 (0.07)	0.32 (0.07)	0.27 (0.07)	0.29 (0.11)
4	12	2.42 (0.49)	2.33 (0.62)	2.0 (0.71)	1263.67 (1364.91)	730.5 (829.63)	533.17 (601.07)	0.46 (0.15)	0.82 (0.27)	0.02 (0.01)	46.86 (50.74)	0.17 (0.37)	0.08 (0.28)	0.17 (0.37)	2.0 (0.58)	0.0 (0.0)	0.17 (0.37)	0.25 (0.43)	1.58 (0.76)	0.06 (0.12)	0.04 (0.14)	0.07 (0.16)	0.83 (0.2)
The entire data set																							
0	81353	8.04 (20.18)	6.86 (17.09)	6.86 (18.67)	175.77 (679.4)	87.89 (342.49)	87.89 (340.55)	0.41 (0.29)	0.04 (0.14)	0.0 (0.0)	45.88 (64.64)	1.82 (4.83)	3.33 (8.79)	1.98 (5.72)	0.91 (3.11)	2.14 (6.64)	2.98 (8.42)	1.28 (3.93)	0.47 (1.82)	0.37 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.17)

Table C.15: BIRCH on data from the 4th month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	63079	3.57 (4.53)	3.03 (4.05)	2.82 (4.2)	45.07 (112.46)	22.74 (57.97)	22.33 (56.88)	0.4 (0.3)	0.04 (0.14)	0.0 (0.0)	36.81 (64.4)	1.0 (1.3)	1.51 (2.29)	0.79 (1.57)	0.27 (0.71)	0.98 (1.58)	1.23 (2.16)	0.48 (1.08)	0.12 (0.42)	0.39 (0.41)	0.39 (0.39)	0.17 (0.28)	0.05 (0.17)
1	312	112.63 (30.32)	89.27 (26.2)	103.36 (27.55)	1303.2 (704.11)	649.38 (355.65)	653.83 (363.36)	0.5 (0.06)	0.03 (0.03)	0.0 (0.0)	47.13 (14.97)	27.18 (19.41)	51.57 (15.03)	26.72 (10.75)	7.16 (4.54)	38.32 (20.88)	47.45 (15.62)	14.79 (7.67)	2.8 (2.58)	0.23 (0.11)	0.46 (0.09)	0.24 (0.08)	0.06 (0.04)
2	39	172.64 (78.05)	154.28 (65.42)	162.31 (73.29)	9732.36 (3777.73)	4884.9 (1897.04)	4847.46 (1967.49)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	45.22 (10.55)	23.31 (17.31)	56.28 (33.24)	52.49 (25.1)	40.56 (12.36)	31.46 (19.95)	65.15 (35.91)	40.62 (21.76)	25.08 (7.35)	0.12 (0.06)	0.31 (0.06)	0.3 (0.06)	0.26 (0.09)
3	46	219.26 (67.56)	181.59 (53.25)	206.93 (63.49)	2626.78 (977.73)	1318.11 (543.76)	1308.67 (462.41)	0.5 (0.05)	0.02 (0.02)	0.0 (0.0)	39.88 (12.37)	38.54 (29.4)	96.89 (37.63)	69.48 (23.88)	14.35 (8.82)	56.22 (33.26)	108.07 (41.36)	37.91 (19.13)	4.74 (3.74)	0.16 (0.08)	0.44 (0.11)	0.33 (0.11)	0.07 (0.04)
4	2772	29.83 (11.6)	26.58 (10.7)	27.45 (10.9)	1263.99 (981.36)	625.41 (498.24)	638.58 (497.39)	0.51 (0.05)	0.05 (0.05)	0.0 (0.0)	46.65 (23.64)	4.39 (3.27)	10.98 (5.69)	8.44 (4.52)	6.03 (3.44)	6.42 (3.91)	11.17 (5.8)	6.19 (3.52)	3.66 (2.5)	0.14 (0.09)	0.35 (0.12)	0.27 (0.1)	0.23 (0.16)
5	3	880.33 (174.27)	505.0 (195.67)	848.67 (138.0)	2761.33 (1033.72)	1296.0 (714.49)	1465.33 (333.31)	0.56 (0.08)	0.01 (0.0)	0.0 (0.0)	25.58 (4.08)	378.33 (42.03)	450.67 (156.42)	47.0 (14.76)	4.33 (4.78)	625.33 (143.79)	210.33 (14.06)	10.67 (5.44)	2.33 (2.05)	0.44 (0.09)	0.5 (0.1)	0.05 (0.01)	0.0 (0.0)
6	9	2.11 (0.31)	2.11 (0.31)	2.0 (0.47)	829.22 (566.06)	416.67 (289.83)	412.56 (284.07)	0.5 (0.05)	0.96 (0.1)	0.03 (0.01)	18.35 (38.15)	0.0 (0.0)	0.11 (0.31)	0.11 (0.31)	1.89 (0.31)	0.0 (0.0)	0.0 (0.0)	0.22 (0.42)	1.78 (0.42)	0.0 (0.0)	0.06 (0.16)	0.04 (0.1)	0.91 (0.18)
7	279	91.8 (26.51)	83.04 (24.13)	85.67 (24.71)	3227.96 (957.97)	1595.18 (506.62)	1632.78 (494.86)	0.51 (0.04)	0.04 (0.03)	0.0 (0.0)	48.84 (13.28)	11.57 (7.6)	33.33 (13.73)	28.28 (9.48)	18.63 (5.0)	17.74 (8.61)	36.17 (13.96)	20.98 (7.23)	10.77 (3.63)	0.12 (0.06)	0.35 (0.07)	0.31 (0.07)	0.21 (0.07)
8	110	78.13 (30.95)	72.15 (28.86)	72.05 (30.45)	6697.15 (1299.2)	3409.35 (725.76)	3287.8 (661.44)	0.49 (0.04)	0.03 (0.02)	0.0 (0.0)	48.26 (15.04)	9.42 (8.44)	24.86 (12.96)	21.39 (11.25)	22.45 (8.6)	13.01 (7.27)	26.95 (14.85)	16.61 (8.9)	15.48 (5.92)	0.12 (0.1)	0.31 (0.09)	0.26 (0.07)	0.31 (0.12)
9	1833	46.39 (18.18)	38.8 (18.12)	40.81 (17.56)	606.51 (533.11)	305.32 (271.88)	301.18 (268.45)	0.5 (0.09)	0.04 (0.04)	0.0 (0.0)	48.04 (20.1)	10.75 (6.83)	20.38 (9.0)	11.46 (7.76)	3.8 (3.95)	14.15 (6.83)	18.24 (9.89)	6.81 (5.86)	1.62 (2.1)	0.26 (0.16)	0.44 (0.12)	0.23 (0.11)	0.07 (0.06)
Clustering 2: 7 Clusters																							
0	358	126.33 (51.58)	101.13 (43.79)	116.67 (48.79)	1473.27 (866.64)	735.3 (445.31)	737.97 (436.53)	0.5 (0.06)	0.03 (0.03)	0.0 (0.0)	46.2 (14.86)	28.64 (21.3)	57.4 (24.68)	32.22 (19.46)	8.08 (5.81)	40.62 (23.62)	55.24 (29.05)	17.76 (12.58)	3.05 (2.84)	0.22 (0.11)	0.46 (0.09)	0.25 (0.09)	0.07 (0.04)
1	389	87.94 (28.51)	79.96 (26.02)	81.82 (27.16)	4208.96 (1891.14)	2108.18 (1000.26)	2100.78 (924.6)	0.5 (0.04)	0.04 (0.03)	0.0 (0.0)	48.68 (13.8)	10.96 (7.9)	30.94 (14.05)	26.33 (10.48)	19.71 (6.47)	16.4 (8.52)	33.57 (14.81)	19.74 (7.99)	12.11 (4.89)	0.12 (0.07)	0.34 (0.08)	0.3 (0.07)	0.24 (0.09)
2	39	172.64 (78.05)	154.28 (65.42)	162.31 (73.29)	9732.36 (3777.73)	4884.9 (1897.04)	4847.46 (1967.49)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	45.22 (10.55)	23.31 (17.31)	56.28 (33.24)	52.49 (25.1)	40.56 (12.36)	31.46 (19.95)	65.15 (35.91)	40.62 (21.76)	25.08 (7.35)	0.12 (0.06)	0.31 (0.06)	0.3 (0.06)	0.26 (0.09)
3	63079	3.57 (4.53)	3.03 (4.05)	2.82 (4.2)	45.07 (112.46)	22.74 (57.97)	22.33 (56.88)	0.4 (0.3)	0.04 (0.14)	0.0 (0.0)	36.81 (64.4)	1.0 (1.3)	1.51 (2.29)	0.79 (1.57)	0.27 (0.71)	0.98 (1.58)	1.23 (2.16)	0.48 (1.08)	0.12 (0.42)	0.39 (0.41)	0.39 (0.39)	0.17 (0.28)	0.05 (0.17)
4	4605	36.42 (16.68)	31.45 (15.34)	32.77 (15.4)	1002.28 (892.43)	498.0 (451.0)	504.28 (452.64)	0.51 (0.07)	0.04 (0.04)	0.0 (0.0)	47.21 (22.31)	6.92 (5.89)	14.72 (8.53)	9.64 (6.2)	5.14 (3.81)	9.5 (6.48)	13.99 (8.44)	6.44 (4.61)	2.85 (2.55)	0.19 (0.14)	0.39 (0.13)	0.26 (0.11)	0.17 (0.15)
5	3	880.33 (174.27)	505.0 (195.67)	848.67 (138.0)	2761.33 (1033.72)	1296.0 (714.49)	1465.33 (333.31)	0.56 (0.08)	0.01 (0.0)	0.0 (0.0)	25.58 (4.08)	378.33 (42.03)	450.67 (156.42)	47.0 (14.76)	4.33 (4.78)	625.33 (143.79)	210.33 (14.06)	10.67 (5.44)	2.33 (2.05)	0.44 (0.09)	0.5 (0.1)	0.05 (0.01)	0.0 (0.0)
6	9	2.11 (0.31)	2.11 (0.31)	2.0 (0.47)	829.22 (566.06)	416.67 (289.83)	412.56 (284.07)	0.5 (0.05)	0.96 (0.1)	0.03 (0.01)	18.35 (38.15)	0.0 (0.0)	0.11 (0.31)	0.11 (0.31)	1.89 (0.31)	0.0 (0.0)	0.0 (0.0)	0.22 (0.42)	1.78 (0.42)	0.0 (0.0)	0.06 (0.16)	0.04 (0.1)	0.91 (0.18)
Clustering 3: 5 Clusters																							
0	428	95.65 (43.45)	86.73 (38.25)	89.15 (41.19)	4712.26 (2660.37)	2361.2 (1369.61)	2351.06 (1324.58)	0.5 (0.04)	0.04 (0.03)	0.0 (0.0)	48.36 (13.58)	12.08 (9.83)	33.25 (18.25)	28.71 (14.63)	21.61 (9.38)	17.78 (11.0)	36.44 (19.99)	21.64 (11.71)	13.29 (6.37)	0.12 (0.07)	0.34 (0.08)	0.3 (0.07)	0.24 (0.09)
1	63088	3.57 (4.53)	3.03 (4.05)	2.82 (4.2)	45.19 (113.04)	22.8 (58.26)	22.39 (57.17)	0.4 (0.3)	0.04 (0.14)	0.0 (0.0)	36.8 (64.4)	1.0 (1.3)	1.51 (2.29)	0.79 (1.57)	0.27 (0.71)	0.98 (1.58)	1.23 (2.16)	0.48 (1.08)	0.12 (0.42)	0.39 (0.41)	0.39 (0.39)	0.17 (0.28)	0.05 (0.17)
2	3	880.33 (174.27)	505.0 (195.67)	848.67 (138.0)	2761.33 (1033.72)	1296.0 (714.49)	1465.33 (333.31)	0.56 (0.08)	0.01 (0.0)	0.0 (0.0)	25.58 (4.08)	378.33 (42.03)	450.67 (156.42)	47.0 (14.76)	4.33 (4.78)	625.33 (143.79)	210.33 (14.06)	10.67 (5.44)	2.33 (2.05)	0.44 (0.09)	0.5 (0.1)	0.05 (0.01)	0.0 (0.0)
3	358	126.33 (51.58)	101.13 (43.79)	116.67 (48.79)	1473.27 (866.64)	735.3 (445.31)	737.97 (436.53)	0.5 (0.06)	0.03 (0.03)	0.0 (0.0)	46.2 (14.86)	28.64 (21.3)	57.4 (24.68)	32.22 (19.46)	8.08 (5.81)	40.62 (23.62)	55.24 (29.05)	17.76 (12.58)	3.05 (2.84)	0.22 (0.11)	0.46 (0.09)	0.25 (0.09)	0.07 (0.04)
4	4605	36.42 (16.68)	31.45 (15.34)	32.77 (15.4)	1002.28 (892.43)	498.0 (451.0)	504.28 (452.64)	0.51 (0.07)	0.04 (0.04)	0.0 (0.0)	47.21 (22.31)	6.92 (5.89)	14.72 (8.53)	9.64 (6.2)	5.14 (3.81)	9.5 (6.48)	13.99 (8.44)	6.44 (4.61)	2.85 (2.55)	0.19 (0.14)	0.39 (0.13)	0.26 (0.11)	0.17 (0.15)
The entire data set																							
0	68482	7.04 (17.05)	6.0 (14.23)	6.0 (15.87)	146.3 (558.59)	73.15 (281.31)	73.15 (280.09)	0.41 (0.29)	0.04 (0.13)	0.0 (0.0)	37.62 (62.16)	1.62 (4.46)	2.91 (7.57)	1.73 (4.76)	0.77 (2.58)	1.89 (6.39)	2.6 (7.07)	1.1 (3.25)	0.4 (1.57)	0.37 (0.4)	0.39 (0.37)	0.18 (0.27)	0.06 (0.17)

