



University of
Zurich^{UZH}

Analysis of BitTorrent Measurements

Romana Pernischova
Zurich, Switzerland
Student ID: 12-917-332

Supervisor: Andri Lareida, Thomas Bocek
Date of Submission: August 22, 2016

Abstract

Das VIOLA BitTorrent Messungssystem stellt Daten über das BitTorrent Netzwerk für eine detaillierte Analyse der Netzwerktopologie zur Verfügung. Um eine solches Gutachten durchführen zu können, wird mittels dem Framework Hadoop ein MapReduce auf die Daten angewendet. Das BitTorrent Netzwerk ist in einer Form vorhanden, in welcher Verbindungen jeweils zwischen einem Teilnehmer und einer Datei, also einem Torrent bestehen. Ein solches Zwei-Moden-Netzwerk lässt sich jedoch nicht analysieren. Die Datentransformation ermöglicht es das Netzwerk in einen Graphen zu verwandeln, in welchem Torrents Knoten darstellen und die Verbindungen durch einen Peer gebildet werden, welche die beiden Torrents im Netzwerk anbietet.

Des weiteren wurde der kreierte Graph anschliessend mit R analysiert. Dafür wurde die igraph Bibliothek verwendet, welche die Berechnung von Kennzahlen der sozialen Netzwerkanalyse (SNA) ermöglicht. Da sich das Netzwerk jedoch ständig verändert und nicht statisch ist, müssen die Kennzahlen in Abhängigkeit mit der Zeit gebracht werden. Somit wurden die R Berechnungen vermehrt auf den Daten ausgeführt, um auch die Entwicklung des Netzwerkes beobachten zu können. Die Daten, welche das Netzwerk dokumentieren, sind dabei stündlich vorhanden.

Auf Grund der Grösse des BitTorrent Netzwerkes wurden zudem auch reduzierte Graphen kreiert. Während der Reduktion wurden Teilnehmer ausgeschlossen, welche nicht wirklich am Austausch der Dateien teilnehmen, sondern zum Beispiel für Monitoring Zwecke Teil des Netzwerkes sind. Solche Peers wurden anhand der Anzahl der Torrents evaluiert, die sie im Netz anzubieten scheinen. Eine eindeutige Grenze konnte nicht auf Anhieb bestimmen werden, jedoch wurde aus den Berechnungen ersichtlich, dass diese zwischen 1000 und 500 Torrents liegt. Wurden Teilnehmer ausgeschlossen die über 200 oder sogar 100 Torrents anbieten, änderten sich die Kennzahlen nicht im Vergleich zu dem Graphen, in welchem Peers ausgeschlossen wurden, welche über 500 Torrents anbieten.

Das aus Torrents geformte Netzwerk weist ein mittlere Pfadlänge von 2 auf. Zur gleichen Zeit hat das Netzwerk einen hohen Clustering-Koeffizient von 0.3, wenn man ihn mit dem Koeffizienten eines Zufallsnetzwerkes der gleichen Grösse vergleicht, welcher 0.08 ist. Weil die Gradverteilung eine Potenzgesetz-Verteilung hat, wird das BitTorrent Netzwerk als ein skalenfreies Netzwerk klassifiziert. In einem solchen Netzwerk gilt das Prinzip der Vorzugsbefestigung. Wenn ein neues Torrent in das Netzwerk kommt, dann wird es sich mit grösster Wahrscheinlichkeit an ein Torrent binden, das bereits viele Verbindungen hat. Aus der sicht der Netzwerkteilnehmer heisst dies, dass ein Peer ein neues Torrent zusammen mit einem beliebten Torrent anbietet.

Die Dichte des BitTorrent Netzwerkes liegt unter 10% und sinkt sogar unter 7%, wenn man nur echte Teilnehmer in Betracht zieht. Diese Veränderung correliert direkt mit der Anzahl der Verbindungen im Netzwerk, welches wiederum mit dem Ausschluss von Teilnehmern zusammenhängt. Auch wenn die Dichte tief ist, ist die mittlere Pfadlänge sehr kurz, was zur Schlussfolgerung führt, dass die Knoten sehr zentral sind und es keine Gruppierungen gibt.

Die Grad-, Zwischen- und Nähezentralität unterstützen die Aussage, dass die Knoten lokal und global zentral sind. Im Netzwerk hat es Knoten, welche extrem gut verbunden sind und somit auf vielen kürzesten Pfade. Diese Knoten haben somit viel Einfluss und Popularität im Netzwerk. Dessen Popularität wird in einem Ranking nach der Anzahl der Teilnehmer, die diesen Torrent anbieten, aufgezeigt.

Das Kommen und Gehen von Teilnehmern im BitTorrent Netzwerk ist normal. Verlässt ein Peer das Netz, dann sind die von ihm angebotenen Torrents nicht mehr verfügbar. War der Teilnehmer zudem der einzige, der dieses Torrent angeboten hat, dann verschwindet das Torrent aus dem Netzwerk durch das Auflösen der Verbindung. Auf diese Weise schrumpft und wächst das BitTorrent Netzwerk ständig. Die Momentaufnahmen des Netzwerkes sind jeweils stündlich verfügbar und die SNA Kennzahlen, welche für die Beantwortung von Klassifikation, Kompaktheit und Zentralität gebraucht wurden, wurden für Daten berechnet, die zwei Tage abdecken. Jede Kennzahl correliert entweder mit der Veränderung der Anzahl Torrents oder Verbindungen, wegen der Art der Berechnung. Beispielsweise verläuft die Kurve des durchschnittlichen Grads gleich mit der Kurve der Dichte und der Anzahl der Verbindungen, weil sowohl der Grad, als auch die Dichte mit der Anzahl der Verbindungen berechnet werden.

Acknowledgments

I want to thank Professor Stiller for giving me the opportunity to write my thesis at the Communication Systems Group and Dr. Thomas Bocek for the supervision during the entire process.

I also want to acknowledge the great support of Andri Lareida as my supervisor. Thank you for the advice, feedback and each and every meeting in regard of this thesis. All of it was very valuable and avoided a lot of stress and unnecessary worries. Also, thank you for such an interesting topic, which made it easy and fun to work on.

Further I want to thank Annatina Vinzens, who provided a lot of helpful insight regarding the English grammar and the structure and content of my thesis. Both Annatina Vinzens and Michael Weiss deserve a thank you for the guidance and answering of so many of my questions concerning scientific writing and the process of writing a thesis in general.

Contents

Abstract	i
Acknowledgments	iii
1 Introduction	1
1.1 Motivation	1
1.2 Description of Work	2
1.3 Thesis Outline	2
2 Background and Related Work	3
2.1 Peer-to-Peer and its Applications	3
2.1.1 Classification of Peer-to-Peer Networks	3
2.1.2 BitTorrent as an Application of Peer-to-Peer Networks	4
2.2 Data Acquisition for the BitTorrent Network	5
2.2.1 Macroscopic Techniques	5
2.2.2 Microscopic Techniques	5
2.2.3 VIOLA as a Complementary Approach	6
2.3 Social Network Analysis	6
2.3.1 Assessing Centrality	7
2.3.2 Assessing Compactness	7
2.3.3 Assessing Clustering	8
2.4 Network Models	9
2.4.1 Random Network	9
2.4.2 Small-World Network	9
2.4.3 Scale-Free Network	9
2.5 Related Work	10
3 Research Questions and Implementation	11
3.1 Questions and their Relevance	11
3.2 Chosen Methods	12
3.3 Implementation of Analysis	13
3.3.1 Data Transformation	14

3.3.2	Using R for Analysis	15
4	Results and Evaluation	19
4.1	Analyzing the Full and Reduced Network	19
4.2	RQ 1: Classification	20
4.2.1	Average Path Length	20
4.2.2	Clustering Coefficient	21
4.2.3	Degree Distribution	22
4.2.4	Conclusion	24
4.3	RQ 2: Compactness	24
4.3.1	Size of the Network	25
4.3.2	Density	25
4.3.3	Conclusion	26
4.4	RQ 3: Popularity of Torrents	26
4.4.1	Degree	26
4.4.2	Betweenness	27
4.4.3	Closeness	27
4.4.4	Popular Torrents	27
4.4.5	Conclusion	29
4.5	RQ 4: Evolution	29
4.5.1	Size of the Network	29
4.5.2	Density	31
4.5.3	Centrality	31
4.5.4	Average Path Length	34
4.5.5	Clustering Coefficient	34
4.5.6	Conclusion	36
5	Summary and Conclusions	37
5.1	Future Work	38
	Bibliography	39
	List of Figures	40
	List of Tables	43
	List of Listings	45
A	Contents of the CD	47

Chapter 1

Introduction

This thesis presents an attempt to make a first assesment of the structure of the BitTorrent Network. It will not only show a classifiction to a scale free network but will also present the reader with further structural insights. This will be achieved with the use of Social Network Analysis, simple graph theory and statistical methods.

1.1 Motivation

Peer-to-peer (P2P) file sharing applications like BitTorrent account for a large amount of the total internet traffic, although video streaming is taking up an even bigger portion [5]. Internet Service Providers (ISPs) face the challenge of network congestion during peak traffic hours, which is made more difficult by the P2P traffic. Furthermore, overlay networks used by file sharing applications such as BitTorrent [7] cause unnecessary inter-domain traffic, because they are typically not aware of the underlying physical network. Several solutions to reduce inter-domain traffic have been proposed by the research community [1, 4, 16, 18, 11]. The evaluation of those solutions is based on limited data which is not necessarily representative.

Social Network Analysis (SNA) traditionally investigates individuals and their social relations. Individuals and relations are modelled as graphs consisting of nodes (individuals) and edges (relations, e.g. friendship). Modelling a social network as a graph enables the calculation and interpretation of several graph metrics, such as centrality metrics, symmetry measures, and entropy measures [5]. Libraries for the calculation of these measures are available for certain platforms [11] and can, therefore, easily be applied to an accordingly prepared data set. However, the interpretation of these measures strongly depends on the mapping chosen to model the graph.

SNA methods are not limited to social networks [2], in fact, these can be applied to any network that can be modelled by nodes and edges. The Orgnet definition of SNA: "Social network analysis [SNA] is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities." [4] hints to the applicability of SNA to areas beyond social networks. This

thesis will apply SNA methodologies to data collected by the VIOLA BitTorrent measurement system to make an assessment of the structure of the BitTorrent network. This type of analysis has not been attempted before, because VIOLA is the first that tracks almost the entire network.

1.2 Description of Work

This thesis applies statistical and SNA methods to the data collected by the VIOLA BitTorrent measurement system. The work starts with literature research on SNA and its applications, thus, in-depth knowledge is gained about BitTorrent and SNA. From the VIOLA data, a graph is created in which torrents (or files) are represented as nodes connected by peers which represent edges. Additionally questions about the structure, clustering, popular torrents, and classification of the network are answered. These questions are discussed in greater detail in Chapter 3.

First, a set of methods to answer the questions is defined and then applied to the VIOLA data set. Second, the outcomes, which are the answers to the questions, will be compared in detail. Finally, all the findings are documented and discussed in a detailed, written, and visualized set of results.

1.3 Thesis Outline

Starting with a description of the theoretical background and related work in Chapter 2, the reader will receive a brief introduction to BitTorrent as a P2P network, the VIOLA measurement system as a way of data acquisition for the BitTorrent network, and Social Network Analysis for assessing the structure of a network in general.

Chapter 3 states the research questions together with their relevance to the thesis and why they were chosen to be answered in this setting. Further the chapter defines methods which are used to answer the research questions and describes the implementation procedure. Data transformation before the analysis and its automation is discussed and further details are given to the usage of R in calculating SNA measures.

The actual results are then shown and discussed in Chapter 4. In this chapter, answers to the research questions that have been stated in Chapter 3 and reasons for these answers are stated. Chapter 5 summarizes the content of the thesis and suggests future work.

Chapter 2

Background and Related Work

The BitTorrent network can be seen as a graph, where torrents are nodes and the edges are constructed when torrents are shared by the same peer in the network.