Table C.16: BIRCH on data from the 5th month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
Clustering 1: 10 Clusters																							
0	83271	5.98 (10.87)	5.11 (9.53)	4.95 (9.85)	104.38 (331.51)	52.29 (168.67)	52.09 (166.89)	0.4 (0.3)	0.04 (0.14)	0.0 (0.0)	49.3 (68.18)	1.48 (2.74)	2.51 (4.94)	1.42 (3.22)	0.57 (1.59)	1.61 (3.39)	2.17 (4.64)	0.89 (2.2)	0.28 (0.92)	0.38 (0.41)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)
1	157	69.75 (34.14)	64.25 (30.5)	64.27 (31.92)	7935.35 (1669.28)	3949.21 (957.84)	3986.14 (883.27)	0.5 (0.05)	0.04 (0.03)	0.0 (0.0)	57.89 (19.15)	8.58 (7.37)	21.72 (13.67)	18.74 (11.19)	20.71 (9.61)	12.22 (8.56)	22.8 (14.33)	14.92 (8.81)	14.34 (6.55)	0.12 (0.07)	0.3 (0.1)	0.26 (0.08)	0.32 (0.11)
2	167	199.01 (56.45)	171.83 (46.96)	188.11 (55.11)	4622.72 (1439.03)	2298.5 (776.34)	2324.23 (715.51)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	56.09 (12.94)	30.88 (19.12)	80.46 (30.88)	59.38 (20.74)	28.29 (8.5)	48.99 (26.59)	84.63 (32.81)	40.56 (13.71)	13.93 (6.23)	0.15 (0.06)	0.4 (0.07)	0.3 (0.07)	0.15 (0.06)
3	778	76.01 (20.29)	69.11 (18.71)	70.02 (19.9)	3171.29 (1145.76)	1582.61 (602.41)	1588.68 (578.75)	0.5 (0.04)	0.04 (0.03)	0.0 (0.0)	60.74 (16.9)	10.14 (6.31)	26.67 (9.92)	22.48 (8.75)	16.72 (5.5)	14.74 (6.75)	28.03 (10.38)	17.32 (7.27)	9.93 (4.03)	0.13 (0.07)	0.35 (0.08)	0.29 (0.07)	0.23 (0.08)
4	27	341.7 (79.8)	224.37 (92.68)	315.41 (78.88)	2555.78 (1916.04)	1246.0 (1048.78)	1309.78 (901.69)	0.54 (0.07)	0.01 (0.01)	0.0 (0.0)	38.74 (11.49)	130.96 (49.7)	150.19 (45.67)	49.85 (34.73)	10.7 (13.0)	169.93 (68.82)	118.63 (58.68)	22.63 (21.91)	4.22 (6.34)	0.39 (0.13)	0.44 (0.11)	0.14 (0.08)	0.03 (0.03)
5	487	129.77 (31.3)	104.85 (29.58)	119.07 (30.63)	1664.83 (917.98)	825.8 (478.76)	839.03 (458.83)	0.51 (0.07)	0.03 (0.03)	0.0 (0.0)	57.72 (17.91)	28.99 (19.38)	59.18 (18.76)	31.69 (12.58)	9.91 (6.58)	41.48 (20.91)	54.8 (19.9)	18.59 (9.21)	4.19 (3.79)	0.22 (0.11)	0.46 (0.1)	0.25 (0.09)	0.08 (0.05)
6	57	180.33 (61.66)	162.77 (53.43)	168.49 (60.16)	12077.6 (3419.55)	6082.86 (1829.05)	5994.74 (1693.48)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	56.39 (10.56)	23.21 (15.16)	58.18 (21.32)	52.33 (23.09)	46.61 (16.08)	33.04 (15.85)	64.0 (25.37)	41.56 (20.08)	29.89 (10.6)	0.12 (0.05)	0.32 (0.06)	0.28 (0.05)	0.27 (0.07)
7	1	2.0 (0.0)	2.0 (0.0)	2.0 (0.0)	2544.0 (0.0)	1163.0 (0.0)	1381.0 (0.0)	0.54 (0.0)	1.0 (0.0)	0.1 (0.0)	23.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)
8	21	2.24 (0.43)	2.19 (0.39)	2.19 (0.5)	1258.86 (1485.71)	631.52 (743.57)	627.33 (767.91)	0.51 (0.08)	0.9 (0.22)	0.02 (0.01)	33.9 (43.52)	0.05 (0.21)	0.1 (0.29)	0.14 (0.35)	1.95 (0.58)	0.05 (0.21)	0.19 (0.39)	0.19 (0.39)	1.76 (0.68)	0.02 (0.07)	0.05 (0.15)	0.06 (0.16)	0.87 (0.21)
9	2	924.5 (17.5)	648.5 (43.5)	843.0 (60.0)	7571.5 (1667.5)	4025.0 (628.0)	3546.5 (1039.5)	0.46 (0.04)	0.01 (0.0)	0.0 (0.0)	30.7 (0.14)	326.5 (23.5)	423.0 (79.0)	144.5 (77.5)	30.5 (7.5)	466.5 (152.5)	313.0 (178.0)	51.0 (31.0)	12.5 (3.5)	0.35 (0.02)	0.46 (0.08)	0.16 (0.09)	0.03 (0.01)
Clustering 2: 7 Clusters																							
0	22	2.23 (0.42)	2.18 (0.39)	2.18 (0.49)	1317.27 (1476.03)	655.68 (734.86)	661.59 (766.5)	0.51 (0.08)	0.9 (0.22)	0.03 (0.02)	33.41 (42.58)	0.05 (0.21)	0.09 (0.29)	0.14 (0.34)	1.95 (0.56)	0.05 (0.21)	0.18 (0.39)	0.18 (0.39)	1.77 (0.67)	0.02 (0.07)	0.05 (0.14)	0.06 (0.16)	0.88 (0.2)
1	935	74.96 (23.32)	68.29 (21.23)	69.05 (22.48)	3971.24 (2175.16)	1979.99 (1112.9)	1991.25 (1101.26)	0.5 (0.04)	0.04 (0.03)	0.0 (0.0)	60.26 (17.33)	9.88 (6.53)	25.84 (10.8)	21.85 (9.31)	17.39 (6.55)	14.31 (7.15)	27.15 (11.31)	16.92 (7.6)	10.67 (4.85)	0.13 (0.07)	0.34 (0.08)	0.29 (0.08)	0.24 (0.09)
2	167	199.01 (56.45)	171.83 (46.96)	188.11 (55.11)	4622.72 (1439.03)	2298.5 (776.34)	2324.23 (715.51)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	56.09 (12.94)	30.88 (19.12)	80.46 (30.88)	59.38 (20.74)	28.29 (8.5)	48.99 (26.59)	84.63 (32.81)	40.56 (13.71)	13.93 (6.23)	0.15 (0.06)	0.4 (0.07)	0.3 (0.07)	0.15 (0.06)
3	29	381.9 (166.61)	253.62 (140.28)	351.79 (154.64)	2901.69 (2285.85)	1437.66 (1243.84)	1464.03 (1073.65)	0.53 (0.07)	0.01 (0.01)	0.0 (0.0)	38.19 (11.28)	144.45 (69.23)	169.0 (84.56)	56.38 (45.96)	12.07 (13.65)	190.38 (107.99)	132.03 (88.41)	24.59 (23.77)	4.79 (6.53)	0.39 (0.12)	0.44 (0.1)	0.14 (0.08)	0.03 (0.03)
4	83271	5.98 (10.87)	5.11 (9.53)	4.95 (9.85)	104.38 (331.51)	52.29 (168.67)	52.09 (166.89)	0.4 (0.3)	0.04 (0.14)	0.0 (0.0)	49.3 (68.18)	1.48 (2.74)	2.51 (4.94)	1.42 (3.22)	0.57 (1.59)	1.61 (3.39)	2.17 (4.64)	0.89 (2.2)	0.28 (0.92)	0.38 (0.41)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)
5	487	129.77 (31.3)	104.85 (29.58)	119.07 (30.63)	1664.83 (917.98)	825.8 (478.76)	839.03 (458.83)	0.51 (0.07)	0.03 (0.03)	0.0 (0.0)	57.72 (17.91)	28.99 (19.38)	59.18 (18.76)	31.69 (12.58)	9.91 (6.58)	41.48 (20.91)	54.8 (19.9)	18.59 (9.21)	4.19 (3.79)	0.22 (0.11)	0.46 (0.1)	0.25 (0.09)	0.08 (0.05)
6	57	180.33 (61.66)	162.77 (53.43)	168.49 (60.16)	12077.6 (3419.55)	6082.86 (1829.05)	5994.74 (1693.48)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	56.39 (10.56)	23.21 (15.16)	58.18 (21.32)	52.33 (23.09)	46.61 (16.08)	33.04 (15.85)	64.0 (25.37)	41.56 (20.08)	29.89 (10.6)	0.12 (0.05)	0.32 (0.06)	0.28 (0.05)	0.27 (0.07)
Clustering 3: 5 Clusters																							
0	1422	93.73 (37.01)	80.81 (29.95)	86.18 (34.88)	3181.35 (2144.17)	1584.71 (1092.18)	1596.64 (1080.97)	0.5 (0.05)	0.04 (0.03)	0.0 (0.0)	59.39 (17.57)	16.43 (15.46)	37.26 (21.15)	25.22 (11.53)	14.83 (7.46)	23.62 (18.7)	36.62 (19.8)	17.49 (8.23)	8.45 (5.46)	0.16 (0.1)	0.38 (0.1)	0.27 (0.08)	0.19 (0.11)
1	224	194.26 (58.39)	169.53 (48.84)	183.12 (57.08)	6519.72 (3881.07)	3261.48 (2004.39)	3258.24 (1915.04)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	56.17 (12.38)	28.93 (18.5)	74.79 (30.35)	57.59 (21.58)	32.95 (13.54)	44.93 (25.28)	79.38 (32.36)	40.81 (15.59)	17.99 (10.29)	0.14 (0.06)	0.38 (0.07)	0.3 (0.06)	0.18 (0.08)
2	22	2.23 (0.42)	2.18 (0.39)	2.18 (0.49)	1317.27 (1476.03)	655.68 (734.86)	661.59 (766.5)	0.51 (0.08)	0.9 (0.22)	0.03 (0.02)	33.41 (42.58)	0.05 (0.21)	0.09 (0.29)	0.14 (0.34)	1.95 (0.56)	0.05 (0.21)	0.18 (0.39)	0.18 (0.39)	1.77 (0.67)	0.02 (0.07)	0.05 (0.14)	0.06 (0.16)	0.88 (0.2)
3	29	381.9 (166.61)	253.62 (140.28)	351.79 (154.64)	2901.69 (2285.85)	1437.66 (1243.84)	1464.03 (1073.65)	0.53 (0.07)	0.01 (0.01)	0.0 (0.0)	38.19 (11.28)	144.45 (69.23)	169.0 (84.56)	56.38 (45.96)	12.07 (13.65)	190.38 (107.99)	132.03 (88.41)	24.59 (23.77)	4.79 (6.53)	0.39 (0.12)	0.44 (0.1)	0.14 (0.08)	0.03 (0.03)
4	83271	5.98 (10.87)	5.11 (9.53)	4.95 (9.85)	104.38 (331.51)	52.29 (168.67)	52.09 (166.89)	0.4 (0.3)	0.04 (0.14)	0.0 (0.0)	49.3 (68.18)	1.48 (2.74)	2.51 (4.94)	1.42 (3.22)	0.57 (1.59)	1.61 (3.39)	2.17 (4.64)	0.89 (2.2)	0.28 (0.92)	0.38 (0.41)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)
The entire data set																							
0	84968	8.07 (20.58)	6.9 (17.38)	6.9 (19.07)	174.06 (701.32)	87.03 (353.69)	87.03 (351.63)	0.4 (0.29)	0.04 (0.14)	0.0 (0.0)	49.48 (67.55)	1.85 (5.14)	3.34 (8.89)	1.98 (5.73)	0.9 (3.17)	2.16 (6.9)	2.99 (8.59)	1.29 (3.95)	0.46 (1.88)	0.38 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)

Table C.17: Mean shift on data from the 1st month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
24 Clusters																							
0	86293	6.7 (12.92)	5.72 (11.28)	5.66 (11.9)	133.28 (428.12)	66.66 (215.43)	66.62 (215.79)	0.41 (0.29)	0.04 (0.13)	0.0 (0.0)	33.65 (44.8)	1.56 (2.9)	2.79 (5.74)	1.63 (3.85)	0.72 (2.11)	1.78 (3.86)	2.47 (5.54)	1.05 (2.66)	0.37 (1.26)	0.38 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)
1	111	172.41 (39.72)	142.95 (35.12)	161.85 (38.47)	2493.24 (981.73)	1226.11 (496.81)	1267.14 (507.41)	0.51 (0.04)	0.02 (0.02)	0.0 (0.0)	44.06 (10.13)	31.72 (14.62)	77.16 (26.51)	49.28 (14.87)	14.25 (6.65)	47.32 (18.65)	79.76 (27.42)	28.64 (11.02)	6.13 (4.17)	0.18 (0.07)	0.44 (0.09)	0.29 (0.08)	0.09 (0.05)
2	11	2.27 (0.45)	2.18 (0.39)	2.0 (0.74)	539.09 (378.58)	337.27 (297.05)	201.82 (173.98)	0.42 (0.15)	0.89 (0.23)	0.02 (0.0)	12.45 (11.41)	0.09 (0.29)	0.09 (0.29)	0.27 (0.45)	1.82 (0.39)	0.18 (0.39)	0.0 (0.0)	0.55 (0.5)	1.27 (0.62)	0.03 (0.1)	0.03 (0.1)	0.12 (0.2)	0.82 (0.21)
3	109	101.83 (34.53)	92.59 (30.24)	96.22 (33.06)	7165.71 (2050.55)	3557.69 (1069.2)	3608.02 (1080.42)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	42.93 (9.46)	11.83 (8.3)	31.94 (15.08)	29.33 (12.52)	28.72 (8.4)	17.28 (9.32)	36.72 (17.17)	23.22 (9.18)	19.0 (6.21)	0.11 (0.05)	0.31 (0.07)	0.28 (0.06)	0.3 (0.09)
4	14	186.29 (30.99)	88.43 (30.88)	182.64 (30.38)	850.43 (462.13)	367.86 (213.9)	482.57 (256.4)	0.58 (0.07)	0.01 (0.01)	0.0 (0.0)	32.02 (10.29)	97.14 (27.56)	67.64 (24.2)	17.5 (8.67)	4.0 (3.91)	122.93 (27.31)	49.36 (23.89)	8.93 (6.4)	1.43 (2.77)	0.53 (0.14)	0.36 (0.11)	0.09 (0.04)	0.02 (0.02)
5	2	327.0 (5.0)	294.5 (18.5)	313.5 (1.5)	7762.0 (278.0)	3835.0 (560.0)	3927.0 (282.0)	0.51 (0.05)	0.01 (0.0)	0.0 (0.0)	43.22 (0.89)	35.0 (4.0)	151.5 (4.5)	115.5 (3.5)	25.0 (1.0)	76.5 (1.5)	169.0 (9.0)	55.0 (9.0)	13.0 (0.0)	0.11 (0.01)	0.46 (0.01)	0.35 (0.01)	0.08 (0.0)
6	2	295.0 (16.0)	207.0 (24.0)	290.0 (13.0)	2525.0 (179.0)	1258.0 (171.0)	1267.0 (8.0)	0.5 (0.03)	0.0 (0.0)	0.0 (0.0)	24.78 (2.77)	84.0 (3.0)	92.0 (13.0)	108.5 (6.5)	10.5 (0.5)	92.5 (1.5)	155.5 (9.5)	41.0 (5.0)	1.0 (0.0)	0.29 (0.03)	0.31 (0.03)	0.37 (0.0)	0.04 (0.0)
7	35	139.14 (23.47)	116.97 (41.15)	86.26 (34.71)	1205.14 (1058.56)	669.09 (574.01)	536.06 (496.21)	0.43 (0.16)	0.03 (0.02)	0.0 (0.0)	46.16 (13.33)	60.63 (17.61)	54.66 (15.5)	18.26 (11.35)	5.6 (5.88)	41.6 (24.5)	32.91 (14.96)	9.4 (8.77)	2.34 (3.23)	0.45 (0.14)	0.39 (0.09)	0.13 (0.07)	0.04 (0.04)
8	1	684.0 (0.0)	291.0 (0.0)	682.0 (0.0)	4715.0 (0.0)	1776.0 (0.0)	2939.0 (0.0)	0.62 (0.0)	0.0 (0.0)	0.0 (0.0)	20.75 (0.0)	320.0 (0.0)	307.0 (0.0)	38.0 (0.0)	19.0 (0.0)	338.0 (0.0)	309.0 (0.0)	20.0 (0.0)	15.0 (0.0)	0.47 (0.0)	0.45 (0.0)	0.06 (0.0)	0.03 (0.0)
9	1	424.0 (0.0)	335.0 (0.0)	423.0 (0.0)	2766.0 (0.0)	1193.0 (0.0)	1573.0 (0.0)	0.57 (0.0)	0.01 (0.0)	0.0 (0.0)	43.47 (0.0)	90.0 (0.0)	161.0 (0.0)	173.0 (0.0)	0.0 (0.0)	100.0 (0.0)	289.0 (0.0)	34.0 (0.0)	0.0 (0.0)	0.21 (0.0)	0.38 (0.0)	0.41 (0.0)	0.0 (0.0)
10	1	380.0 (0.0)	266.0 (0.0)	304.0 (0.0)	1103.0 (0.0)	673.0 (0.0)	430.0 (0.0)	0.39 (0.0)	0.01 (0.0)	0.0 (0.0)	33.96 (0.0)	173.0 (0.0)	189.0 (0.0)	15.0 (0.0)	3.0 (0.0)	252.0 (0.0)	49.0 (0.0)	3.0 (0.0)	0.0 (0.0)	0.46 (0.0)	0.5 (0.0)	0.04 (0.0)	0.01 (0.0)
11	1	367.0 (0.0)	208.0 (0.0)	350.0 (0.0)	3472.0 (0.0)	1563.0 (0.0)	1909.0 (0.0)	0.55 (0.0)	0.02 (0.0)	0.0 (0.0)	45.72 (0.0)	122.0 (0.0)	158.0 (0.0)	59.0 (0.0)	28.0 (0.0)	133.0 (0.0)	163.0 (0.0)	45.0 (0.0)	9.0 (0.0)	0.33 (0.0)	0.43 (0.0)	0.16 (0.0)	0.08 (0.0)
12	1	364.0 (0.0)	274.0 (0.0)	350.0 (0.0)	1975.0 (0.0)	977.0 (0.0)	998.0 (0.0)	0.51 (0.0)	0.01 (0.0)	0.0 (0.0)	32.78 (0.0)	103.0 (0.0)	175.0 (0.0)	83.0 (0.0)	3.0 (0.0)	120.0 (0.0)	215.0 (0.0)	14.0 (0.0)	1.0 (0.0)	0.28 (0.0)	0.48 (0.0)	0.23 (0.0)	0.01 (0.0)
13	1	357.0 (0.0)	222.0 (0.0)	325.0 (0.0)	1623.0 (0.0)	801.0 (0.0)	822.0 (0.0)	0.51 (0.0)	0.01 (0.0)	0.0 (0.0)	42.53 (0.0)	164.0 (0.0)	160.0 (0.0)	28.0 (0.0)	5.0 (0.0)	184.0 (0.0)	131.0 (0.0)	8.0 (0.0)	2.0 (0.0)	0.46 (0.0)	0.45 (0.0)	0.08 (0.0)	0.01 (0.0)
14	1	347.0 (0.0)	225.0 (0.0)	342.0 (0.0)	1520.0 (0.0)	629.0 (0.0)	891.0 (0.0)	0.59 (0.0)	0.01 (0.0)	0.0 (0.0)	36.48 (0.0)	105.0 (0.0)	206.0 (0.0)	28.0 (0.0)	8.0 (0.0)	215.0 (0.0)	104.0 (0.0)	22.0 (0.0)	1.0 (0.0)	0.3 (0.0)	0.59 (0.0)	0.08 (0.0)	0.02 (0.0)
15	1	331.0 (0.0)	303.0 (0.0)	243.0 (0.0)	3023.0 (0.0)	1654.0 (0.0)	1369.0 (0.0)	0.45 (0.0)	0.01 (0.0)	0.0 (0.0)	48.33 (0.0)	93.0 (0.0)	154.0 (0.0)	67.0 (0.0)	17.0 (0.0)	96.0 (0.0)	112.0 (0.0)	27.0 (0.0)	8.0 (0.0)	0.28 (0.0)	0.47 (0.0)	0.2 (0.0)	0.05 (0.0)
16	1	331.0 (0.0)	212.0 (0.0)	331.0 (0.0)	2615.0 (0.0)	899.0 (0.0)	1716.0 (0.0)	0.66 (0.0)	0.01 (0.0)	0.0 (0.0)	14.09 (0.0)	34.0 (0.0)	183.0 (0.0)	104.0 (0.0)	10.0 (0.0)	43.0 (0.0)	230.0 (0.0)	56.0 (0.0)	2.0 (0.0)	0.1 (0.0)	0.55 (0.0)	0.31 (0.0)	0.03 (0.0)
17	2	276.0 (42.0)	247.0 (48.0)	266.0 (37.0)	8627.0 (674.0)	4696.0 (397.0)	3931.0 (277.0)	0.46 (0.0)	0.01 (0.0)	0.0 (0.0)	34.16 (1.21)	32.0 (1.0)	107.5 (28.5)	80.5 (12.5)	56.0 (2.0)	60.0 (4.0)	119.5 (23.5)	59.5 (7.5)	27.0 (2.0)	0.12 (0.02)	0.38 (0.05)	0.29 (0.0)	0.21 (0.02)
18	1	309.0 (0.0)	262.0 (0.0)	276.0 (0.0)	7027.0 (0.0)	3949.0 (0.0)	3078.0 (0.0)	0.44 (0.0)	0.02 (0.0)	0.0 (0.0)	44.21 (0.0)	70.0 (0.0)	154.0 (0.0)	61.0 (0.0)	24.0 (0.0)	95.0 (0.0)	138.0 (0.0)	29.0 (0.0)	14.0 (0.0)	0.23 (0.0)	0.5 (0.0)	0.2 (0.0)	0.08 (0.0)
19	1	271.0 (0.0)	176.0 (0.0)	271.0 (0.0)	7490.0 (0.0)	3525.0 (0.0)	3965.0 (0.0)	0.53 (0.0)	0.01 (0.0)	0.0 (0.0)	48.21 (0.0)	89.0 (0.0)	73.0 (0.0)	52.0 (0.0)	57.0 (0.0)	107.0 (0.0)	81.0 (0.0)	52.0 (0.0)	31.0 (0.0)	0.33 (0.0)	0.27 (0.0)	0.19 (0.0)	0.21 (0.0)
20	18	183.56 (37.93)	163.11 (35.96)	178.22 (35.13)	4846.11 (1231.32)	2431.11 (651.23)	2415.0 (598.07)	0.5 (0.02)	0.02 (0.01)	0.0 (0.0)	41.6 (9.67)	21.17 (10.45)	65.5 (21.12)	63.39 (13.8)	33.5 (6.1)	36.61 (13.66)	76.61 (18.45)	49.78 (13.4)	15.22 (3.22)	0.11 (0.04)	0.36 (0.07)	0.35 (0.05)	0.19 (0.02)
21	3	241.33 (19.48)	199.0 (22.99)	226.67 (11.84)	9174.0 (1218.89)	4788.0 (590.68)	4386.0 (651.81)	0.48 (0.01)	0.01 (0.01)	0.0 (0.0)	36.87 (8.4)	49.0 (7.79)	88.33 (22.57)	59.0 (4.9)	45.0 (5.72)	75.0 (22.11)	76.67 (7.54)	47.67 (4.92)	27.33 (0.47)	0.2 (0.02)	0.36 (0.06)	0.25 (0.04)	0.19 (0.04)
22	12	188.5 (26.6)	145.08 (32.15)	167.0 (25.96)	968.67 (364.94)	507.33 (211.99)	461.33 (160.92)	0.48 (0.05)	0.01 (0.01)	0.0 (0.0)	38.82 (10.24)	54.5 (12.32)	110.75 (22.17)	19.83 (6.48)	3.42 (2.33)	101.17 (26.75)	57.58 (10.82)	6.92 (3.71)	1.33 (1.11)	0.29 (0.05)	0.59 (0.06)	0.11 (0.04)	0.02 (0.01)
23	15	2.47 (0.72)	2.47 (0.72)	2.4 (0.8)	882.73 (1034.38)	447.33 (590.26)	435.4 (470.41)	0.53 (0.06)	0.84 (0.25)	0.01 (0.0)	27.58 (23.2)	0.0 (0.0)	0.27 (0.44)	0.27 (0.44)	1.93 (0.77)	0.07 (0.25)	0.2 (0.4)	0.47 (0.72)	1.67 (0.7)	0.0 (0.0)	0.11 (0.19)	0.11 (0.18)	0.78 (0.21)
24	1	2.0 (0.0)	2.0 (0.0)	2.0 (0.0)	1187.0 (0.0)	722.0 (0.0)	465.0 (0.0)	0.39 (0.0)	1.0 (0.0)	0.05 (0.0)	2.5 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)
The entire data set																							
0	86639	7.25 (16.34)	6.17 (13.85)	6.17 (15.12)	148.14 (522.07)	74.07 (262.25)	74.07 (263.05)	0.41 (0.29)	0.04 (0.13)	0.0 (0.0)	33.68 (44.72)	1.68 (4.06)	3.01 (7.21)	1.77 (4.66)	0.79 (2.5)	1.94 (5.27)	2.68 (7.04)	1.13 (3.14)	0.41 (1.5)	0.38 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)