This chapter introduces the theoretical background and related work. Basics to P2P networks and BitTorrent as an application are given in Section 2.1. Further, the section explains how such networks are classified. Next, the different ways of collecting information about the BitTorrent network are shown. In this context, the VIOLA BitTorrent measurement system is explained as well. In addition, this chapter also contains information about SNA in Section 2.3 and how its measures are calculated and assessed. In the last section of this chapter, information about related work is given.

2.1 Peer-to-Peer and its Applications

P2P networks are networks, where participants connect directly with each other, with no centralizing entity. However, hybrid systems exist. A classification is simple but necessary to understand why the BitTorrent Network falls in this hybrid category.

First, a definition of P2P is given and it is clarified how networks of such manner are classified. There are many applications to P2P networks, however, since BitTorrent is the subject of the analysis, this specific P2P network is explained.

2.1.1 Classification of Peer-to-Peer Networks

There are three main characteristics that mark a P2P network, namely the client/server functionality of each peer, the direct communication among peers and the autonomy of each peer [21]. The ability of a peer to act both as a server and client distinguishes it from a client/server network, where a participant is either a server or client but never both at the same time [22]. Kellerer proposes a definition of P2P networks in [12]. When participants share their hardware resources in the network with other participants, then

such a network is called a P2P network. The services and contents that are offered by the network need those shared resources. Peers access those directly without any intermediary entities. Further, an entity can be removed without the network failing [22].

On the contrary, a hybrid P2P network has some sort of central unit [22]. A network is classified as a hybrid P2P network, if it firstly is a P2P network but also a central participant is necessary for parts of the coordination. The central unit does not take part in the content sharing.

2.1.2 BitTorrent as an Application of Peer-to-Peer Networks

P2P networks have various application like instant messaging, file-sharing, grid computing or other forms of collaboration. The most popular file-sharing network is BitTorrent.

The BitTorrent file distribution system is a hybrid network, which is stated and justified in [7, 6, 13]. It uses trackers that keep records of the files available from each peer. To access a file over BitTorrent, one first downloads a tracker file of the desired file from a BitTorrent portal. In this file, the peer will find IP-Adresses of other peers from which the torrent can be download. The need of these trackers and therefore a central passive coordination, makes the network a hybrid P2P network.

A network of peers that is formed around a torrent file is called a swarm [13]. The tracker maintains a list of peers participating in a specific swarm. Sometimes a tracker maintains more than one list or swarms at the same time. On the other hand, a BitTorrent portal makes torrent files available, which in return have IP addresses of trackers, so that the client can then access a swarm. A client does however not have a list of all the peers participating in a swarm but only between 40 and 200 peers, which form its neighborhood. The actual logistics of down- and uploading are only handled by the peers that are connected to each other [6, 13]. The tracker is not involved in this process and its only purpose is for the peers to find each other. The tracker does not have any more information than that. Even though the tracker is the sole way of peers finding each other, the tracker runs an algorithm that returns a random list of peers. The reason behind creating a random network of peers is its robustness.

To make this system work, two types of peers are needed: seeders and leechers [6, 13]. Seeders are peers that already have a complete version of a file and are uploading it to other peers at the same time. Leechers do not have the requested file or only parts of it and therefore are seeking to download the missing parts from other peers. While leechers are downloading the missing parts, they are at the same time uploading parts of the file that they already have. Once a leecher has the complete file, it turns into a seeder.

A peer sends a list to all its neighbors with the list of available files it is ready to upload for exchange. If a peer has completed downloading a file or parts of it, it informs all its neighbors about the completion, so that at any time, the neighbors have a complete list of available file parts from all its neighbors.

Is a peer waiting to download a file, it is in the state of being choked. Consequently, a peer can only download from a peer when it is unchoked. At any time, a peer only allows four peers to download at the same time. This means that there are always maximum of four peers unchoked for a specific file and all others are choked. The selection of

which peer should be unchoked is based on a tit-for-tat algorithm. It should therefore also prevent freeriding, since it prefers peers that have a large upload-rate for unchoking.

2.2 Data Acquisition for the BitTorrent Network

Kryczka and co. provide a classification of data acquisition methods for BitTorrent Networks [13]. The two main categories are macroscopic and microscopic acquisitions depending on the detail of the retrieved information. Macroscopic data retrieves information about the demographics and the high level performance, whereas in microscopic approach focuses on peer level performance.

2.2.1 Macroscopic Techniques

Macroscopic techniques for data retrieval are used to understand the demographics of the BitTorrent network. Demographic information includes type of content published, popularity of this content, distribution of BitTorrent users per country, and relevance of different portals and trackers [13, 14]. This techniques also allow to study some performance aspects like the ration of seeders/leechers, session time of BitTorrent users, arrival rate of peers, seedless state duration (period the torrent is without seeder) and others. To further divide and classify macroscopic techniques, a distinction between BitTorrent portal crawling and BitTorrent tracker crawling can be made [13].

Portal crawling enables finding detailed information about millions of torrents, which are indexed and structured in the portal that is to be crawled [13]. Once information about different torrents has been retrieved, one needs to crawl them in a systematic way to retrieve more information than only the number of seeders and leechers and its size. These techniques provide measures like content popularity distribution, distribution of the number of published content per category and torrents publishing rate per date. Information which is retrieved when crawling a tracker is of a different kind. While looking for more detailed demographic information, the crawling of a portal does not suffice. The IP addresses of the peers participating in a swarm need to be recorded to get arrival and departure rates. However, IP addresses can only be obtained from a tracker. There are several ways of doing so: tracker logs (requires trackers collaboration), publicly available information (only minor trackers do this) or by using measuring techniques.

All the described methods retrieve exclusively the peers' IP addresses. This only gives insight on the presence or absence of a peer and not the upload or download rates. For such details, microscopic techniques are needed [13, 14].

2.2.2 Microscopic Techniques

In the contrary to macroscopic techniques, microscopic techniques give highly detailed insides into smaller parts of the BitTorrent network like a swarm or group [13, 14]. Dif-

ferent parts of the BitTorrent peer wire protocol need to be implemented. Any crawler has to implement and therefore perform the handshake procedure for connecting to other peers. Passive handshake means, that the crawler waits until a peer starts a handshake and an active handshake on the other hand means, that the crawler connects actively to a peer and therefore starts a handshake itself. After the handshake, when the crawler is connected to the peer, it sends different messages to find out the specified parameters.

2.2.3 VIOLA as a Complementary Approach

In the two presented techniques, information about the network is collected without knowledge of the content of the torrent that is being shared. The information that is gathered is therefore mainly about the interconnectivity of peers. A third way for data acquisition is a complementary approach. VIOLA BitTorrent measurement systems (VIOLA) is a such a technique and is described below [14].

VIOLA measures the BitTorrent system in a distributed manner [14]. It measures the dimensions of content, time and peers and is not concerned with the connection. It can therefore examine a large number of swarms over a longer time period.

The measurement system is designed as a master-slave system. First, the master crawls a BitTorrent portal and then goes on to examine the discovered torrents. Hence, the master is responsible for discovering new torrents and communicating them to the slaves. The slaves in return have to check in with the master, querying trackers for the torrents which they have received from the master and then sending the results back to the master.

2.3 Social Network Analysis

In [2] and also in [23] all basics of Social Network Analysis (SNA) are explained and used as reference for the theoretical background. SNA provides ways to quantify characteristics of complex networks. Measures can be divided into two main groups, depending on the point of view. If we are looking at individual nodes, we talk about node level measures. On the other hand, if the network as a whole is to be captured then we refer to network level measures.

However, all measures have a graph as their reference [2, 23]. One can distinguish between a directed and an undirected graph. The term directed and undirected is associated with the edge type. We speak of a directed edge (u, v) if the graph has a connection from vertex u to v . An undirected edge (w, z) is a connection from nodes w to z and the other direction as well. It means that when edge (w, z) exists then so does (z, w) . SNA measures take into account only undirected graphs.

Another distinction of graphs are the weights on the edges [2, 23]. A graph can either treat unweighted or weighted edges. An unweighted graph would have the weight of 1 for all edges and weighted graph would have different weights for different edges. For this theses unweighted and undirected graphs are considered and are from now on assumed.

2.3.1 Assessing Centrality

The assessment of centrality attempts to answer questions of who is most popular in the network or who is at the center of it. Centrality is concerned with the relative centrality of the individual points in the graph and is therefore also called point centrality as it is illustrated in [23]. Local and global point centrality also differs. The first shows how central the point is in its neighborhood and how many connections it has. The latter addresses the influence in the entire graph.

Node level measures, which are used to measure point centrality in different ways, are listed below as defined by Scott in [23]:

- **Degree** is the number of direct connections to other nodes. This measure is used for the assessment of local centrality. However, this measure is closely related to the size of the graph and therefore comparison between graphs is not recommended.
- **Closeness** is based on the inverse of the distance of each node to every other node in the network. If this measure is high for a node, the node is considered to be close to every other node in the network. The measure gives the idea of global centrality. The results often yield different most central points than local centrality, that was assessed with the measurement of degree. Low numbers indicate high levels of closeness.
- **Betweenness** counts the number of times this node lies on a shortest path. It is a measure for communication flow. A node with high betweenness has high influence in the network. This measure is built around the concept of local dependency instead of centrality. The nodes with high betweenness measures are depended on other nodes, because they lie on different and high numbers of shortest paths.

2.3.2 Assessing Compactness

When assessing compactness, the general linkage of a graph is of interest. A complete graph is one, in which all nodes are linked with all other nodes. This would result in the density of 1 or 100%. Therefore, the assessment of compactness through density is the calculation of how far a graph is from completion and how distributed the lines are in the network [23]. The integration of the graph adds to this assessment by calculating centralization.

To assess the compactness of a network, two main measures are needed as they are described in [2].

- **Density** is defined as the number of lines present to the maximum number of possible lines in the graph. This measure however is not sufficient for the assessment

of the density in general. Other measures are needed to be set into context because density is relative to the size of the graph. Comparison between graphs is not recommended unless the measure is normalized before comparison.

- **Centralization** calculates the difference between the centrality scores of the most central nodes and the scores of all other points. It is therefore the ratio of the actual sum of differences to the maximum possible sum of differences.

Finding boundaries between central, marginal and peripheral nodes is a matter of the researchers judgment like stated in [23]. However, to identify central and peripheral nodes is an important analytical method to describe the structure of a given network. The search for cliques or clusters should be helpful to identify the boundaries to the structural center. Not all networks will show one clear cluster in the center. It is also possible that a graph has different points that are locally central, but do not have a direct connection to each other, which leads to multiple centers.

Another possibility when talking about center and periphery is, that there might be an absolute center in a graph. This would be a focal point around which the graph is structured. The criterion of a set of points for a structural center is not met when considering an absolute center which consists of only one point. The absolute center is the node that is closest to all the other nodes in the graph in terms of distance [23].

2.3.3 Assessing Clustering

Various theoretical models of subgroups have emerged, which are clarified in [23]. Groups are seen as individuals that have mutual connections and therefore show a high density. Cliques, clusters, and components are the different models used, however they have little in common. The corresponding model is chosen depending on the aspect for identification of the group.

Components identify groups that are not or only weakly connected to each other. Cliques are used for groups that are complete. Clusters are judged on similarity of the nodes, which can be relative.

The clustering coefficient gives an idea about the potential of cliques being formed. The term is misleading when considering the definitions above. It measures the degree to which the nodes in a network can contain cliques. In real world networks, nodes tend to form cliques which are then connected to other cliques by a few links. The global clustering coefficient is calculated by counting all triangles in a network and dividing this count by the number of triplets. A triangle is therefore a closed triplet and is counted three times, when calculating the amount of triplets.

2.4 Network Models

A network structure can be classified based on a theoretical model. There are three established concepts for network structure as they are shown in [9, 17]. Random, small-world and scale-free networks differ in various aspects and are characterized based on SNA Measures. Each of them implements different real world phenomenons like skewed degree distribution, clustering or the small-world effect.

2.4.1 Random Network

Random networks have been introduced by Erdős and Renyi in 1959 [9, 17]. Such a random network is created by an empty graph with a predefined number of nodes. A fixed number of edges is then placed randomly between these nodes, thus creating a random network. Due to the fact that the number of nodes and edges is predefined, random networks of any size and density can be constructed. Because of the random decisions, the degree distribution is a Poisson distribution and therefore the majority of the nodes have a degree close to the mean.