Table C.18: Mean shift on data from the 2nd month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
22 Clusters																							
0	83795	7.04 (14.58)	6.02 (12.71)	5.98 (13.38)	144.38 (486.26)	72.13 (244.6)	72.25 (245.31)	0.41 (0.29)	0.04 (0.13)	0.0 (0.0)	38.47 (46.96)	1.61 (3.27)	2.93 (6.4)	1.73 (4.3)	0.77 (2.34)	1.87 (4.32)	2.61 (6.2)	1.11 (2.98)	0.39 (1.38)	0.37 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.17)
1	2	489.0 (7.0)	276.0 (6.0)	481.5 (6.5)	1818.0 (14.0)	799.0 (8.0)	1019.0 (22.0)	0.56 (0.01)	0.01 (0.0)	0.0 (0.0)	29.34 (4.91)	204.5 (7.5)	232.5 (24.5)	47.5 (11.5)	4.5 (1.5)	307.0 (0.0)	160.5 (11.5)	13.0 (6.0)	1.0 (1.0)	0.42 (0.02)	0.47 (0.04)	0.1 (0.02)	0.01 (0.0)
2	11	199.09 (43.1)	193.91 (43.84)	62.18 (40.58)	1095.73 (767.87)	709.73 (423.56)	386.0 (357.96)	0.28 (0.16)	0.01 (0.01)	0.0 (0.0)	48.99 (7.61)	103.82 (35.5)	72.18 (21.46)	18.55 (13.02)	4.55 (5.26)	27.55 (18.59)	26.36 (18.58)	6.73 (6.24)	1.55 (2.35)	0.53 (0.15)	0.36 (0.08)	0.09 (0.06)	0.02 (0.03)
3	44	161.82 (36.33)	145.0 (31.41)	154.77 (35.74)	6437.34 (1654.57)	3203.41 (878.09)	3233.93 (852.5)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	50.35 (10.53)	18.41 (9.19)	53.68 (17.2)	52.95 (12.24)	36.77 (7.49)	28.95 (11.33)	63.36 (18.32)	40.55 (9.64)	21.91 (5.47)	0.11 (0.04)	0.33 (0.05)	0.33 (0.05)	0.23 (0.04)
4	2	129.5 (1.5)	124.0 (3.0)	118.0 (3.0)	20408.0 (292.0)	11478.5 (323.5)	8929.5 (31.5)	0.44 (0.01)	0.02 (0.01)	0.0 (0.0)	38.48 (0.04)	14.0 (2.0)	33.0 (11.0)	36.0 (1.0)	46.5 (8.5)	16.0 (1.0)	46.0 (14.0)	24.0 (6.0)	32.0 (6.0)	0.11 (0.02)	0.25 (0.08)	0.28 (0.0)	0.36 (0.07)
5	1	573.0 (0.0)	369.0 (0.0)	561.0 (0.0)	4906.0 (0.0)	2600.0 (0.0)	2306.0 (0.0)	0.47 (0.0)	0.01 (0.0)	0.0 (0.0)	31.33 (0.0)	211.0 (0.0)	254.0 (0.0)	87.0 (0.0)	21.0 (0.0)	339.0 (0.0)	179.0 (0.0)	35.0 (0.0)	8.0 (0.0)	0.37 (0.0)	0.44 (0.0)	0.15 (0.0)	0.04 (0.0)
6	1	555.0 (0.0)	238.0 (0.0)	553.0 (0.0)	1611.0 (0.0)	609.0 (0.0)	1002.0 (0.0)	0.62 (0.0)	0.01 (0.0)	0.0 (0.0)	41.57 (0.0)	309.0 (0.0)	208.0 (0.0)	35.0 (0.0)	3.0 (0.0)	376.0 (0.0)	165.0 (0.0)	12.0 (0.0)	0.0 (0.0)	0.56 (0.0)	0.37 (0.0)	0.06 (0.0)	0.01 (0.0)
7	1	359.0 (0.0)	355.0 (0.0)	161.0 (0.0)	9592.0 (0.0)	5814.0 (0.0)	3778.0 (0.0)	0.39 (0.0)	0.03 (0.0)	0.0 (0.0)	47.06 (0.0)	124.0 (0.0)	148.0 (0.0)	53.0 (0.0)	34.0 (0.0)	57.0 (0.0)	65.0 (0.0)	23.0 (0.0)	16.0 (0.0)	0.35 (0.0)	0.41 (0.0)	0.15 (0.0)	0.09 (0.0)
8	1	359.0 (0.0)	187.0 (0.0)	357.0 (0.0)	5771.0 (0.0)	2388.0 (0.0)	3383.0 (0.0)	0.59 (0.0)	0.01 (0.0)	0.0 (0.0)	46.69 (0.0)	173.0 (0.0)	146.0 (0.0)	24.0 (0.0)	16.0 (0.0)	236.0 (0.0)	94.0 (0.0)	17.0 (0.0)	10.0 (0.0)	0.48 (0.0)	0.41 (0.0)	0.07 (0.0)	0.04 (0.0)
9	21	249.9 (39.44)	198.19 (41.18)	223.43 (39.07)	1783.52 (599.36)	910.33 (325.24)	873.19 (307.38)	0.49 (0.06)	0.02 (0.02)	0.0 (0.0)	41.93 (11.14)	66.05 (19.75)	126.24 (19.87)	49.67 (16.86)	7.95 (5.58)	90.67 (16.73)	108.1 (25.51)	22.43 (10.48)	2.24 (2.37)	0.26 (0.06)	0.51 (0.06)	0.2 (0.06)	0.03 (0.02)
10	1	339.0 (0.0)	283.0 (0.0)	336.0 (0.0)	11951.0 (0.0)	6190.0 (0.0)	5761.0 (0.0)	0.48 (0.0)	0.02 (0.0)	0.0 (0.0)	46.64 (0.0)	52.0 (0.0)	113.0 (0.0)	95.0 (0.0)	79.0 (0.0)	72.0 (0.0)	144.0 (0.0)	78.0 (0.0)	42.0 (0.0)	0.15 (0.0)	0.33 (0.0)	0.28 (0.0)	0.23 (0.0)
11	3	300.67 (26.34)	231.33 (29.33)	295.67 (25.22)	5477.33 (287.46)	2691.0 (280.97)	2786.33 (161.54)	0.51 (0.03)	0.01 (0.0)	0.0 (0.0)	36.06 (2.83)	56.33 (12.76)	112.33 (20.53)	94.33 (21.91)	37.67 (6.13)	73.33 (11.9)	149.67 (5.73)	56.33 (9.46)	16.33 (5.44)	0.19 (0.04)	0.37 (0.04)	0.32 (0.09)	0.12 (0.01)
12	3	288.0 (57.35)	242.67 (52.95)	275.0 (55.72)	3105.67 (782.37)	1525.67 (425.13)	1580.0 (398.33)	0.51 (0.05)	0.01 (0.01)	0.0 (0.0)	41.85 (12.23)	29.0 (16.27)	150.33 (26.23)	90.33 (22.4)	18.33 (7.13)	36.67 (25.85)	183.0 (35.9)	50.33 (14.27)	5.0 (3.27)	0.1 (0.04)	0.53 (0.06)	0.31 (0.02)	0.06 (0.02)
13	1	331.0 (0.0)	197.0 (0.0)	330.0 (0.0)	1093.0 (0.0)	484.0 (0.0)	609.0 (0.0)	0.56 (0.0)	0.01 (0.0)	0.0 (0.0)	32.56 (0.0)	128.0 (0.0)	176.0 (0.0)	26.0 (0.0)	1.0 (0.0)	202.0 (0.0)	121.0 (0.0)	7.0 (0.0)	0.0 (0.0)	0.39 (0.0)	0.53 (0.0)	0.08 (0.0)	0.0 (0.0)
14	1	306.0 (0.0)	246.0 (0.0)	270.0 (0.0)	12087.0 (0.0)	5416.0 (0.0)	6671.0 (0.0)	0.55 (0.0)	0.03 (0.0)	0.0 (0.0)	46.62 (0.0)	81.0 (0.0)	129.0 (0.0)	50.0 (0.0)	46.0 (0.0)	110.0 (0.0)	89.0 (0.0)	39.0 (0.0)	32.0 (0.0)	0.26 (0.0)	0.42 (0.0)	0.16 (0.0)	0.15 (0.0)
15	105	186.42 (34.58)	160.59 (30.82)	174.43 (33.67)	3711.13 (1509.94)	1867.45 (808.85)	1843.69 (732.71)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	48.92 (12.76)	30.78 (14.77)	78.5 (21.99)	55.03 (15.15)	22.1 (7.79)	47.3 (20.26)	81.08 (20.39)	36.22 (11.86)	9.83 (5.18)	0.16 (0.06)	0.42 (0.08)	0.3 (0.07)	0.12 (0.05)
16	13	201.77 (39.03)	110.38 (34.81)	191.85 (37.31)	1043.38 (516.9)	477.54 (251.99)	565.85 (275.16)	0.55 (0.06)	0.02 (0.02)	0.0 (0.0)	41.08 (16.07)	94.38 (21.53)	80.54 (21.03)	21.62 (9.64)	5.23 (4.73)	127.92 (27.8)	51.92 (17.97)	10.46 (7.38)	1.54 (1.6)	0.47 (0.08)	0.4 (0.09)	0.11 (0.04)	0.02 (0.02)
17	2	181.0 (23.0)	170.5 (19.5)	173.0 (23.0)	15342.5 (1742.5)	8105.0 (713.0)	7237.5 (1029.5)	0.47 (0.01)	0.02 (0.0)	0.0 (0.0)	42.2 (2.66)	14.5 (2.5)	65.5 (14.5)	53.5 (8.5)	47.5 (2.5)	34.5 (4.5)	65.0 (20.0)	39.5 (0.5)	34.0 (1.0)	0.08 (0.0)	0.36 (0.03)	0.29 (0.01)	0.27 (0.05)
18	48	106.04 (25.4)	97.79 (24.03)	100.04 (23.3)	8869.6 (1778.24)	4369.27 (887.66)	4500.33 (1056.69)	0.51 (0.04)	0.03 (0.02)	0.0 (0.0)	48.94 (10.67)	11.29 (6.56)	32.44 (12.16)	30.38 (8.88)	31.94 (9.02)	18.1 (7.92)	35.94 (10.81)	24.38 (8.32)	21.62 (5.74)	0.1 (0.05)	0.3 (0.07)	0.29 (0.05)	0.31 (0.08)
19	2	36.5 (20.5)	35.5 (19.5)	36.0 (20.0)	16477.0 (3558.0)	8582.0 (1843.0)	7895.0 (1715.0)	0.48 (0.0)	0.03 (0.0)	0.0 (0.0)	48.41 (10.16)	1.0 (1.0)	4.0 (3.0)	6.5 (5.5)	25.0 (11.0)	2.0 (2.0)	5.0 (4.0)	8.0 (5.0)	21.0 (9.0)	0.02 (0.02)	0.09 (0.03)	0.14 (0.07)	0.75 (0.12)
20	51	24.0 (25.36)	22.06 (23.23)	22.18 (23.4)	5336.27 (4034.53)	2711.75 (2062.78)	2624.53 (2023.26)	0.49 (0.09)	0.39 (0.43)	0.01 (0.01)	35.86 (24.18)	2.92 (4.03)	7.18 (8.63)	6.14 (7.47)	7.76 (7.51)	4.25 (4.93)	7.39 (9.29)	4.76 (5.83)	5.76 (5.47)	0.08 (0.1)	0.2 (0.18)	0.21 (0.17)	0.5 (0.27)
21	4	2.0 (0.0)	1.5 (0.5)	2.0 (0.0)	911.75 (758.01)	462.5 (433.64)	449.25 (324.97)	0.53 (0.06)	1.0 (0.0)	0.02 (0.01)	18.88 (17.02)	0.0 (0.0)	0.25 (0.43)	0.25 (0.43)	1.5 (0.5)	0.0 (0.0)	0.25 (0.43)	0.5 (0.5)	1.25 (0.43)	0.0 (0.0)	0.12 (0.22)	0.12 (0.22)	0.75 (0.25)
The entire data set																							
0	84113	7.6 (18.05)	6.49 (15.4)	6.49 (16.61)	163.07 (614.17)	81.54 (310.03)	81.54 (308.01)	0.41 (0.29)	0.04 (0.14)	0.0 (0.0)	38.49 (46.88)	1.73 (4.46)	3.15 (7.83)	1.87 (5.11)	0.85 (2.79)	2.02 (5.83)	2.82 (7.54)	1.21 (3.51)	0.44 (1.65)	0.37 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.17)

Table C.19: Mean shift on data from the 3rd month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. de-gree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neigh-bors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 mes-sage out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 mes-sage (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
29 Clusters																							
0	80933	7.17 (14.81)	6.15 (12.99)	6.06 (13.61)	152.32 (536.01)	76.17 (269.66)	76.16 (269.71)	0.41 (0.29)	0.04 (0.14)	0.0 (0.0)	45.84 (64.79)	1.64 (3.23)	2.97 (6.45)	1.76 (4.37)	0.8 (2.48)	1.89 (4.29)	2.63 (6.23)	1.13 (3.07)	0.41 (1.49)	0.37 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.17)
1	36	252.92 (42.73)	181.86 (43.36)	235.92 (43.38)	2534.78 (1151.15)	1204.44 (587.3)	1330.33 (593.03)	0.53 (0.04)	0.02 (0.01)	0.0 (0.0)	49.99 (11.77)	76.67 (20.62)	112.53 (30.07)	50.64 (16.06)	13.08 (7.81)	96.69 (28.94)	108.39 (24.38)	26.08 (12.18)	4.75 (4.02)	0.3 (0.06)	0.44 (0.07)	0.2 (0.07)	0.05 (0.03)
2	16	2.69 (0.77)	2.56 (0.7)	2.38 (0.99)	700.38 (575.47)	378.88 (374.75)	321.5 (322.3)	0.48 (0.15)	0.83 (0.21)	0.02 (0.0)	43.62 (46.81)	0.19 (0.39)	0.12 (0.33)	0.06 (0.24)	2.31 (0.68)	0.12 (0.33)	0.12 (0.33)	0.38 (0.48)	1.75 (0.97)	0.06 (0.12)	0.03 (0.08)	0.03 (0.12)	0.88 (0.19)
3	14	283.07 (54.47)	232.07 (46.63)	263.71 (45.59)	5862.07 (1286.79)	2899.5 (698.6)	2962.57 (634.27)	0.51 (0.03)	0.02 (0.02)	0.0 (0.0)	48.77 (15.07)	55.79 (22.76)	98.07 (34.21)	90.21 (17.34)	39.0 (8.24)	71.36 (22.72)	110.29 (19.68)	65.57 (16.68)	16.5 (5.17)	0.19 (0.06)	0.34 (0.07)	0.33 (0.06)	0.14 (0.03)
4	59	104.93 (37.25)	97.12 (33.54)	98.92 (38.25)	10668.17 (2420.77)	5383.95 (1426.28)	5284.22 (1152.64)	0.5 (0.04)	0.03 (0.03)	0.0 (0.0)	54.05 (13.45)	11.14 (6.58)	31.86 (13.57)	29.59 (13.83)	32.34 (11.64)	16.71 (8.19)	36.08 (17.31)	24.54 (11.48)	21.58 (7.05)	0.11 (0.07)	0.3 (0.06)	0.27 (0.06)	0.31 (0.08)
5	59	159.47 (23.06)	122.24 (29.52)	147.07 (26.05)	1745.32 (1070.39)	871.05 (518.64)	874.27 (570.36)	0.5 (0.07)	0.03 (0.03)	0.0 (0.0)	55.55 (16.86)	41.93 (15.99)	79.54 (20.81)	29.41 (13.17)	8.59 (5.95)	68.76 (29.05)	58.78 (16.74)	15.71 (8.54)	3.81 (3.56)	0.27 (0.1)	0.5 (0.08)	0.18 (0.08)	0.05 (0.04)
6	5	220.2 (49.6)	77.4 (50.23)	216.2 (46.91)	830.6 (375.59)	321.2 (221.95)	509.4 (167.4)	0.66 (0.14)	0.01 (0.01)	0.0 (0.0)	38.73 (18.57)	125.8 (18.48)	78.8 (32.36)	13.8 (9.28)	1.8 (1.47)	144.2 (33.05)	64.0 (17.62)	7.2 (3.31)	0.8 (0.75)	0.59 (0.09)	0.35 (0.08)	0.06 (0.03)	0.01 (0.01)
7	2	2.5 (0.5)	2.0 (1.0)	2.5 (0.5)	3182.0 (1656.0)	1786.0 (1128.0)	1396.0 (528.0)	0.48 (0.09)	0.67 (0.33)	0.03 (0.0)	4.08 (2.58)	0.0 (0.0)	0.5 (0.5)	0.5 (0.5)	1.5 (0.5)	0.0 (0.0)	0.5 (0.5)	0.5 (0.5)	1.5 (0.5)	0.0 (0.0)	0.25 (0.25)	0.17 (0.17)	0.58 (0.08)
8	1	777.0 (0.0)	524.0 (0.0)	726.0 (0.0)	1792.0 (0.0)	897.0 (0.0)	895.0 (0.0)	0.5 (0.0)	0.01 (0.0)	0.0 (0.0)	33.76 (0.0)	260.0 (0.0)	506.0 (0.0)	10.0 (0.0)	1.0 (0.0)	616.0 (0.0)	107.0 (0.0)	3.0 (0.0)	0.0 (0.0)	0.33 (0.0)	0.65 (0.0)	0.01 (0.0)	0.0 (0.0)
9	1	499.0 (0.0)	351.0 (0.0)	482.0 (0.0)	2220.0 (0.0)	1155.0 (0.0)	1065.0 (0.0)	0.48 (0.0)	0.01 (0.0)	0.0 (0.0)	40.79 (0.0)	158.0 (0.0)	299.0 (0.0)	30.0 (0.0)	12.0 (0.0)	344.0 (0.0)	120.0 (0.0)	14.0 (0.0)	4.0 (0.0)	0.32 (0.0)	0.6 (0.0)	0.06 (0.0)	0.02 (0.0)
10	1	385.0 (0.0)	220.0 (0.0)	379.0 (0.0)	2407.0 (0.0)	1192.0 (0.0)	1215.0 (0.0)	0.5 (0.0)	0.02 (0.0)	0.0 (0.0)	56.52 (0.0)	168.0 (0.0)	167.0 (0.0)	38.0 (0.0)	12.0 (0.0)	204.0 (0.0)	152.0 (0.0)	17.0 (0.0)	6.0 (0.0)	0.44 (0.0)	0.43 (0.0)	0.1 (0.0)	0.03 (0.0)
11	2	319.0 (46.0)	253.0 (17.0)	317.0 (48.0)	3369.0 (78.0)	1569.0 (163.0)	1800.0 (241.0)	0.53 (0.06)	0.01 (0.0)	0.0 (0.0)	43.29 (1.03)	63.5 (27.5)	102.0 (15.0)	141.5 (6.5)	12.0 (3.0)	75.5 (26.5)	156.0 (6.0)	83.5 (14.5)	2.0 (1.0)	0.19 (0.06)	0.32 (0.0)	0.45 (0.04)	0.04 (0.02)
12	1	334.0 (0.0)	322.0 (0.0)	149.0 (0.0)	959.0 (0.0)	674.0 (0.0)	285.0 (0.0)	0.3 (0.0)	0.01 (0.0)	0.0 (0.0)	54.49 (0.0)	138.0 (0.0)	174.0 (0.0)	22.0 (0.0)	0.0 (0.0)	91.0 (0.0)	53.0 (0.0)	5.0 (0.0)	0.0 (0.0)	0.41 (0.0)	0.52 (0.0)	0.07 (0.0)	0.0 (0.0)
13	7	261.43 (36.97)	242.57 (35.3)	248.0 (36.36)	8189.29 (1798.91)	4174.71 (1123.83)	4014.57 (703.43)	0.5 (0.04)	0.03 (0.03)	0.0 (0.0)	54.84 (11.45)	21.57 (8.4)	107.57 (20.69)	85.71 (12.68)	46.57 (6.76)	44.57 (15.55)	115.14 (26.32)	64.86 (13.11)	23.43 (3.29)	0.08 (0.03)	0.41 (0.03)	0.33 (0.04)	0.18 (0.03)
14	9	225.56 (39.42)	200.11 (36.22)	214.11 (32.73)	7302.89 (1081.31)	3531.78 (355.41)	3771.11 (807.56)	0.51 (0.04)	0.03 (0.02)	0.0 (0.0)	57.19 (7.57)	27.33 (13.93)	71.89 (14.2)	74.22 (12.39)	52.11 (8.62)	40.22 (10.65)	82.11 (15.8)	63.78 (11.57)	28.0 (6.07)	0.12 (0.04)	0.32 (0.03)	0.33 (0.03)	0.23 (0.03)
15	120	176.72 (29.74)	154.25 (24.91)	165.12 (29.8)	3478.0 (1311.73)	1724.08 (699.16)	1753.92 (649.43)	0.51 (0.05)	0.04 (0.04)	0.0 (0.0)	57.88 (16.49)	25.51 (9.89)	74.37 (19.19)	55.08 (13.43)	21.77 (8.02)	39.05 (12.4)	79.74 (19.86)	36.77 (10.12)	9.55 (5.4)	0.14 (0.04)	0.42 (0.07)	0.31 (0.06)	0.13 (0.05)
16	6	212.83 (29.43)	198.33 (32.02)	208.33 (30.33)	11284.17 (2329.45)	5867.33 (1152.07)	5416.83 (1269.15)	0.48 (0.03)	0.04 (0.02)	0.0 (0.0)	46.18 (13.8)	14.5 (4.03)	68.17 (13.13)	66.17 (13.2)	64.0 (7.68)	31.67 (7.09)	79.17 (14.89)	61.67 (9.86)	35.83 (4.52)	0.07 (0.02)	0.32 (0.03)	0.31 (0.03)	0.3 (0.02)
17	31	161.71 (35.9)	138.06 (28.24)	149.65 (33.9)	6430.48 (1012.14)	3137.68 (626.99)	3292.81 (548.83)	0.51 (0.05)	0.03 (0.03)	0.0 (0.0)	56.0 (15.39)	29.23 (17.75)	55.16 (13.66)	43.35 (10.22)	33.97 (7.49)	39.42 (19.23)	56.42 (13.02)	33.03 (9.26)	20.77 (5.32)	0.17 (0.07)	0.34 (0.05)	0.27 (0.04)	0.22 (0.05)
18	1	1028.0 (0.0)	690.0 (0.0)	1001.0 (0.0)	7681.0 (0.0)	3582.0 (0.0)	4099.0 (0.0)	0.53 (0.0)	0.01 (0.0)	0.0 (0.0)	28.56 (0.0)	324.0 (0.0)	490.0 (0.0)	178.0 (0.0)	36.0 (0.0)	458.0 (0.0)	449.0 (0.0)	77.0 (0.0)	17.0 (0.0)	0.32 (0.0)	0.48 (0.0)	0.17 (0.0)	0.04 (0.0)
19	1	498.0 (0.0)	400.0 (0.0)	469.0 (0.0)	7770.0 (0.0)	4153.0 (0.0)	3617.0 (0.0)	0.47 (0.0)	0.05 (0.0)	0.0 (0.0)	48.74 (0.0)	119.0 (0.0)	189.0 (0.0)	121.0 (0.0)	69.0 (0.0)	183.0 (0.0)	182.0 (0.0)	83.0 (0.0)	21.0 (0.0)	0.24 (0.0)	0.38 (0.0)	0.24 (0.0)	0.14 (0.0)
20	1	492.0 (0.0)	444.0 (0.0)	488.0 (0.0)	5428.0 (0.0)	2736.0 (0.0)	2692.0 (0.0)	0.5 (0.0)	0.0 (0.0)	0.0 (0.0)	31.32 (0.0)	45.0 (0.0)	135.0 (0.0)	288.0 (0.0)	24.0 (0.0)	54.0 (0.0)	356.0 (0.0)	71.0 (0.0)	7.0 (0.0)	0.09 (0.0)	0.27 (0.0)	0.59 (0.0)	0.05 (0.0)
21	2	381.5 (30.5)	328.0 (36.0)	342.5 (5.5)	4273.5 (1773.5)	2104.5 (958.5)	2169.0 (815.0)	0.52 (0.02)	0.01 (0.0)	0.0 (0.0)	37.87 (10.53)	82.0 (25.0)	171.0 (9.0)	114.5 (15.5)	14.0 (12.0)	96.0 (19.0)	201.0 (33.0)	38.5 (1.5)	7.0 (7.0)	0.21 (0.05)	0.45 (0.01)	0.31 (0.07)	0.03 (0.03)
22	1	343.0 (0.0)	132.0 (0.0)	342.0 (0.0)	1913.0 (0.0)	841.0 (0.0)	1072.0 (0.0)	0.56 (0.0)	0.01 (0.0)	0.0 (0.0)	31.82 (0.0)	210.0 (0.0)	121.0 (0.0)	9.0 (0.0)	3.0 (0.0)	218.0 (0.0)	119.0 (0.0)	2.0 (0.0)	3.0 (0.0)	0.61 (0.0)	0.35 (0.0)	0.03 (0.0)	0.01 (0.0)
23	1	339.0 (0.0)	300.0 (0.0)	330.0 (0.0)	8821.0 (0.0)	3388.0 (0.0)	5433.0 (0.0)	0.62 (0.0)	0.07 (0.0)	0.0 (0.0)	60.27 (0.0)	40.0 (0.0)	105.0 (0.0)	113.0 (0.0)	81.0 (0.0)	70.0 (0.0)	105.0 (0.0)	113.0 (0.0)	42.0 (0.0)	0.12 (0.0)	0.31 (0.0)	0.33 (0.0)	0.24 (0.0)
24	1	287.0 (0.0)	219.0 (0.0)	276.0 (0.0)	2121.0 (0.0)	858.0 (0.0)	1263.0 (0.0)	0.6 (0.0)	0.01 (0.0)	0.0 (0.0)	35.28 (0.0)	12.0 (0.0)	201.0 (0.0)	67.0 (0.0)	7.0 (0.0)	4.0 (0.0)	236.0 (0.0)	33.0 (0.0)	3.0 (0.0)	0.04 (0.0)	0.7 (0.0)	0.23 (0.0)	0.02 (0.0)
25	12	178.08 (39.84)	164.25 (45.99)	152.33 (28.54)	9081.17 (1561.02)	4697.92 (946.22)	4383.25 (724.51)	0.48 (0.04)	0.03 (0.02)	0.0 (0.0)	55.66 (12.48)	29.92 (14.38)	69.5 (20.28)	48.5 (12.56)	30.17 (8.84)	36.83 (12.39)	68.17 (16.41)	29.83 (7.95)	17.5 (5.32)	0.17 (0.07)	0.39 (0.04)	0.28 (0.06)	0.17 (0.04)
26	28	145.93 (27.09)	115.11 (46.28)	94.75 (38.3)	1582.39 (1455.1)	861.46 (802.55)	720.93 (663.33)	0.48 (0.11)	0.03 (0.04)	0.0 (0.0)	58.07 (17.73)	67.68 (15.65)	52.14 (15.78)	19.68 (12.94)	6.43 (6.06)	48.07 (29.39)	33.71 (16.42)	10.07 (9.1)	2.89 (3.48)	0.48 (0.11)	0.35 (0.07)	0.13 (0.07)	0.04 (0.04)
27	1	51.0 (0.0)	49.0 (0.0)	50.0 (0.0)	24008.0 (0.0)	11983.0 (0.0)	12025.0 (0.0)	0.5 (0.0)	0.02 (0.0)	0.0 (0.0)	44.98 (0.0)	3.0 (0.0)	10.0 (0.0)	8.0 (0.0)	30.0 (0.0)	7.0 (0.0)	9.0 (0.0)	7.0 (0.0)	27.0 (0.0)	0.06 (0.0)	0.2 (0.0)	0.16 (0.0)	0.59 (0.0)
28	1	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)	3133.0 (0.0)	1579.0 (0.0)	1554.0 (0.0)	0.5 (0.0)	0.67 (0.0)	0.04 (0.0)	11.67 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	3.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	3.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)
The entire data set																							
0	81353	8.04 (20.18)	6.86 (17.09)	6.86 (18.67)	175.77 (679.4)	87.89 (342.49)	87.89 (340.55)	0.41 (0.29)	0.04 (0.14)	0.0 (0.0)	45.88 (64.64)	1.82 (4.83)	3.33 (8.79)	1.98 (5.72)	0.91 (3.11)	2.14 (6.64)	2.98 (8.42)	1.28 (3.93)	0.47 (1.82)	0.37 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.17)