2.4.2 Small-World Network

The small-world model was introduced by Watts and Strogatz [17]. The idea is the implementation of the occurrence of the small-world effect, where each node is reachable in six steps on average. To accomplish this form of closeness, clusters are present which are distanced from each other but connected by few links. The links that connect the clusters are therefore long distance and form a short cut between the clusters.

To visualize, one could take a lattice, where the nodes are positioned in a circle and each node is connected only to its neighbors. Getting around in a lattice takes much more steps than the small-world effect would allow. With the presence of short cuts in the lattice, the path shortens down, depending on the density of the long distance connections. In the real world, this would represent a person that has spent some time abroad, for example has done an exchange year overseas and has formed long distance friendships.

2.4.3 Scale-Free Network

A scale-free network is based on the idea of preferential attachment as explained in [9, 17]. As the network evolves and nodes join in, there is a certain probability to which edges will be formed. Nodes that are already important in the network and have many edges of their own, have a high probability of gaining new edges with incoming nodes than nodes with less influence in the network. This way popular nodes become even more popular as the network grows.

This phenomenon can be also seen in the degree distribution, where the degree is plotted to its probability of occurring in the network. Because of the preferential attachment, the distribution is not a poisson distribution. It differs from it due to a higher peak and a long tail. It is the typical power law distribution, because it shows that there are many nodes with a few links and a few nodes with many links.

2.5 Related Work

So far, the analysis of the BitTorrent Network has been focused on performance of downloads, prevention of free riding and other topics not related to the structure of the network, but to performance, scalability, and efficiency [19, 15]. The structure of the network has not yet been attempted to be analyzed, because no other measurement system like VIOLA has been developed for the BitTorrent network. Other measurement methods do not focus on collecting metadata to show the entire network and therefore a structural analysis has not been possible [14].

However, the Gnutella network, which is also a P2P file sharing network, has been analyzed regarding topology in [20]. Similar to VIOLA, a crawler was used by the authors to gain information about the Gnutella network which was then analyzed in terms of degree distribution and clustering. Their findings conclude that Gnutella does not present a pure power law distribution but benefits from it nonetheless. These findings provided the base for the analysis of the internet traffic and its efficiency and scalability. The authors therefore proposed improvements to the protocol that is used by Gnutella based on the research conducted.

The choice of using SNA for this matter is based on many other applications, in which SNA has been used in similar cases, just like in the case of the mapping of Gnutella [20]. SNA has also been used to analyze networks, like the innovators network that was analyzed by Cantner and co. in [3]. The focus of the paper was put on the evolution of the network and its findings included the insight that new innovators entered the network close to the core, following the idea of preferential attachment [3]. In [24], Sternitzke and co use SNA as a tool to visualize the patent citation network and to gain insight on centrality of popular inventors. They distinguish between weak and strong citations, which depend on the extensiveness of the citation and origin either from strong rivalry or from colleagues inside of the same company [24].

Daly and Haahr, authors of [8] designed a routing for MANETS, where SNA is used to determine the best route. MANETs are mobile ad-hoc networks, which are self-configuring and without an infrastructure. If a node is looking for a destination but does not know the route, it forwards the message to a node that is structurally more central and therefore has a higher probability of knowing the destination. For the comparison of betweenness was used. The results of the experiment showed that the SimBet Routing was as successful as Epidemic routing without the overhead of sending the message to all nodes known to one.

Chapter 3

Research Questions and Implementation

The goal of this thesis is the successful answering of the research questions, which will be described and discussed in this chapter. The statement and relevance of the research questions are presented in Section 3.1. Furthermore, each question will require different methods of analysis which will be shown in 3.2. In Section 3.3 the analysis procedures will be explained and it will explain how the methods will be used.

3.1 Questions and their Relevance

Properties like clustering and average path length have to be compared to a random, small world and scale free network. Such a comparison enables the ability to characterize and classify the network. Beside these simple measures, another option is to plot the degrees on the x-axis and the cumulative frequency on the y-axis. Such a diagram is called the degree distribution and enables for the comparison with the characteristics of a random, small world and scale free network. The resulting research question is stated as follows.

RQ 1 *How can the BitTorrent network be classified using a comparison to different network models?*

The second research question is about the density of the network. It is relevant for its characterization. A very dense network is hard to analyze, because the calculation of other measures will increase in time and they also lose their relevance. The betweenness of nodes in an extremely dense network is high for every node and therefore does not lead to new conclusions. Best case for this thesis is the presence of a compactness less than 25%, otherwise the interpretation of all other measures is limited.

RQ 2 *How compact is the BitTorrent network?*

Centrality metrics can help identify the important and/or popular torrents in a network. The further interpretation of what an important or popular torrent in such a case means is another part of the research for the third question. Furthermore, the popularity can be assessed using the BitTorrent portal and then compared to the important torrents identified, using centrality measures such as degree, betweenness and closeness.

RQ 3 *Which are the important torrents in contrary to popular torrents?*

The BitTorrent network, however, is not static, but changes with time as peers come and go. It evolves, shrinks, and grows. This depends not only on the time of the day but also on the geolocation of the peers. VIOLA is not able to capture the entire network within seconds. It takes approximately 20 minutes to register the network. This leads to a clocking of the network each hour once. This fact can not be ignored and an analysis of the network over time is necessary. The result of the fourth question is diagrams showing the development of all measures mentioned above over time with a possible comparison to the size of the network.

RQ 4 *How do Social Network Analysis measures change over time as the network evolves?*

The calculation of measures is not an answer to any of the research questions posed. An interpretation and explanation of the context is necessary. SNA provides a base for such an interpretation and will therefore function as such.

3.2 Chosen Methods

In this section, all previously stated research questions are being discussed in terms of how they are going to be answered and which methods will be applied.

The chosen measures have to be put into a context that makes sense with the BitTorrent network. Since the network is not primarily a social network but rather an anonymous file exchange protocol, an interpretation has to consider the lack of social interaction between nodes. However, there is still a small social component, which is however optimized by the algorithm that is used to select a peer for exchange. The decision is not done based on experience or human knowledge and therefore, merely the scientific aspects of the measures have to be taken into account.

Research Question 1, characterization of the network, can be answered with basic SNA measures and compared to a random network to receive a classification. For this, the average path length, the clustering coefficient and the degree distribution are compared to those of a random network.

How to classify the network based on this comparison is well explained in [2, 23]. If BitTorrent's average path length is smaller or equal to the average path length of the random graph, the network can either be characterized as a small-world network or as a scale-free network. If the clustering coefficient is much bigger than the one of the random network, consequently the network is a small-world network. There is no accurate comparison of

the clustering coefficient of a random graph and a scale-free network. At last, if the degree distribution is similar, then we have a small-world network. However, if the degree distribution is a power law distribution, then the network is a scale-free network.

Hence, the answer to Research Questions 1 is a statement of the BitTorrent network being a random, small world or scale-free network. An interpretation of the average path length and the clustering coefficient is also required.

When addressing the question of compactness, which is Research Question 2, the most important measure is density. It is calculated by taking the actual amount of vertices divided by the number of potential vertices. This way, the ratio of how many of the possible connections exist in the particular graph is calculated. The formula for this calculation looks as follows, when assuming that N is the number of nodes and E is the number of vertices: $\frac{E}{(N*(N-1))/2}$ [2, 23]. How this ration affects the network is also part of the answer to Reasearch Question 2.

To be able to assess the importance and popularity of particular nodes, as it is stated in Research Question 3, it is neccessary to collect the 50 nodes with the highest degree. The amount of how many nodes to collect was chosen arbitrarily and does not have any further meaning. These nodes will then be compared to the 100 most popular or most owned torrents. The goal is to see if the popular torrents according to the download rate are also important in the network. Since the network evolves, the measures of degree, betweenness, and closeness will be calculated at the same time of the network evolution.

The time evaluation of the measures is done by applying the methods mentioned to separate datasets which have been gathered at different times. This enables the author to create a graph showing measures over time and also their correlation. An additional interpretation of the evolution is also needed to fully answer Research Question 4.

3.3 Implementation of Analysis

The analysis of a big data set can not be tackedled at once and with no prior knowledge of the methods, tools and its limits. Firstly, all measures will be calculated using a reduced data set for testing purposes before the code will then be applied to the complete datasets. Now, to be able to apply the functions to a full set, time is required. The calculation will take several minutes to hours. Due to this time component, a script will be developed that enables to calculate all functions on a server.

Such a procedure can be applied to the data transformation and to the analysis as well. This way, all the needed steps can be executed by the server, without having to monitor the steps or the server. It also enables the user to gather measures for Research Question 4 and then in a manual step, draw the diagrams which show their evolution. The interpretation is done as last, once all measures have been calculated and visualized.

In Section 3.3.1 it is described, how a hadoop MapReduce was used to transform the data into the format, that was then later analyzed using R. The automation of the data

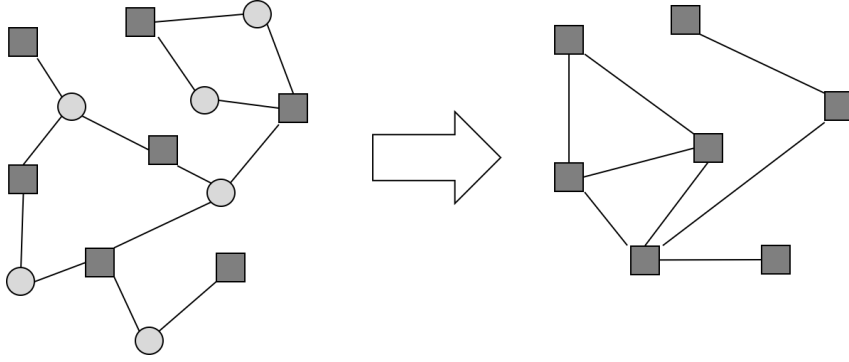


Figure 3.1: Transformation from a two mode to a one mode network.

transformation is also described. In Section 3.3.2, the usage of R in the process of calculating measures as well as the used commands and their automation for all data sets are illustrated.

3.3.1 Data Transformation

The data model which is used by VIOLA has to be restructured so that a network connecting torrents can be created. For the MapReduce a flat avro file is used. Such a file stores information of an *ANNOUNCE*. Those are results that were recorded by the slaves during the measurements of VIOLA. It consists of the torrent in question, the slave id and a list of peers, represented by IP addresses, that the tracker has returned for the torrent.

The goal is to transform the given data into the form of torrent to torrent connection with the weight of how many peers are forming this edge. This procedure is shown in Figure 3.1, where circles represent peers and squares torrents. A two mode network is one, where there are two different kinds of nodes. For the SNA analysis a one mode network is needed which means, that only one kind of nodes are present in the network, in this case torrents. The data transformation is performed with the help of a MapReduce in three steps. The Hadoop framework was used because of its simple to use implementation and an already functioning cluster at the University of Zurich.

The first MapReduce job is the transformation of the data that was retrieved by VIOLA into a form where the infohash of the torrent is associated with one and each IP address separately. An infohash is basically the identification of a torrent. For that, the *ANNOUNCE* is taken, as well as the infohash and the list of peers from each *ANNOUNCE* entry. In a first step, pairs are formed from each entry with the infohash and each IP address for the peers recored in that entry. For the pairs, a new dataclass is needed which stores a text value for the infohash and a long value for the IP address. They are then used in the mapper for the removal of the duplicates. The mapper stores the values in the .avro format, so that the system can deal with the big amount of data.

The second MapReduce job reads the unique infohash-IP-pairs and constructs infohash-to-infohash edges for the network. An edge is formed when two unique infohash IP pairs

have the same IP address. Then, the two different infohashes are taken to form the edge. The case of the infohashes being the same will not occur, because that would mean that the unique infohash IP pairs had duplicates. However, those duplicates have already been eliminated in the previous MapReduce job. There still will be duplicates formed during this second MapReduce, but will be dealt with in the third MapReduce job.

The last MapReduce job takes all the edges that were previously produced and reduces them to weighted edges. Duplicates created in the previous step are used to calculate the weights of the edges. Therefore, it counts how many times an edge has been built by unique IP addresses. This last step then yields the finished edge list of the network that is to be the subject of the analysis.