Table C.20: Mean shift on data from the 4th month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
22 Clusters																							
0	68032	6.21 (11.78)	5.32 (10.32)	5.24 (10.9)	120.37 (388.82)	60.11 (194.75)	60.27 (196.74)	0.41 (0.29)	0.04 (0.13)	0.0 (0.0)	37.58 (62.35)	1.46 (2.66)	2.58 (5.23)	1.51 (3.57)	0.66 (1.94)	1.67 (3.55)	2.28 (5.08)	0.96 (2.46)	0.34 (1.18)	0.37 (0.4)	0.39 (0.37)	0.18 (0.28)	0.06 (0.17)
1	12	2.17 (0.37)	2.17 (0.37)	2.08 (0.49)	633.75 (563.07)	323.83 (289.74)	309.92 (277.75)	0.49 (0.04)	0.92 (0.2)	0.03 (0.01)	30.18 (46.69)	0.0 (0.0)	0.17 (0.37)	0.17 (0.37)	1.83 (0.37)	0.0 (0.0)	0.08 (0.28)	0.58 (0.76)	1.42 (0.76)	0.0 (0.0)	0.07 (0.16)	0.07 (0.16)	0.86 (0.2)
2	35	159.31 (47.41)	145.71 (40.62)	149.97 (41.49)	8055.43 (1962.0)	4075.26 (1088.65)	3980.17 (942.46)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	47.84 (9.23)	18.66 (14.29)	52.69 (20.44)	48.17 (15.63)	39.8 (8.38)	26.66 (15.05)	60.49 (21.17)	37.77 (12.39)	25.06 (6.52)	0.11 (0.06)	0.32 (0.05)	0.3 (0.05)	0.27 (0.08)
3	5	283.6 (38.48)	186.8 (27.65)	277.4 (42.89)	2950.2 (234.46)	1365.4 (121.56)	1584.8 (137.15)	0.54 (0.02)	0.01 (0.01)	0.0 (0.0)	27.4 (7.91)	94.2 (17.08)	95.6 (23.95)	74.0 (23.87)	19.8 (5.23)	116.8 (14.99)	107.8 (20.17)	45.6 (11.64)	7.2 (2.23)	0.33 (0.04)	0.34 (0.08)	0.26 (0.07)	0.07 (0.02)
4	29	180.07 (43.07)	158.28 (37.28)	168.59 (39.68)	4157.03 (1003.85)	2094.76 (543.45)	2062.28 (496.26)	0.5 (0.03)	0.03 (0.02)	0.0 (0.0)	44.83 (11.52)	26.1 (13.2)	72.76 (21.82)	55.62 (16.04)	25.59 (5.96)	41.48 (16.33)	77.28 (24.1)	37.66 (11.26)	12.17 (4.07)	0.14 (0.05)	0.4 (0.05)	0.31 (0.05)	0.15 (0.05)
5	69	152.65 (30.94)	131.29 (27.7)	145.06 (29.1)	2052.84 (864.28)	1026.16 (460.76)	1026.68 (431.8)	0.5 (0.06)	0.03 (0.03)	0.0 (0.0)	45.26 (11.69)	19.42 (12.05)	75.61 (23.57)	46.14 (12.63)	11.48 (5.82)	34.58 (17.22)	81.43 (24.9)	24.36 (7.27)	4.68 (3.76)	0.12 (0.06)	0.49 (0.09)	0.31 (0.08)	0.08 (0.04)
6	1	1121.0 (0.0)	761.0 (0.0)	1038.0 (0.0)	4223.0 (0.0)	2302.0 (0.0)	1921.0 (0.0)	0.45 (0.0)	0.01 (0.0)	0.0 (0.0)	23.11 (0.0)	405.0 (0.0)	639.0 (0.0)	66.0 (0.0)	11.0 (0.0)	815.0 (0.0)	203.0 (0.0)	15.0 (0.0)	5.0 (0.0)	0.36 (0.0)	0.57 (0.0)	0.06 (0.0)	0.01 (0.0)
7	1	806.0 (0.0)	468.0 (0.0)	795.0 (0.0)	2008.0 (0.0)	875.0 (0.0)	1133.0 (0.0)	0.56 (0.0)	0.0 (0.0)	0.0 (0.0)	22.3 (0.0)	319.0 (0.0)	457.0 (0.0)	30.0 (0.0)	0.0 (0.0)	594.0 (0.0)	198.0 (0.0)	3.0 (0.0)	0.0 (0.0)	0.4 (0.0)	0.57 (0.0)	0.04 (0.0)	0.0 (0.0)
8	1	714.0 (0.0)	286.0 (0.0)	713.0 (0.0)	2053.0 (0.0)	711.0 (0.0)	1342.0 (0.0)	0.65 (0.0)	0.01 (0.0)	0.0 (0.0)	31.32 (0.0)	411.0 (0.0)	256.0 (0.0)	45.0 (0.0)	2.0 (0.0)	467.0 (0.0)	230.0 (0.0)	14.0 (0.0)	2.0 (0.0)	0.58 (0.0)	0.36 (0.0)	0.06 (0.0)	0.0 (0.0)
9	1	524.0 (0.0)	450.0 (0.0)	453.0 (0.0)	3372.0 (0.0)	1604.0 (0.0)	1768.0 (0.0)	0.52 (0.0)	0.0 (0.0)	0.0 (0.0)	27.02 (0.0)	126.0 (0.0)	254.0 (0.0)	136.0 (0.0)	8.0 (0.0)	125.0 (0.0)	288.0 (0.0)	36.0 (0.0)	4.0 (0.0)	0.24 (0.0)	0.48 (0.0)	0.26 (0.0)	0.02 (0.0)
10	1	399.0 (0.0)	367.0 (0.0)	354.0 (0.0)	7414.0 (0.0)	3749.0 (0.0)	3665.0 (0.0)	0.49 (0.0)	0.04 (0.0)	0.0 (0.0)	39.82 (0.0)	59.0 (0.0)	146.0 (0.0)	125.0 (0.0)	69.0 (0.0)	87.0 (0.0)	139.0 (0.0)	106.0 (0.0)	22.0 (0.0)	0.15 (0.0)	0.37 (0.0)	0.31 (0.0)	0.17 (0.0)
11	1	397.0 (0.0)	326.0 (0.0)	389.0 (0.0)	10111.0 (0.0)	4149.0 (0.0)	5962.0 (0.0)	0.59 (0.0)	0.04 (0.0)	0.0 (0.0)	41.91 (0.0)	50.0 (0.0)	168.0 (0.0)	114.0 (0.0)	65.0 (0.0)	69.0 (0.0)	188.0 (0.0)	99.0 (0.0)	33.0 (0.0)	0.13 (0.0)	0.42 (0.0)	0.29 (0.0)	0.16 (0.0)
12	1	331.0 (0.0)	248.0 (0.0)	327.0 (0.0)	3141.0 (0.0)	1229.0 (0.0)	1912.0 (0.0)	0.61 (0.0)	0.01 (0.0)	0.0 (0.0)	33.85 (0.0)	75.0 (0.0)	125.0 (0.0)	122.0 (0.0)	9.0 (0.0)	86.0 (0.0)	139.0 (0.0)	101.0 (0.0)	1.0 (0.0)	0.23 (0.0)	0.38 (0.0)	0.37 (0.0)	0.03 (0.0)
13	1	308.0 (0.0)	303.0 (0.0)	77.0 (0.0)	913.0 (0.0)	763.0 (0.0)	150.0 (0.0)	0.16 (0.0)	0.01 (0.0)	0.0 (0.0)	35.88 (0.0)	175.0 (0.0)	112.0 (0.0)	19.0 (0.0)	2.0 (0.0)	47.0 (0.0)	28.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.57 (0.0)	0.36 (0.0)	0.06 (0.0)	0.01 (0.0)
14	4	259.5 (30.6)	201.0 (30.27)	254.0 (31.03)	1915.25 (471.82)	900.25 (196.31)	1015.0 (279.22)	0.53 (0.02)	0.01 (0.01)	0.0 (0.0)	33.06 (9.84)	57.25 (8.79)	120.5 (26.09)	76.0 (19.09)	5.75 (5.45)	80.25 (24.54)	148.75 (23.07)	22.75 (12.36)	2.25 (2.28)	0.22 (0.04)	0.46 (0.06)	0.29 (0.07)	0.02 (0.02)
15	1	294.0 (0.0)	221.0 (0.0)	294.0 (0.0)	9730.0 (0.0)	4441.0 (0.0)	5289.0 (0.0)	0.54 (0.0)	0.02 (0.0)	0.0 (0.0)	39.44 (0.0)	71.0 (0.0)	71.0 (0.0)	86.0 (0.0)	66.0 (0.0)	75.0 (0.0)	103.0 (0.0)	82.0 (0.0)	34.0 (0.0)	0.24 (0.0)	0.24 (0.0)	0.29 (0.0)	0.22 (0.0)
16	6	202.0 (25.99)	114.33 (28.78)	188.33 (26.6)	1583.83 (1645.12)	774.33 (831.59)	809.5 (816.96)	0.54 (0.05)	0.01 (0.01)	0.0 (0.0)	31.17 (5.21)	96.17 (23.0)	82.67 (15.07)	21.0 (10.82)	2.17 (2.34)	131.0 (24.53)	51.83 (19.09)	4.17 (4.63)	1.33 (1.97)	0.48 (0.08)	0.41 (0.07)	0.1 (0.05)	0.01 (0.01)
17	3	176.67 (16.54)	160.0 (14.31)	172.67 (18.62)	2406.67 (236.0)	1126.33 (64.61)	1280.33 (171.63)	0.53 (0.02)	0.01 (0.0)	0.0 (0.0)	31.05 (5.94)	15.0 (2.16)	44.67 (10.14)	107.67 (20.04)	9.33 (3.09)	18.0 (5.72)	78.33 (10.53)	76.0 (21.6)	0.33 (0.47)	0.09 (0.01)	0.25 (0.05)	0.61 (0.08)	0.05 (0.02)
18	193	118.18 (21.61)	97.38 (20.36)	106.39 (23.36)	2709.59 (1469.1)	1351.41 (760.07)	1358.18 (732.88)	0.5 (0.05)	0.03 (0.03)	0.0 (0.0)	45.25 (13.26)	27.33 (17.74)	45.48 (14.27)	29.59 (9.6)	15.78 (8.41)	34.06 (19.86)	43.5 (13.56)	20.5 (8.35)	8.33 (5.85)	0.22 (0.12)	0.38 (0.08)	0.26 (0.08)	0.14 (0.09)
19	3	132.67 (23.75)	123.0 (15.94)	121.33 (26.95)	18677.67 (3339.65)	9447.33 (1001.46)	9230.33 (2559.01)	0.49 (0.05)	0.02 (0.01)	0.0 (0.0)	38.5 (7.75)	17.33 (7.13)	37.33 (7.76)	32.67 (7.59)	45.33 (9.74)	19.67 (6.55)	42.0 (8.6)	25.67 (6.18)	34.0 (6.48)	0.13 (0.04)	0.28 (0.01)	0.25 (0.06)	0.34 (0.02)
20	81	67.02 (26.79)	62.33 (25.09)	60.86 (24.77)	7602.04 (1832.12)	3880.2 (934.28)	3721.84 (974.85)	0.49 (0.04)	0.04 (0.02)	0.0 (0.0)	48.06 (16.85)	8.48 (9.32)	20.84 (11.29)	17.51 (9.52)	20.2 (7.94)	11.15 (6.77)	22.11 (11.71)	13.4 (6.89)	14.21 (5.79)	0.12 (0.11)	0.3 (0.1)	0.25 (0.07)	0.33 (0.12)
21	1	2.0 (0.0)	2.0 (0.0)	2.0 (0.0)	1074.0 (0.0)	471.0 (0.0)	603.0 (0.0)	0.56 (0.0)	1.0 (0.0)	0.06 (0.0)	2.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)
The entire data set																							
0	68482	7.04 (17.05)	6.0 (14.23)	6.0 (15.87)	146.3 (558.59)	73.15 (281.31)	73.15 (280.09)	0.41 (0.29)	0.04 (0.13)	0.0 (0.0)	37.62 (62.16)	1.62 (4.46)	2.91 (7.57)	1.73 (4.76)	0.77 (2.58)	1.89 (6.39)	2.6 (7.07)	1.1 (3.25)	0.4 (1.57)	0.37 (0.4)	0.39 (0.37)	0.18 (0.27)	0.06 (0.17)