Hadoop Script

Applying the transformation to many data sets by hand was not an option. That would require manually setting the parameters of about 24 sets over 4 days and in between triggering of the jobs and therefore waiting for them to be completed. An automation had to be made, so that the calculation could run over night.

For the ability to trigger a terminal command after a previous one is completed, a simple bash script is used. It is set up as a double for-loop, first going through the available days and then hours for each day. The terminal command is then called with the appropriate parameters. The bash script was launched on a hadoop cluster. The script can be seen in Listing 3.1.

```

1 INPUT="/mnt/hdfs/user/lareida/database/partitioned/2016/3"
2 DIR="/mnt/hdfs/user/pernischova/"
3 OUTPUT="mapreduce_16_03_"
4 #!/bin/bash
5 DAYS='ls -l $INPUT | egrep '^d'| awk '{print $9}''
6 for DAY in DAYS
7 do
8   HOURS='ls -l $INPUT/$DAY | egrep '^d'| awk '{print $9}''
9   for HOUR in HOURS
10  do
11    #execute hadoop request
12    hadoop jar MapReduce.jar org/pernischova/ViolaWeightedEdgeList
13    /user/lareida/database/partitioned/2016/3/$DAY/$HOUR
14    $OUTPUT${DAY}-${HOUR}
15  done
16 done

```

Listing 3.1: Hadoop script used to execute all MapReduce jobs.

Once all MapReduce jobs are finished, the results are copied on a different server, on which they are further processed using R, as it is explained in Section 3.3.2.

3.3.2 Using R for Analysis

The programming language R provides a system for statistical computation and graphics. It is primarily a command line tool, but graphic open source interfaces are available. Since

Number of vertices	<code>length(V(graph))</code>
Number of edges	<code>length(E(graph))</code>
Density of graph	<code>density(graph)</code>
Mean of degree of vertices	<code>mean(degree(graph))</code>
Mean of betweenness of vertices	<code>mean(betweenness(graph))</code>
Mean of closeness of vertices	<code>mean(closness(graph))</code>
Clustering coefficient	<code>transitivity(graph)</code>
Average path length	<code>average.path.length(graph)</code>
Number of clusters	<code>clusters(graph)\$no</code>
Mean size of clusters	<code>mean(cluster(graph)\$csize)</code>

Table 3.1: R commands used for calculations.

the community around R is well developed and big, there are many libraries, which provide the user with well defined and also efficient functions [10]. The library *igraph* includes all function needed to calculate SNA measures using graphs.

First, the data has to be imported into the R environment and cast into a dataframe. Out of the constructed dataframe, which consist of edge information, the graph is generated using the command `graph <- graph.data.frame(relation, directed = FALSE)`. On the *graph* object, futher methods can be applied. Table 3.1 shows the functions that have been used.

R Script

Each R command has to be executed for every data set and the results have to be documented in a way that they can easily be used for interpretation. R is capable of running scripts. One such script has been written to execute all functions after each other and then write the output into a text file. The script is shown in Listing 3.2.

```

1 #loop through all available folders in data
2 library("igraph",
3       lib.loc="/home/pernischova/R/x86_64-pc-linux-gnu-library/3.2")
4 sink("/home/pernischova/data/excluded_over200/results_13_3.txt")
5 setwd("/home/pernischova")
6 for(day in 13:14)
7 {
8   for(hour in 0:23)
9   {
10    setwd("/home/pernischova")
11    string <- paste("/home/pernischova/data/mapreduce_16_03_", day, "_",
12    hour, "/weightededgelist", sep="")
13    setwd(string)
14    # clear relation for new Data
15    relation <- NULL
16    # write data into structure
17
18    # loop -> do this for every file in dir!
19    file.names <- dir(pattern = "part-r-")
20    for(i in 1:length(file.names)){
21      file <- read.table(file.names[i], header=FALSE, sep="\t")
22      relation <- rbind(relation, file)
23    }
24  }
25 }

```

```

22
23 output <- NULL
24 output <- rbind(day, hour)
25 cat("\n")
26 cat(output, sep="/")
27 cat("\n")
28 # make a graph object
29 graph <- graph.data.frame(relation, directed = FALSE)
30 # get measurements
31 output <- NULL
32
33 # Number of Vertices, Number of Edges:
34 output <- rbind(length(V(graph)), length(E(graph)))
35
36 # Density of graph:
37 output <- rbind(output, graph.density(graph))
38
39 # Mean Degree:
40 output <- rbind(output, mean(degree(graph)))
41
42 # Mean Betweenness:
43 output <- rbind(output, mean(betweenness(graph)))
44
45 # Mean Closeness:
46 output <- rbind(output, mean(closeness(graph)))
47
48 # Centralization Score based on Degree:
49 output <- rbind(output, centralization.degree(graph)$centralization)
50
51 # Clustering Coefficient:
52 output <- rbind(output, transitivity(graph))
53
54 # Average Path Length:
55 output <- rbind(output, average.path.length(graph))
56
57 cluster <- clusters(graph)
58 output <- rbind(output, cluster$no)
59 cat(output, sep=", ")
60 cat("\n")
61 cat(cluster$size, sep=", ")
62 cat("\n")
63 }
64 }

```

Listing 3.2: R Script used, to calculate SNA measures.

Chapter 4

Results and Evaluation

The goal of this chapter is to answer the research questions defined in Chapter 3. Since each question can be answered with the help of particular measures, the sections are labeled accordingly. Section 4.2 answers the first research questions about the classification of the network. The second research question is answered in Section 4.3, where the compactness of the BitTorrent network is analyzed.

4.1 Analyzing the Full and Reduced Network

The first assessment of a test network revealed that the density is over 91%. The test data was used for testing purposes. Because of this initial finding, the network had to be reduced, so that a more accurate picture could be portrayed. IP addresses that hold a vast number of torrents do not represent real participants in the network and are therefore excluded. This way, the network is reduced.

To be able to find these IP addresses, a query is used on the database holding the data. The result of this query is a list of all IP addresses that had more than a specific amount of torrents, see Figure 4.1. The figure shows IP addresses on the x-axis with their number of shared torrents on the y-axis. Only the first 100 torrents are displayed, so that the big fall in the amount can be easily visible. This list is then implemented as a filter in the MapReduce procedure, reducing the amount of edges that are then built. To be able to see the difference between the full and the reduced network, the SNA measures are calculated for all the data sets.

Since there is no true threshold on how many torrents a peer normally offers, the limits 1000, 500, 200 and 100 torrents are used. The SNA measures were calculated for all these networks and the diagrams are shown and discussed in the corresponding sections below. During the comparing, it turned out that the results for the exclusions of over 500, 200 and 100 yield the same results. This leads to the conclusion that the significant peers, which owned an unreal amount of torrents, were excluded in the first two steps. Further exclusions did not change the network enough to see a difference in the measures.

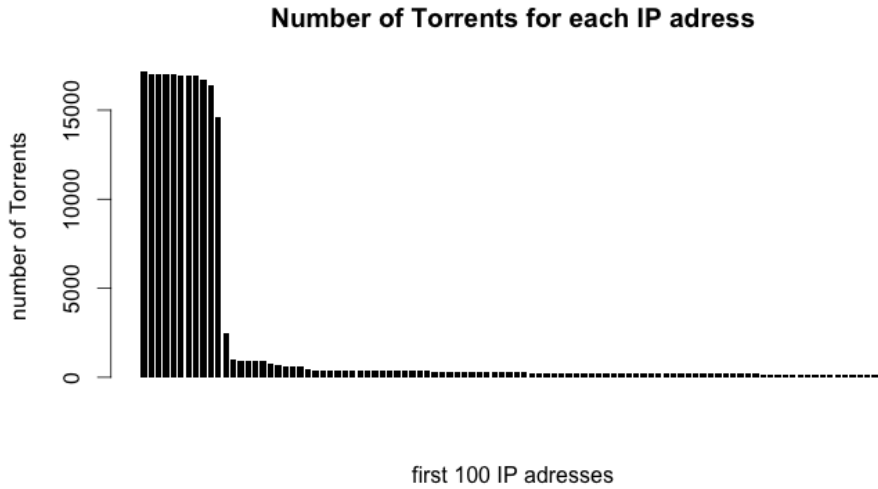


Figure 4.1: First 100 IP addresses and their number of torrents.

Therefore, only the results of the full networks and the networks of exclusions over 1000 and 500 are presented.

The number of torrents in the graph did not change significantly. When talking about 16'000 nodes, the difference of 60 is very small. The nodes that are excluded are peers that track the network, just like the VIOLA measurement system does. The importance of this exclusion is to be seen in different measures like the number of edges, density, path length or mean degree.

4.2 RQ 1: Classification

The goal of Research Question 1 is to classify the BitTorrent network. The following subsection compare the measures that are necessary for such a classification to a random network.

The random network was created using the model of Erdos and Renyi with R . To be able to compare measures, the size of the network has to be equal. Therefore, the number of nodes and edges are identical in the comparison.

4.2.1 Average Path Length

The average path length refers to the average of all shortest paths in a network. This measure shows how connected the graph is. The result is shown in Table 4.1 showing the reduced networks with a comparison to a random network.

This number is low for a network of this size. Having a path length of two means that through two connections, each node can be reached on average. This points to the graph

Network	BitTorrent	Random
Full network	1.95	1.91
Excluded over 1000	1.96	1.92
Excluded over 500	1.98	1.93

Table 4.1: The average path length for all networks in comparison to a random network of the same size.

Network	BitTorrent	Random
Full network	0.429	0.091
Excluded over 1000	0.361	0.079
Excluded over 500	0.323	0.066

Table 4.2: The clustering coefficient for all networks in comparison to a random network of the same size.

having a few nodes with have a very high degree and those nodes being the linking point for most of the paths. Because of these well connected nodes, there is a big number of paths which only have a lenght of one, and therefore pushing the average path length a great deal. When looking at the degree distribution in Section 4.2.3 this is very clear. It also shows that there are many nodes with a small degree and therefore probably being connected to a node with a high degree.

4.2.2 Clustering Coefficient

As it was explained in Section 2.3.3, the clustering coefficient gives an idea about the potential of clusters being formed in a network.

The edges that have been excluded through the process of reducing the network were the ones being responsible for a great number of triangles. Therefore, the clustering coefficient is significantly higher in the full than in the reduced network as shown in Table 4.2.

However, the absence of clusters was not expected. Even when looking at the reduced network, no clusters were found that could be reported. This underlines the finding of a very short average path. For the path being so short, there has to be only one cluster. The connecting between clusters would require at least a path length of two added to the average path length, which is present in one cluster. This however would only be possible in a structre with only two clusters. When there are more clusters in a network, the average path length has to be higher as it is suggested by the small-world phenomenon. In a small-world network the average path length is 6 [17].

Clustering can however also be assessed using data mining algorithims. They are typical for a small-world network, but don't exclude a classification as a scale-free network. Surprisingly, there are no clusters persent in the network. This supports the assumption of a scale-free classification.

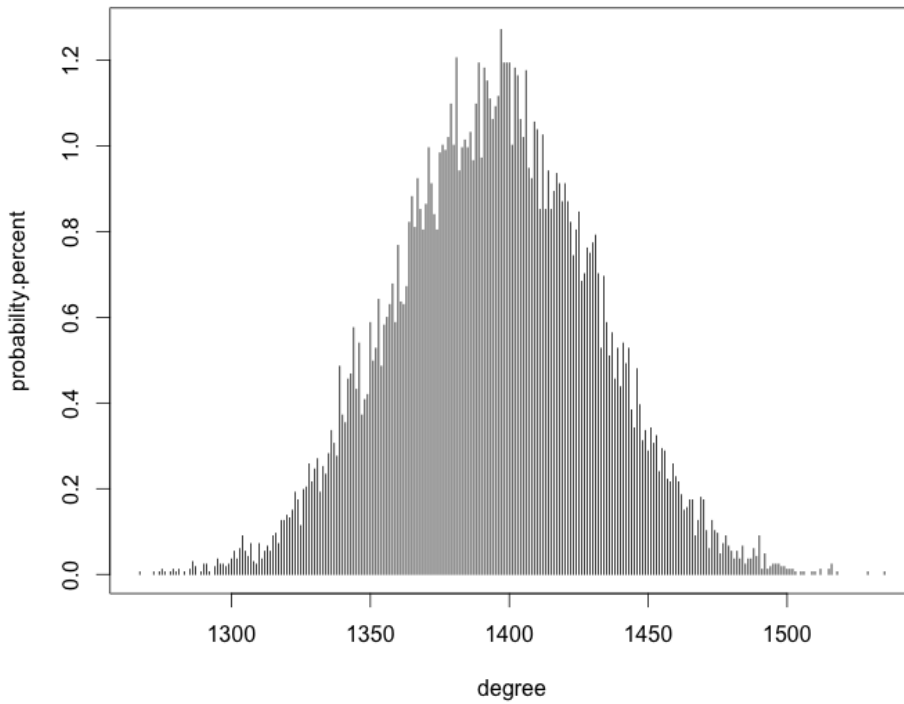


Figure 4.2: Degree Distribution of a random network

4.2.3 Degree Distribution

The degree of a node in a graph is defined by the edges that are connected by this node. Since this thesis is not concerned with the direction of the edges, the incoming and outgoing edges are the same and therefore there is one measure for each node. The distribution of the degrees is a way of showing the probability of a node having that degree.

The degree distribution of a network is crucial for its classification. The comparison to the distribution of a random network of the same size is the way to analyze the BitTorrent network.

In a random network, the degree of the nodes form a normal distribution. An exemplary distribution is shown in Figure 4.2. For this example, information was used corresponding to the size of the BitTorrent network, so that an appropriate comparison can be made. Now comparing Figure 4.2 to Figures 4.3a, 4.4a, and 4.5a, which show the degree distribution of the full graph, excluded over 1000 and over 500 one can see, that they differ. The degree distribution of the BitNetwork is a power law distribution, as it can clearly be seen in Figures 4.3b, 4.4b, and 4.5b. In these figures the distribution was logarithmized to show the descent, which is typical for a power law distribution.

The distribution of degrees in a network are pivotal and a power law distribution signifies a scale free network [9].

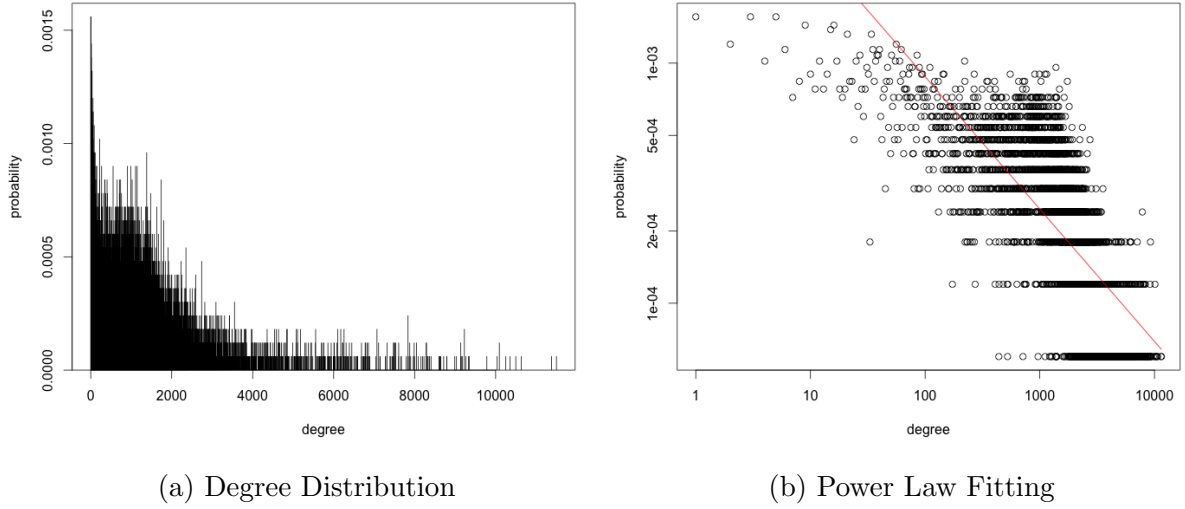


Figure 4.3: Degree distribution and power law fitting for the full BitTorrent network.

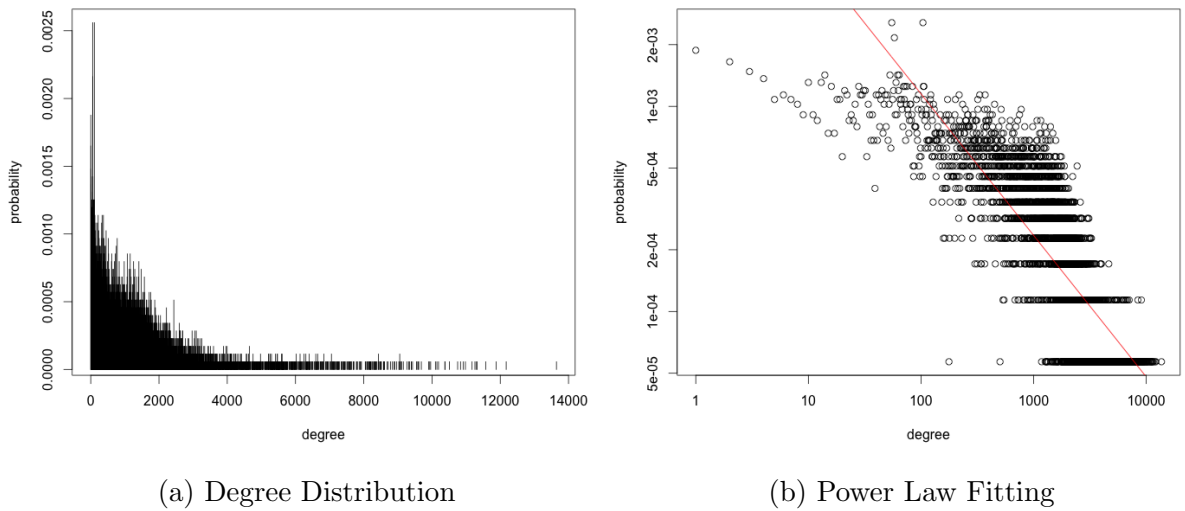


Figure 4.4: Degree distribution and power law fitting for the BitTorrent network excluding IPs with over 1000 torrents.

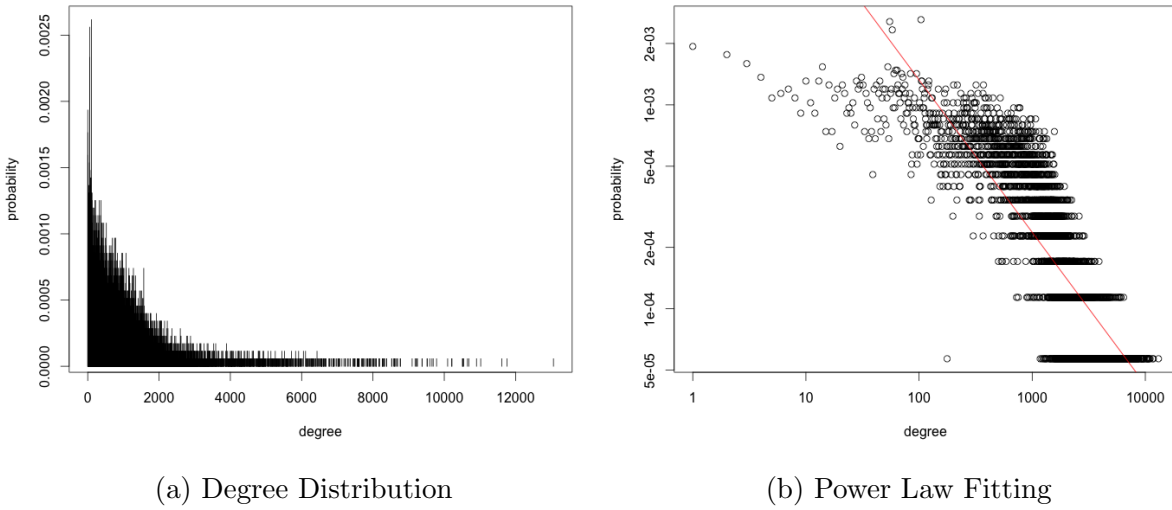


Figure 4.5: Degree distribution and power law fitting for the BitTorrent network excluding IPs with over 500 torrents.

4.2.4 Conclusion

The clustering coefficient of the BitTorrent Network being larger than the one of the random network points to the BitTorrent network being a small world or scale free network. Since the average degree is almost identical, no decision can be based on that. The most important comparison is to be made using the degree distribution.

The degree distribution of the BitTorrent network shows a typical distribution of a scale-free network, the power law distribution. The Poisson distribution which can be seen in the degree distribution of the random network is a pointer, that the BitTorrent is not random or small-world but scale-free.

With the BitTorrent network being a scale-free network, it would mean, that preferential attachment is present. When a new torrent is being added to the network, the probability is big that the peer that offers this new torrent also offers or seeks another that already is offered by many others.

4.3 RQ 2: Compactness

Compactness is analyzed using mainly the density measure. However, density does not say much about the network without knowing the size of it. This means, that to be able to answer the Research Questions 2, a look at the size and density has to be taken.

Number of	Nodes	Edges
Full network	16'940	13'018'406
Excluded over 1000	16'940	11'263'700
Excluded over 500	16'935	9'473'873

Table 4.3: The number of nodes and edges for the BitTorrent network.

4.3.1 Size of the Network

The size of the BitTorrent network is shown in numbers in Table 4.3. The number of nodes does not change with the exclusion of peers, which have many torrents, because those are peers that do not add new torrents to the network. The excluded peers are not real participants in the network. They try to monitor the network like the VIOLA measurement system or try to attract peers to be able to assess down- and uploading rates. Since they are not real participants, they do not offer torrents that would be disregarded through this exclusion.

The number of edges, also shown in Table 4.3, however, changes significantly. The excluded peers create a vast number of edges. When one of these excluded peers offers for example 1000 torrents, then there are $1000 * (1000 - 1)/2 = 499'500$ edges formed. Exactly this change in size was anticipated when excluding such peers. The results excluding peers that offered more than 200 or 100 torrents, are not included, because of the exclusion not reducing the size any further. Therefore, the threshold of which peers can still be considered as real participants has to lie between 1000 and 500. However, the finding of the exact threshold is not subject of this thesis.

Through the exclusion, a more accurate structure is shown since excluded peers do not participate in the file exchange.

4.3.2 Density

The number of torrents and edges gives an idea of how many nodes there are in the graph. Density adds more information in form of a ratio, since density is calculated from the number of edges and nodes. It is the percentage, that shows how many edges there are out of all possible ones. This ratio is shown in Table 4.4. Considering the network having almost 17'000 nodes, this is not a high density. Once IP addresses are excluded that have too many torrents and are therefore not regular peers, the density drops to almost 6%. The excluded IPs constructed almost one third of the network's density.

To paint a better picture, one has to compare this percentage to a full graph. A full graph is one, where each node is connected to every other node and therefore has a density of 100%. So on average, each node is only connected to 1'185 out of almost 17'000 nodes, which accounts for only 7%.

	Density
Full network	9.1 %
Excluded over 1000	7.9 %
Excluded over 500	6.6 %

Table 4.4: The density of the full network and the two reduced networks.

4.3.3 Conclusion

The change in size of edges can be very well explained by the exclusion of peers that do not really participate in the network but only want to monitor it. It also justifies the consistency of the number of nodes in the full and reduced networks. Considering the size of the network, the density of 7% is extremely small and makes further calculations easier. The size of the network, when looking at the number of nodes and edges would not suggest such a low density.

The BitTorrent network shows a compactness of 7%, when looking at the connections between torrents. This means, that there must be many more peers present in the network than torrents. These peers must offer some of the same torrents and therefore not increasing the density.

4.4 RQ 3: Popularity of Torrents

The centrality of the network can be assessed in different ways. However, to receive a good overview of the network and the connections between the nodes, more than one way of assessment has to be consulted [23]. Therefore, degree, closeness, and betweenness has been chosen. The degree was already discussed in Section 4.2.3 together with the degree distribution, even though it talks about centrality as well. As an addition, the popularity of torrents is taken into account and is compared to the measures.