Table C.21: Mean shift on data from the 5th month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
27 Clusters																							
0	84628	7.34 (15.7)	6.3 (13.7)	6.21 (14.44)	152.37 (554.21)	76.16 (279.06)	76.21 (279.08)	0.4 (0.29)	0.04 (0.14)	0.0 (0.0)	49.47 (67.68)	1.69 (3.53)	3.04 (6.84)	1.8 (4.6)	0.8 (2.55)	1.94 (4.66)	2.7 (6.6)	1.16 (3.21)	0.41 (1.52)	0.38 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)
1	20	2.25 (0.43)	2.2 (0.4)	2.2 (0.51)	1160.75 (1454.51)	565.45 (699.18)	595.3 (773.06)	0.51 (0.08)	0.89 (0.23)	0.02 (0.01)	35.52 (43.97)	0.05 (0.22)	0.1 (0.3)	0.15 (0.36)	1.95 (0.59)	0.05 (0.22)	0.2 (0.4)	0.2 (0.4)	1.75 (0.7)	0.02 (0.07)	0.05 (0.15)	0.07 (0.16)	0.87 (0.21)
2	142	184.06 (46.99)	163.68 (40.99)	172.29 (46.49)	6014.8 (1865.03)	3001.12 (979.57)	3013.68 (961.71)	0.5 (0.04)	0.03 (0.02)	0.0 (0.0)	57.7 (11.54)	25.75 (14.12)	67.46 (23.88)	56.41 (18.82)	34.44 (9.14)	38.39 (17.01)	73.15 (24.36)	41.96 (14.8)	18.78 (6.54)	0.14 (0.06)	0.36 (0.07)	0.31 (0.06)	0.2 (0.06)
3	16	148.5 (48.53)	134.06 (43.34)	141.81 (48.9)	13269.81 (2284.07)	6605.62 (1198.85)	6664.19 (1205.46)	0.5 (0.03)	0.03 (0.02)	0.0 (0.0)	57.64 (10.28)	15.94 (6.51)	45.06 (17.53)	42.56 (16.98)	44.94 (13.83)	23.19 (11.72)	54.19 (23.16)	32.5 (12.81)	31.94 (8.79)	0.11 (0.04)	0.3 (0.05)	0.28 (0.04)	0.31 (0.05)
4	7	379.86 (53.55)	274.71 (26.18)	376.0 (53.04)	3232.14 (772.35)	1532.57 (437.73)	1699.57 (353.61)	0.53 (0.05)	0.0 (0.0)	0.0 (0.0)	33.04 (7.31)	102.43 (27.99)	156.43 (30.44)	104.86 (21.3)	16.14 (9.33)	122.43 (31.75)	196.57 (25.02)	53.57 (18.31)	3.43 (3.81)	0.26 (0.04)	0.41 (0.07)	0.28 (0.06)	0.04 (0.02)
5	10	231.2 (43.48)	143.6 (35.41)	226.2 (42.14)	1527.2 (847.65)	756.2 (463.73)	771.0 (389.35)	0.52 (0.06)	0.01 (0.01)	0.0 (0.0)	35.69 (10.11)	87.7 (19.0)	108.3 (30.26)	27.3 (13.02)	7.9 (7.73)	142.3 (42.78)	67.8 (22.68)	13.0 (11.98)	3.1 (4.09)	0.38 (0.06)	0.46 (0.06)	0.12 (0.05)	0.04 (0.04)
6	1	942.0 (0.0)	692.0 (0.0)	783.0 (0.0)	5904.0 (0.0)	3397.0 (0.0)	2507.0 (0.0)	0.42 (0.0)	0.01 (0.0)	0.0 (0.0)	30.84 (0.0)	350.0 (0.0)	502.0 (0.0)	67.0 (0.0)	23.0 (0.0)	619.0 (0.0)	135.0 (0.0)	20.0 (0.0)	9.0 (0.0)	0.37 (0.0)	0.53 (0.0)	0.07 (0.0)	0.02 (0.0)
7	1	907.0 (0.0)	605.0 (0.0)	903.0 (0.0)	9239.0 (0.0)	4653.0 (0.0)	4586.0 (0.0)	0.5 (0.0)	0.0 (0.0)	0.0 (0.0)	30.56 (0.0)	303.0 (0.0)	344.0 (0.0)	222.0 (0.0)	38.0 (0.0)	314.0 (0.0)	491.0 (0.0)	82.0 (0.0)	16.0 (0.0)	0.33 (0.0)	0.38 (0.0)	0.24 (0.0)	0.04 (0.0)
8	1	542.0 (0.0)	537.0 (0.0)	119.0 (0.0)	7625.0 (0.0)	4633.0 (0.0)	2992.0 (0.0)	0.39 (0.0)	0.02 (0.0)	0.0 (0.0)	45.11 (0.0)	302.0 (0.0)	156.0 (0.0)	52.0 (0.0)	32.0 (0.0)	27.0 (0.0)	50.0 (0.0)	28.0 (0.0)	14.0 (0.0)	0.56 (0.0)	0.29 (0.0)	0.1 (0.0)	0.06 (0.0)
9	1	454.0 (0.0)	299.0 (0.0)	431.0 (0.0)	2168.0 (0.0)	997.0 (0.0)	1171.0 (0.0)	0.54 (0.0)	0.01 (0.0)	0.0 (0.0)	40.8 (0.0)	175.0 (0.0)	261.0 (0.0)	13.0 (0.0)	5.0 (0.0)	371.0 (0.0)	54.0 (0.0)	2.0 (0.0)	4.0 (0.0)	0.39 (0.0)	0.57 (0.0)	0.03 (0.0)	0.01 (0.0)
10	1	424.0 (0.0)	278.0 (0.0)	409.0 (0.0)	6499.0 (0.0)	2934.0 (0.0)	3565.0 (0.0)	0.55 (0.0)	0.01 (0.0)	0.0 (0.0)	44.83 (0.0)	153.0 (0.0)	227.0 (0.0)	19.0 (0.0)	25.0 (0.0)	314.0 (0.0)	64.0 (0.0)	15.0 (0.0)	16.0 (0.0)	0.36 (0.0)	0.54 (0.0)	0.04 (0.0)	0.06 (0.0)
11	1	422.0 (0.0)	247.0 (0.0)	417.0 (0.0)	2542.0 (0.0)	1248.0 (0.0)	1294.0 (0.0)	0.51 (0.0)	0.01 (0.0)	0.0 (0.0)	37.14 (0.0)	177.0 (0.0)	178.0 (0.0)	55.0 (0.0)	12.0 (0.0)	245.0 (0.0)	144.0 (0.0)	24.0 (0.0)	4.0 (0.0)	0.42 (0.0)	0.42 (0.0)	0.13 (0.0)	0.03 (0.0)
12	6	341.5 (36.16)	299.83 (39.58)	328.5 (32.89)	6115.5 (1767.44)	3191.0 (1004.19)	2924.5 (807.12)	0.48 (0.04)	0.01 (0.01)	0.0 (0.0)	43.61 (4.44)	45.17 (13.84)	152.0 (32.48)	109.83 (19.39)	34.5 (4.79)	87.33 (22.49)	165.33 (27.6)	63.67 (14.99)	12.17 (2.03)	0.13 (0.04)	0.44 (0.08)	0.32 (0.05)	0.1 (0.02)
13	1	407.0 (0.0)	227.0 (0.0)	401.0 (0.0)	6859.0 (0.0)	3537.0 (0.0)	3322.0 (0.0)	0.48 (0.0)	0.0 (0.0)	0.0 (0.0)	21.99 (0.0)	179.0 (0.0)	114.0 (0.0)	57.0 (0.0)	57.0 (0.0)	214.0 (0.0)	102.0 (0.0)	57.0 (0.0)	28.0 (0.0)	0.44 (0.0)	0.28 (0.0)	0.14 (0.0)	0.14 (0.0)
14	7	307.71 (39.91)	210.0 (52.25)	277.14 (43.63)	3264.43 (1129.22)	1534.43 (572.4)	1730.0 (606.11)	0.54 (0.05)	0.02 (0.01)	0.0 (0.0)	48.27 (10.62)	114.14 (29.12)	124.29 (31.83)	52.29 (15.87)	17.0 (7.35)	132.29 (36.28)	112.43 (31.7)	25.0 (10.17)	7.43 (4.17)	0.37 (0.07)	0.4 (0.07)	0.17 (0.06)	0.06 (0.03)
15	1	333.0 (0.0)	320.0 (0.0)	300.0 (0.0)	21016.0 (0.0)	11702.0 (0.0)	9314.0 (0.0)	0.44 (0.0)	0.04 (0.0)	0.0 (0.0)	56.67 (0.0)	37.0 (0.0)	97.0 (0.0)	107.0 (0.0)	92.0 (0.0)	50.0 (0.0)	113.0 (0.0)	82.0 (0.0)	55.0 (0.0)	0.11 (0.0)	0.29 (0.0)	0.32 (0.0)	0.28 (0.0)
16	1	314.0 (0.0)	277.0 (0.0)	313.0 (0.0)	17991.0 (0.0)	9322.0 (0.0)	8669.0 (0.0)	0.48 (0.0)	0.02 (0.0)	0.0 (0.0)	50.8 (0.0)	33.0 (0.0)	77.0 (0.0)	119.0 (0.0)	85.0 (0.0)	52.0 (0.0)	116.0 (0.0)	101.0 (0.0)	44.0 (0.0)	0.11 (0.0)	0.25 (0.0)	0.38 (0.0)	0.27 (0.0)
17	1	308.0 (0.0)	112.0 (0.0)	308.0 (0.0)	577.0 (0.0)	199.0 (0.0)	378.0 (0.0)	0.66 (0.0)	0.01 (0.0)	0.0 (0.0)	39.93 (0.0)	196.0 (0.0)	108.0 (0.0)	4.0 (0.0)	0.0 (0.0)	270.0 (0.0)	36.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.64 (0.0)	0.35 (0.0)	0.01 (0.0)	0.0 (0.0)
18	65	217.97 (41.03)	174.51 (36.6)	207.23 (40.04)	2220.69 (973.48)	1088.58 (512.18)	1132.11 (493.41)	0.51 (0.05)	0.02 (0.02)	0.0 (0.0)	53.25 (15.22)	42.86 (19.73)	109.52 (27.74)	53.18 (17.38)	12.4 (8.1)	69.32 (25.23)	106.14 (30.31)	27.88 (12.27)	3.89 (3.26)	0.19 (0.08)	0.51 (0.1)	0.24 (0.06)	0.06 (0.04)
19	1	303.0 (0.0)	234.0 (0.0)	303.0 (0.0)	21069.0 (0.0)	9571.0 (0.0)	11498.0 (0.0)	0.55 (0.0)	0.02 (0.0)	0.0 (0.0)	54.92 (0.0)	57.0 (0.0)	80.0 (0.0)	80.0 (0.0)	86.0 (0.0)	62.0 (0.0)	107.0 (0.0)	73.0 (0.0)	61.0 (0.0)	0.19 (0.0)	0.26 (0.0)	0.26 (0.0)	0.28 (0.0)
20	18	114.0 (65.55)	104.39 (60.24)	96.17 (50.45)	11879.83 (1557.8)	6149.17 (817.77)	5730.67 (889.88)	0.48 (0.03)	0.03 (0.01)	0.0 (0.0)	56.51 (17.42)	20.06 (20.11)	43.78 (28.7)	28.06 (16.26)	22.11 (8.81)	26.83 (20.69)	36.0 (20.82)	18.39 (11.89)	14.94 (6.05)	0.15 (0.07)	0.38 (0.07)	0.25 (0.05)	0.23 (0.1)
21	5	193.2 (36.65)	163.4 (59.13)	86.2 (63.85)	1203.4 (923.59)	686.4 (415.63)	517.0 (523.36)	0.31 (0.2)	0.01 (0.01)	0.0 (0.0)	47.3 (8.77)	100.8 (37.25)	69.8 (15.46)	17.0 (9.32)	5.6 (6.56)	39.8 (30.35)	36.6 (24.21)	7.6 (7.71)	2.2 (2.71)	0.51 (0.13)	0.37 (0.08)	0.09 (0.05)	0.03 (0.04)
22	12	196.92 (35.67)	82.83 (22.13)	192.08 (35.35)	873.17 (563.22)	384.33 (320.49)	488.83 (248.09)	0.59 (0.07)	0.02 (0.01)	0.0 (0.0)	43.88 (18.83)	112.0 (23.16)	67.0 (15.28)	15.42 (10.57)	2.5 (2.29)	130.33 (30.8)	54.67 (15.03)	5.75 (5.07)	1.33 (1.7)	0.57 (0.06)	0.34 (0.04)	0.08 (0.05)	0.01 (0.01)
23	13	119.54 (38.0)	112.0 (35.5)	113.85 (37.82)	10059.31 (1306.93)	5162.69 (875.76)	4896.62 (599.76)	0.49 (0.04)	0.05 (0.04)	0.0 (0.0)	56.53 (21.89)	10.15 (3.16)	31.69 (15.01)	33.46 (12.45)	44.23 (12.81)	17.08 (6.15)	36.54 (16.13)	28.38 (12.14)	31.85 (7.74)	0.09 (0.03)	0.25 (0.07)	0.28 (0.05)	0.38 (0.06)
24	6	112.17 (50.46)	105.5 (47.75)	106.5 (49.37)	15779.5 (2696.85)	7712.67 (1922.27)	8066.83 (1086.1)	0.52 (0.06)	0.03 (0.01)	0.0 (0.0)	53.09 (11.75)	10.5 (4.03)	38.0 (25.39)	29.17 (14.25)	34.5 (11.3)	22.67 (15.89)	35.0 (22.05)	24.83 (10.45)	24.0 (7.57)	0.1 (0.03)	0.29 (0.11)	0.25 (0.06)	0.35 (0.12)
25	1	2.0 (0.0)	2.0 (0.0)	2.0 (0.0)	3221.0 (0.0)	1953.0 (0.0)	1268.0 (0.0)	0.39 (0.0)	1.0 (0.0)	0.05 (0.0)	1.5 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)
26	1	2.0 (0.0)	2.0 (0.0)	2.0 (0.0)	2544.0 (0.0)	1163.0 (0.0)	1381.0 (0.0)	0.54 (0.0)	1.0 (0.0)	0.1 (0.0)	23.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	2.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)
The entire data set																							
0	84968	8.07 (20.58)	6.9 (17.38)	6.9 (19.07)	174.06 (701.32)	87.03 (353.69)	87.03 (351.63)	0.4 (0.29)	0.04 (0.14)	0.0 (0.0)	49.48 (67.55)	1.85 (5.14)	3.34 (8.89)	1.98 (5.73)	0.9 (3.17)	2.16 (6.9)	2.99 (8.59)	1.29 (3.95)	0.46 (1.88)	0.38 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)