4.4.1 Degree

As already discussed in Section 4.2.3, the degree distribution is a power law distribution. This means, that there are many nodes with a small degree but also some nodes which are very well connected [9]. These nodes drive the mean degree high, as it can be observed on the difference between the mean and the median of the degree in Table 4.5.

As an example, the degree of 882 means that the particular node is connected to 882 other nodes. In the case of the BitTorrent network, it implies that the torrent is offered together with other 882 torrents. However, one needs to consider, that this does not stand for only one peer, that offers all 883 torrents, but for many peers that offer one torrent together with another torrent and therefore forming an edge. A high degree thus signifies high popularity and demand among peers. The degree of 882 or 712 shows that the torrent is connect to $882/16'940 * 100 = 5\%$ of all torrents. With the density being 6%, this is a high degree.

Degree	Mean	Median
Full network	1'404	882
Excluded over 1000	1'203	882
Excluded over 500	1'012	712

Table 4.5: The mean and median degree of the nodes of the BitTorrent network from 14.03., hour 11.

Betweenness	Mean	Median
Full network	8'008	1'710
Excluded over 1000	8'129	1'710
Excluded over 500	8'283	1'637

Table 4.6: The mean and median betweenness of nodes of the BitTorrent network from 14.03., hour 11.

4.4.2 Betweenness

Betweenness of nodes is consulted for the local dependency instead of centrality. If the betweenness is high, it indicates a high influence in the network [23]. Table 4.6 shows the mean and median betweenness in the network.

Betweenness as a number says that the node lies on that amount of shortest paths [23]. The number is especially high, when the particular node lies on many shortest paths that connect one specific node with any other node. In this network, such a development can be further explained when looking at path length in Section 4.2.1.

The difference between the mean and the median of betweenness is to be explained the same way as with degree. The distribution is not a normal distribution but rather skewed, like the power law distribution. Therefore, the median is located closer to the smaller betweenness and the mean is driven up by the nodes which have extremely high betweenness. These nodes are the ones that also have a high degree and thus are passed many times by the shortest paths.

4.4.3 Closeness

The measure of closeness tells something about the global centrality of a node. For this measure, a low number indicates a high level of closeness like it is in the case of the BitTorrent network. This number is considered small in the context saying that the nodes in the network are close to each other [23]. The reason for closeness can then be explained by the small path length and small clustering coefficient that is discussed in Section 4.2.

4.4.4 Popular Torrents

The popularity of a torrent is defined by the number of peers that offer it in the BitTorrent network. The following query returns the infohash of a torrent and the number of how

Closeness	Mean	Median
Full network	0.0000071	0.0000071
Excluded over 1000	0.0000071	0.0000071
Excluded over 500	0.0000071	0.0000071

Table 4.7: The mean and median closeness of nodes of the BitTorrent network from 14.03. hour 11.

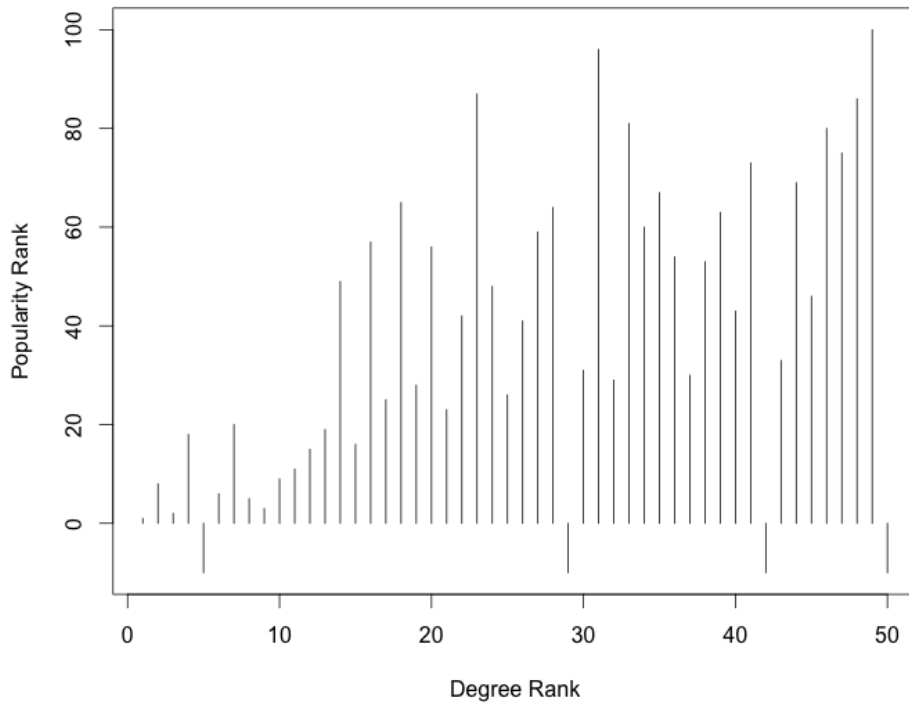


Figure 4.6: Torrents with highest degree, mapped with their popularity rank.

many peers are providing this particular torrent. The first 100 torrents with the highest number of peers are recorded and used for comparison.

```

1 SELECT a.infohash, count(distinct a.ip) as popularity
2 FROM announce a
3 WHERE a.pyear = 2016
4 AND a.pmonth = 03
5 AND a.pday = 14
6 AND a.phour = 11
7 GROUP BY a.infohash
8 ORDER BY popularity DESC;
```

Listing 4.1: Impala query to extract the most popular torrents from the database

Taking a look at the most popular torrents and the torrents with the highest degree, it is not surprising to see, that the lists contain the same torrents. All but three torrents of the top 50 list with the highest degree are to be found in the list of top 100 popular torrents.

Figure 4.6 shows the 50 torrents with the highest degree together with their popularity rank, if such is below 100. As already mentioned, a high degree can be directly correlated with the popularity of the torrent in the network.

4.4.5 Conclusion

The degrees shows a typical power law distribution. Therefore, there are many nodes in the network with a low degree but also some nodes with a very high degree. This correlates with the high betweenness of the nodes. The nodes with a high betweenness have high influence in the network, because many shortest paths happen to pass through them. The small closeness states a high global centrality of the nodes.

The peers that offer popular torrents, at the same time offer other torrents in the network. This leads to the popular torrents also having a high degree and betweenness. For the peers, such torrents presents a way into the network. When a peer can offer a popular torrent to many other peers, it is not seen as a free-rider, when the peer wants to download torrents from other peers.

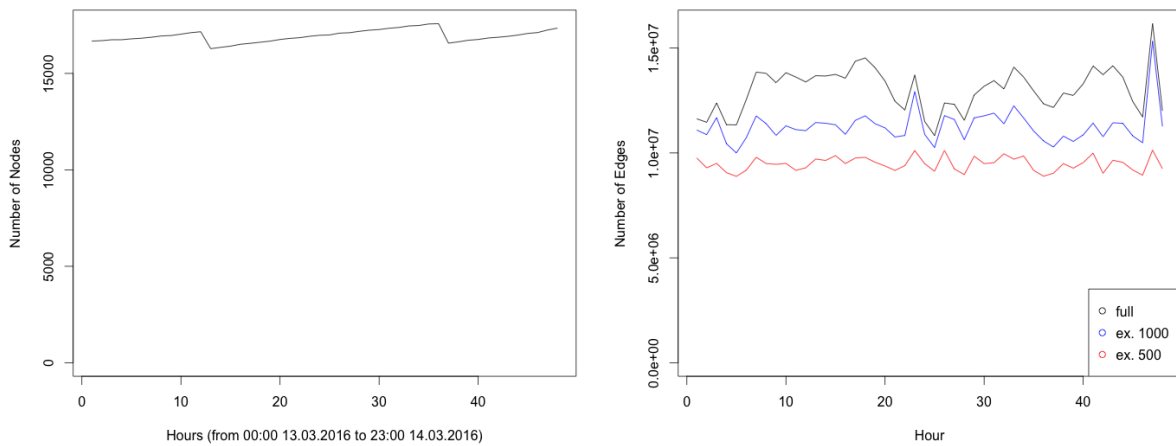
4.5 RQ 4: Evolution

The last research question investigates the development of the network over time. So far, the measures have been explained in the context of the BitTorrent network but since the network is not static, the evolution of it is taken into account as well. The following sections show the evolution of the measures and explain the in- and decrease of them.

All diagrams have the same x-axis which is labeled with hours. The hour 0 stands for midnight on the day 13.03.2016. Each hour then is measured and plotted until day 14.03.2016 hour 23.

4.5.1 Size of the Network

The size of a network can be expressed in different ways. First, let us take a look at the number of torrents, as it is shown in Figure 4.7a. There is a drop in the number of torrents at midday of roughly 1'000 torrents both times. This drop is to be accredited to VIOLA which drops torrents that do not have less than a certain amount of peers. This elimination happens each day before noon. So when measuring for more than two days, this drop would reoccur each day. The torrents, that are excluded between the hour 11 and 12 on day 14 are shown in Figure 4.8. The x-axis shows the popularity of the torrent, this means, how many peers offer the torrent. The y-axis shows the number of torrents that have this popularity and are then not found in the list of torrents for hour 12. The query that is shown in Listing 4.2 is used to aquire the information from the database.



(a) The number of torrents in the network

(b) The number of edges in the network

Figure 4.7: Size of the BitTorrent network over the course of two days.

```

1 SELECT c.pop, count(c.infohash)
2 FROM (
3     SELECT a.infohash, COUNT(DISTINCT a.ip) AS pop
4     FROM announce a
5     WHERE a.pyear = 2016
6     AND a.pmonth = 03
7     AND a.pday = 14
8     AND a.phour = 11
9     AND NOT EXISTS (
10        SELECT b.infohash
11        FROM announce b
12        WHERE b.pyear = 2016
13        AND b.pmonth = 03
14        AND b.pday = 14
15        AND b.phour = 12
16        AND a.infohash = b.infohash)
17     GROUP BY a.infohash) AS c
18 GROUP BY c.pop

```

Listing 4.2: Query to get popularity distribution of torrents making the difference between hours 11 and 12.

The next way of analyzing the size of a network is the number of edges. Figure 4.7b shows the rise and fall of the number of edges over the course of two days. Changes in the size can be explained simply by peers that come and go. Peers are seldomly connected to the network two day in a row. The coming and going of peers explains the rise and fall of the number of edges. Once there is no peer that offers the two torrents together, the edge is eliminated. At the same time though new peers appear, offering several torrents and therefore, new edges are formed. The curve represents only snapshots of the network, because it is only logged each hour. Comparing the lines of the full network and the network where IPs were excluded, one can see that the number of edges remains more stable. There are still changes in the number, but they remain within a certain boundary.

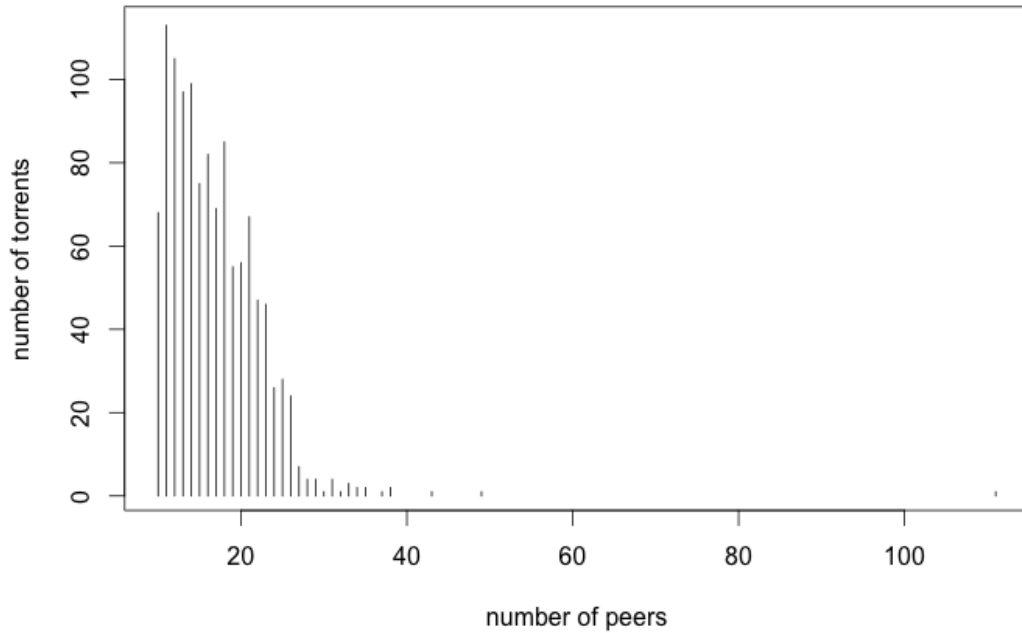


Figure 4.8: Popularity of torrents, that make the difference between hours 11 and 12.

However, when taking a look at the whole network, peaks in either direction are visible. This can be explained by peers parting and coming that have a big amount of torrents, which are then excluded in the reduced network causing the peaks to subside.

4.5.2 Density

Figure 4.9 shows the changing density for the full and reduced networks. The density varies around 10% when the entire network is taken into account. The rise and falls of density are the same as the ones in Figure 4.7b, because the density is directly correlated with the number of edges.

4.5.3 Centrality

Centrality is assessed using different measures, degree, betweenness and closeness. Their development is shown and explained below.

Degree

Just like density, the rise and fall of the mean degree is the same as the developments in the number of edges shown in Figure 4.10. The degree is also directly correlated with the

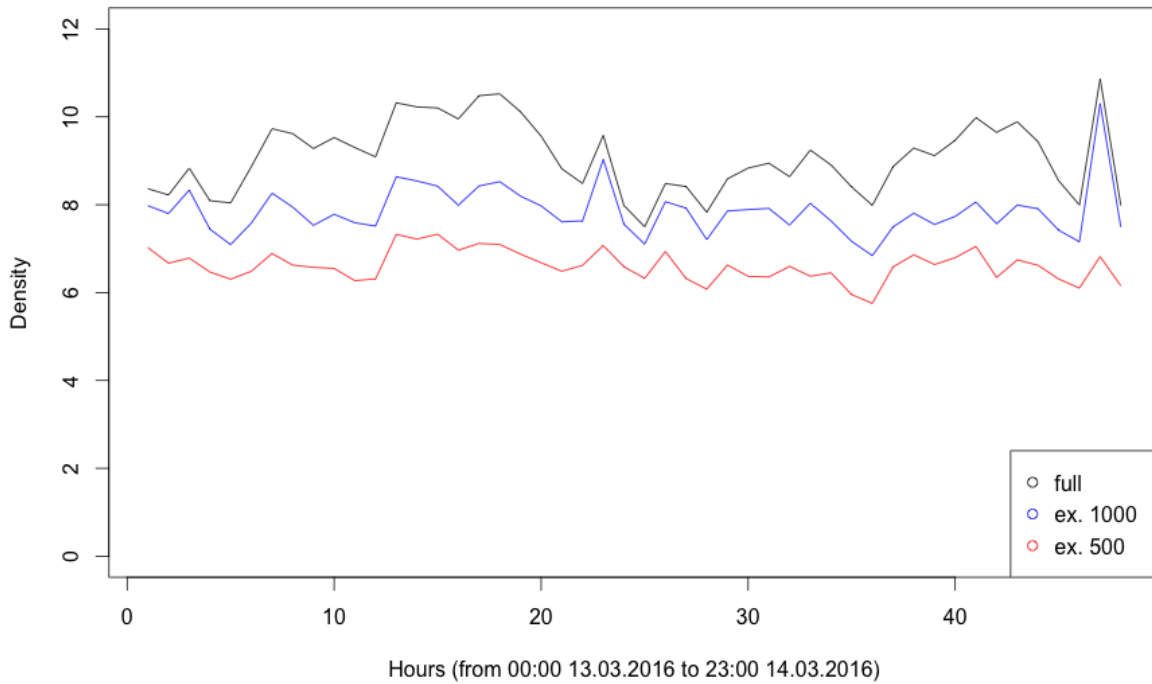


Figure 4.9: The density and its evolution over two days shown for every network.

number of edges, since each edge is counted twice, for each node that it connects. The accumulated degree of the network is therefore twice the number of edges.

Betweenness

The increase in betweenness can easily be explained by the growing number of nodes in the network, like it is visualized in Figure 4.11. As the network grows, the influence of the nodes, which are very well connected increases and therefore, the mean betweenness does as well. The measure is directly correlated with the number of nodes in the network and resembles the rise and fall of the number of nodes in Figure 4.7a.

Closeness

As it can be seen in Figure 4.12, the closeness jumps between approximately 0.00001 and 0.00003. However, the jumps are deceiving, because the number is so small and the scale in the figure is showing just the necessary segment.

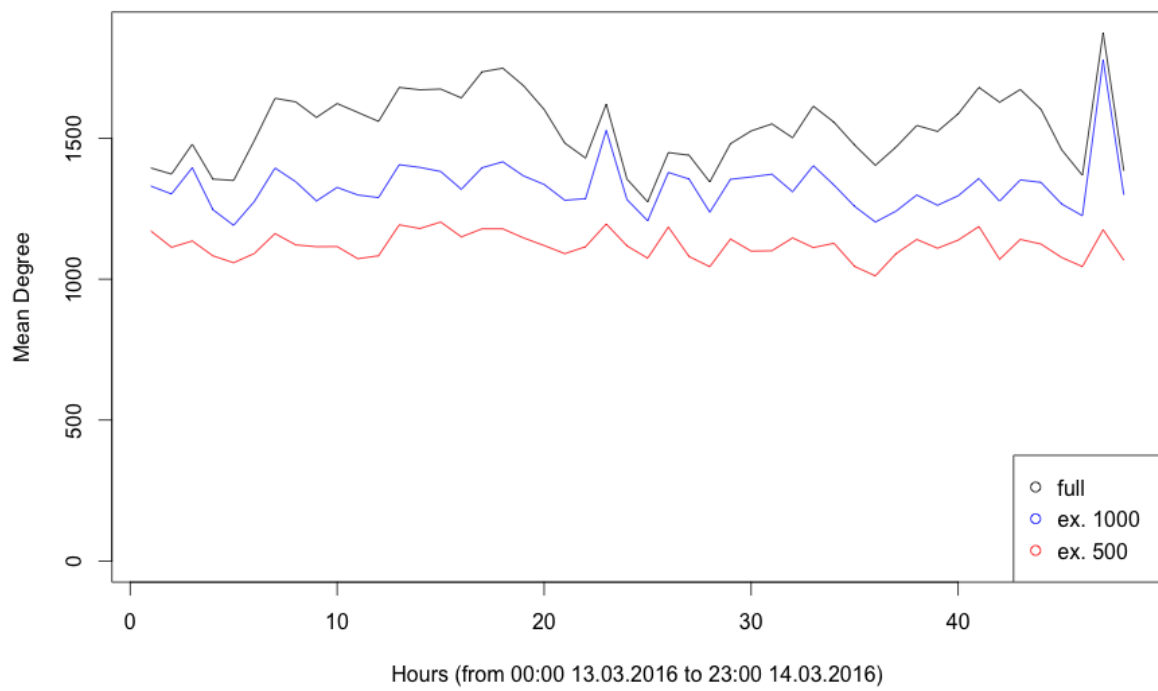


Figure 4.10: Mean degree of nodes in the networks over the course of two days.

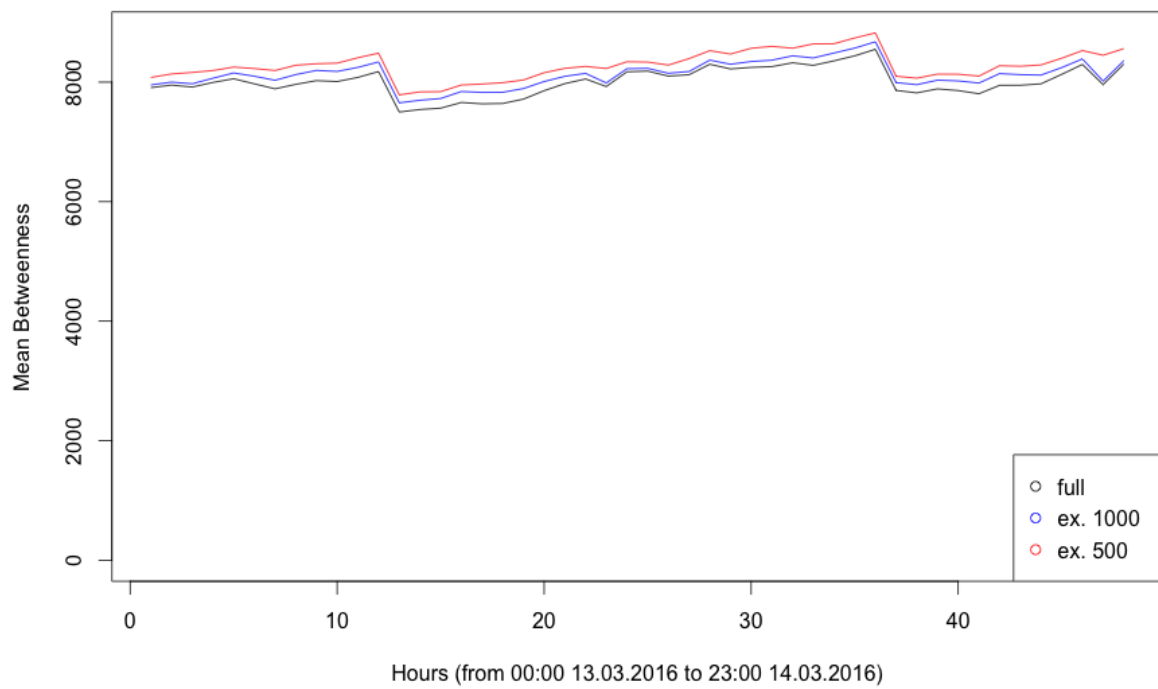


Figure 4.11: Plot showing the mean betweenness of the nodes in the networks.

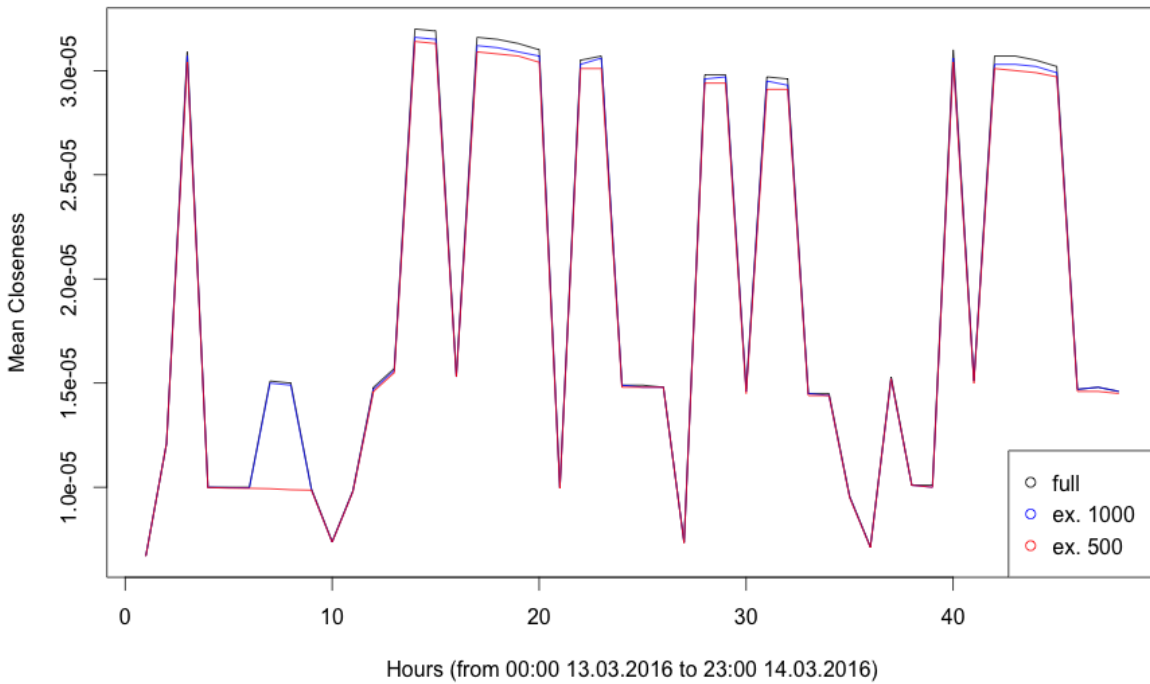


Figure 4.12: Mean closeness for the full and reduced networks.