Table C.22: DBSCAN on data from the 1st month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
2 Clusters + outliers																							
-1	1438	84.69 (61.42)	71.62 (49.17)	76.48 (58.41)	2727.21 (2125.96)	1367.46 (1071.89)	1359.75 (1084.75)	0.49 (0.09)	0.1 (0.23)	0.0 (0.0)	50.08 (71.2)	17.09 (20.91)	34.04 (29.65)	21.9 (17.92)	11.66 (9.5)	22.63 (25.93)	32.9 (29.43)	14.34 (11.85)	6.6 (6.19)	0.17 (0.14)	0.36 (0.16)	0.25 (0.14)	0.22 (0.21)
0	85100	5.95 (10.25)	5.07 (8.98)	4.99 (9.49)	104.74 (294.62)	52.3 (147.84)	52.43 (148.85)	0.41 (0.3)	0.04 (0.13)	0.0 (0.0)	32.63 (38.0)	1.42 (2.31)	2.49 (4.64)	1.43 (3.12)	0.61 (1.68)	1.59 (3.09)	2.18 (4.48)	0.91 (2.16)	0.3 (1.0)	0.38 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)
1	101	1.0 (0.0)	1.0 (0.0)	0.0 (0.0)	1.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	683.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
The entire data set																							
0	86639	7.25 (16.34)	6.17 (13.85)	6.17 (15.12)	148.14 (522.07)	74.07 (262.25)	74.07 (263.05)	0.41 (0.29)	0.04 (0.13)	0.0 (0.0)	33.68 (44.72)	1.68 (4.06)	3.01 (7.21)	1.77 (4.66)	0.79 (2.5)	1.94 (5.27)	2.68 (7.04)	1.13 (3.14)	0.41 (1.5)	0.38 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)

Table C.23: DBSCAN on data from the 2nd month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
1 Cluster + outliers																							
-1	1398	94.41 (67.07)	80.37 (54.67)	85.21 (63.11)	3230.77 (2601.36)	1628.67 (1328.74)	1602.1 (1306.9)	0.49 (0.09)	0.11 (0.24)	0.0 (0.0)	48.5 (32.97)	18.71 (23.37)	37.52 (31.45)	24.77 (19.08)	13.41 (10.71)	24.89 (29.44)	36.42 (29.7)	16.47 (13.26)	7.43 (6.87)	0.17 (0.14)	0.34 (0.15)	0.26 (0.14)	0.23 (0.21)
0	82715	6.14 (11.21)	5.24 (9.84)	5.16 (10.33)	111.22 (327.82)	55.39 (163.6)	55.84 (166.94)	0.41 (0.29)	0.04 (0.13)	0.0 (0.0)	38.32 (47.06)	1.44 (2.46)	2.57 (5.04)	1.48 (3.37)	0.64 (1.81)	1.63 (3.3)	2.25 (4.85)	0.95 (2.36)	0.32 (1.07)	0.37 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)
The entire data set																							
0	84113	7.6 (18.05)	6.49 (15.4)	6.49 (16.61)	163.07 (614.17)	81.54 (310.03)	81.54 (308.01)	0.41 (0.29)	0.04 (0.14)	0.0 (0.0)	38.49 (46.88)	1.73 (4.46)	3.15 (7.83)	1.87 (5.11)	0.85 (2.79)	2.02 (5.83)	2.82 (7.54)	1.21 (3.51)	0.44 (1.65)	0.37 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.17)

Table C.24: DBSCAN on data from the 3rd month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. de-gree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neigh-bors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 mes-sage out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 mes-sage (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
2 Clusters + outliers																							
-1	1204	110.24 (81.06)	92.99 (65.11)	100.13 (77.25)	3800.95 (2969.04)	1901.18 (1520.31)	1899.78 (1482.78)	0.5 (0.1)	0.09 (0.2)	0.0 (0.0)	92.13 (176.79)	22.05 (25.82)	43.4 (38.71)	28.61 (23.95)	16.18 (12.64)	29.55 (36.31)	42.07 (36.39)	19.46 (16.29)	9.05 (7.81)	0.18 (0.15)	0.36 (0.17)	0.24 (0.14)	0.21 (0.2)
0	80083	6.51 (12.48)	5.57 (10.95)	5.47 (11.48)	121.41 (368.53)	60.7 (185.02)	60.72 (186.16)	0.41 (0.3)	0.04 (0.14)	0.0 (0.0)	44.38 (54.29)	1.52 (2.73)	2.73 (5.6)	1.58 (3.72)	0.68 (1.96)	1.73 (3.68)	2.39 (5.37)	1.01 (2.58)	0.34 (1.15)	0.37 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)
1	66	1.0 (0.0)	1.0 (0.0)	0.0 (0.0)	1.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1027.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
The entire data set																							
0	81353	8.04 (20.18)	6.86 (17.09)	6.86 (18.67)	175.77 (679.4)	87.89 (342.49)	87.89 (340.55)	0.41 (0.29)	0.04 (0.14)	0.0 (0.0)	45.88 (64.64)	1.82 (4.83)	3.33 (8.79)	1.98 (5.72)	0.91 (3.11)	2.14 (6.64)	2.98 (8.42)	1.28 (3.93)	0.47 (1.82)	0.37 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.17)

Table C.25: DBSCAN on data from the 4th month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
1 Cluster + outliers																							
-1	1131	87.0 (71.66)	73.52 (54.8)	79.8 (68.06)	2879.1 (2462.18)	1443.26 (1252.44)	1435.84 (1233.32)	0.5 (0.09)	0.11 (0.25)	0.0 (0.0)	69.76 (146.67)	16.82 (25.78)	34.67 (35.2)	22.92 (18.74)	12.58 (10.29)	23.27 (38.29)	34.11 (29.74)	15.21 (12.89)	7.2 (6.75)	0.17 (0.15)	0.36 (0.17)	0.25 (0.13)	0.22 (0.2)
0	67351	5.69 (10.01)	4.87 (8.81)	4.76 (9.24)	100.41 (296.55)	50.14 (148.57)	50.27 (150.07)	0.41 (0.3)	0.04 (0.13)	0.0 (0.0)	37.08 (59.58)	1.37 (2.27)	2.38 (4.5)	1.37 (3.08)	0.57 (1.62)	1.53 (3.01)	2.07 (4.36)	0.87 (2.13)	0.29 (0.97)	0.37 (0.41)	0.39 (0.38)	0.17 (0.28)	0.06 (0.16)
The entire data set																							
0	68482	7.04 (17.05)	6.0 (14.23)	6.0 (15.87)	146.3 (558.59)	73.15 (281.31)	73.15 (280.09)	0.41 (0.29)	0.04 (0.13)	0.0 (0.0)	37.62 (62.16)	1.62 (4.46)	2.91 (7.57)	1.73 (4.76)	0.77 (2.58)	1.89 (6.39)	2.6 (7.07)	1.1 (3.25)	0.4 (1.57)	0.37 (0.4)	0.39 (0.37)	0.18 (0.27)	0.06 (0.17)

Table C.26: DBSCAN on data from the 5th month

Cluster	# of nodes	Degree	Out-degree	In-degree	W. degree	W. in-degree	W. out-degree	$\frac{Out}{Total}$	CC	W. CC	Non-common neighbors	1 msg	Less than 7 msgs	Less than 32 msgs	More than 32 msgs	1 message out	Less than 7 msgs out	Less than 32 msgs out	More than 32 msgs out	1 message (%)	Less than 7 msgs (%)	Less than 32 msgs (%)	More than 32 msgs (%)
3 Cluster + outliers																							
-1	1329	112.18 (81.63)	94.24 (65.32)	102.6 (77.4)	3912.82 (3053.51)	1961.73 (1561.11)	1951.09 (1535.12)	0.5 (0.09)	0.1 (0.23)	0.0 (0.0)	63.47 (81.83)	22.48 (28.36)	44.51 (38.16)	28.77 (22.8)	16.43 (12.85)	30.86 (37.56)	42.94 (36.86)	19.52 (15.73)	9.27 (8.28)	0.18 (0.15)	0.35 (0.15)	0.25 (0.13)	0.23 (0.21)
0	83524	6.42 (12.23)	5.52 (10.77)	5.38 (11.25)	114.81 (355.0)	57.32 (177.96)	57.49 (179.21)	0.4 (0.3)	0.04 (0.14)	0.0 (0.0)	48.05 (58.82)	1.52 (2.69)	2.69 (5.46)	1.56 (3.68)	0.66 (1.93)	1.71 (3.57)	2.36 (5.26)	1.0 (2.55)	0.32 (1.12)	0.38 (0.41)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)
1	67	1.0 (0.0)	1.0 (0.0)	0.0 (0.0)	1.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	931.6 (15.51)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
2	48	1.0 (0.0)	0.98 (0.14)	1.0 (0.0)	3.21 (1.59)	1.54 (0.87)	1.67 (1.01)	0.52 (0.13)	0.0 (0.0)	0.0 (0.0)	930.79 (15.91)	0.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.6 (0.49)	0.4 (0.49)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)	0.0 (0.0)	0.0 (0.0)
The entire data set																							
0	84968	8.07 (20.58)	6.9 (17.38)	6.9 (19.07)	174.06 (701.32)	87.03 (353.69)	87.03 (351.63)	0.4 (0.29)	0.04 (0.14)	0.0 (0.0)	49.48 (67.55)	1.85 (5.14)	3.34 (8.89)	1.98 (5.73)	0.9 (3.17)	2.16 (6.9)	2.99 (8.59)	1.29 (3.95)	0.46 (1.88)	0.38 (0.4)	0.39 (0.37)	0.17 (0.27)	0.06 (0.16)

Anna Fridtun Aarekol

A graph theoretical approach to online predator detection

NTNU
Norwegian University of
Science and Technology