4.5.4 Average Path Length

Figure 4.13 shows the path length for the full and the reduced networks. The number seems to jump significantly, but taking a closer look at the labeling of the y-axis, which shows a very small section of the axis from 1.92 to 2.00. When looking at the lines from a distance, the in- and decrease of the line can be seen, which equals to the number of torrents in Figure 4.7a. The nodes that are new are not as well connected and cause the average path length to augment. However, this amount is extremely small and not of great significance.

4.5.5 Clustering Coefficient

Figure 4.14, the clustering coefficient is shown for the full and reduced network. Since it is interdependent of the number of edges, it is absolutely not surprising that the incline and down fall in the diagram is extremely similar to the one in the number of edges in Figure 4.7b. It can be explained by peers joining and leaving the network and therefore edges being dropped or added.

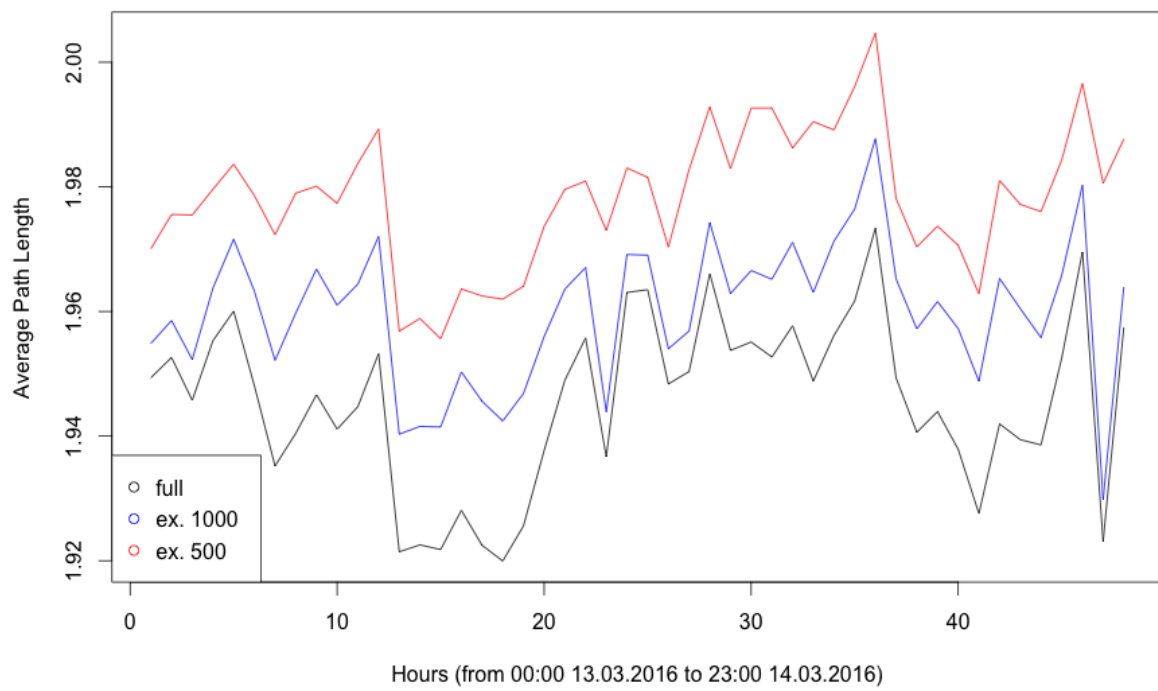


Figure 4.13: Average path length for all networks.

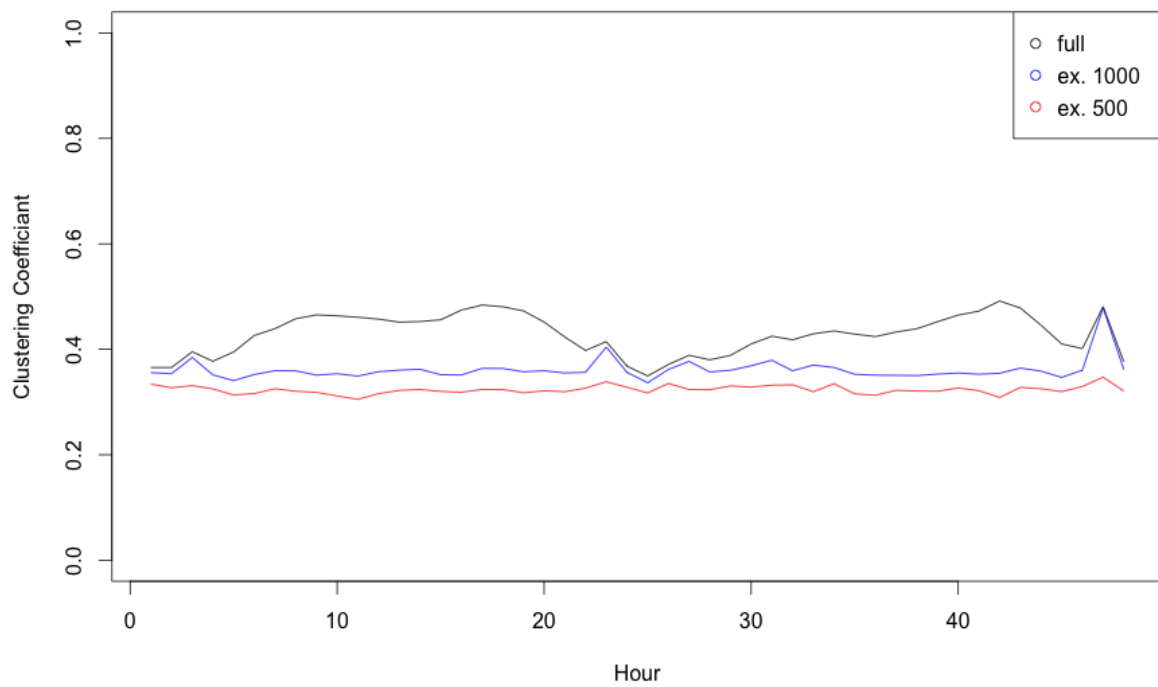


Figure 4.14: Plot with the clustering coefficient of the full and reduced networks.

4.5.6 Conclusion

The development of the measures is either dependent on the number of nodes or the number of edges and therefore, the diagrams look extremely similar when looking at in- and decline of the lines. The explanations of those are correlated with the coming and going of peers and those offering new torrents, that previously weren't offerend in the network.

When peers join or leave the BitTorrent network they offer other torrents and therefore edges are created or torn respectively. All the changes in the the measures are explained by this phenomenon.

Chapter 5

Summary and Conclusions

The VIOLA BitTorrent measurement system provided the data for the analysis of the network. Using the Hadoop MapReduce framework, a transformation was then performed. Through this, the data was transformed from a two mode network, torrents and peers as nodes, to a one mode network, only having torrents as nodes. The edges were formed based on the information of one peer providing two or more torrents. Further, a graph object was created in R, on which numerous measurements were calculated and documented for the discussion. These steps were executed several times, because the network is in constant change and snapshot of it would not be representative.

Due to the vastness of the network a filter was implemented, which reduced the network based on the peers and the torrents they provided. If a peer provides an unrealistic amount of torrents, it was to be excluded and no edges were formed through this particular peer. The threshold of the unrealistic number of torrents lies between 1000 and 500 torrents. This is due to the fact that calculations made for networks which excluded peers with more than 200 or 100 torrents yield the same results. Therefore, they offer no further information about the structure of the network.

The BitTorrent network formed from torrents has a very short average path length of almost 2. At the same time it has a relatively high clustering coefficient of 0.3 compared to the one of a random network of the same size, that is 0.08. Because of the power law distribution of the degrees in the network, the BitTorrent network is classified as a scale-free network. This means that in the BitTorrent network preferential attachment is in practice. When a new torrent joins the network it is most likely to connect to a torrent, that already has many other connections. Stated in the terms of peers, the peer which offers a new torrent, that has not been in the network so far, offers it together with a torrent that many other peers offer as well.

The density of the network lies below 10% for the full network and below 7% for the reduced network. Hence, a torrent is connected to 5% of all the torrents on average. This change is directly correlated with the decrease of edges, that are associated with the reduction of peers. Even though the density is very low, the average path length is very short, which leads to the conclusion of the nodes being central and no clusters being present.

The centrality measures of degree, betweenness and closeness supported the statement

of the small average path length. In the BitTorrent Network there are nodes, which are extremely well connected and make therefore lie on numerous shortest paths. This gives the nodes a lot of influence and popularity in the network. That these well connected nodes are popular is shown by their ranking on the amount of peers, that offer these torrents.

The coming and going of peers is normal in the BitTorrent network. When a peer leaves the network, the torrents that he offered are no longer available, if he was the only participant to do so. This way the BitTorrent network grows or shrinks in number of torrents constantly. The snapshots of the network are available for each hour and the SNA measures that were used to answer the questions of classification, compactness and centrality were calculated throughout the course of two days. Each SNA measure showed a curve either similar to the curve of the number of edges or nodes. Each measure is either calculated with the number of nodes or edges. For example, the mean degree shows the exact same curve as the density and number of edges, because both the degree and density are correlated with the number of edges.

5.1 Future Work

The results of the data transformation calls for the calculation of further measures, which were not subject of this thesis. No ego-centric analysis was conducted. Such analysis would throw light on the neighborhood of nodes. Further, other methods like data mining could be used to analyze the neighborhood to gain knowledge about what torrents are available together. Is there some kind of pattern, which is followed when new torrents join the network? Are they made available together with torrents of the same category, language, year of issue or is there some other influence all together?

An more indepth analysis about clusters and cliques would be a further step. The conclusion about no significant clusters was unexpected. Other methods than the igraph library provides could be applied to see, if those bring about the same results. When using data mining, clustering can be examined using specific attributes of nodes. Such properties were not used in this thesis but are available through VIOLA.

For both, the analysis of the neighborhoods and the in-depth analysis of cluster, further characteristic traits of torrents are needed. VIOLA already collects such, since the master crawls a BitTorrent portal to discover torrents. The attributes include the name of the movie, size, length, language and category. Further information about the movies could also be obtained using the international movie database.

Bibliography

- [1] R. Bindal, P. Cao, W. Chan, J. Medved, G. Suwala, T. Bates, and A. Zhang: Improving Traffic Locality in BitTorrent via Biased Neighbor Selection, 26th IEEE International Conference on Distributed Computing Systems (ICDCS 2006), Lisboa, Portugal, July 2006.
- [2] V. Burger, D. Hock, I. Scholtes, T. Hossfeld, D. Garcia, Mi. Seufert: Social Network analysis in the enterprise: Challenges and Opportunities, Socioinformatics-The Sicoal Impact of Interactions between Humans and IT, Springer International Publishing, 2014. 95-120.
- [3] U. Cantner, H. Graf: The network of innovators in Jena: An application of social network analysis, Research Policy 35/4, 2006.
- [4] D.R. Choffnes and F.E. Bustamante: Taming the Torrent: A Practical Approach to Reducing Cross-ISP Traffic in Peer-to-Peer Systems, ACM SIGCOMM Computer Communication Review, Vol. 38, No. 4, pp 363-374, October 2008.
- [5] Cisco Inc.: Cisco Visual Networking Index: Forecast and Methodology, 2012-2017. http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html, last visited April 2016.
- [6] B. Cohen: Incentives build robustness in BitTorrent, Workshop on Economics of Peer-to-Peer systems. Vol. 6. 2003.
- [7] B. Cohen: BitTorrent for Developers, http://www.bittorrent.org/beps/bep_0003.html, last visited in March 2016.
- [8] E.M. Daly, M. Haahr: Social Network Analysis for Routing in Disconnected Delay-tolerant MANETs, Proceedings of the 8th ACM International Symposium on Mobile Ad Hoc Networking and Computing, New York, USA, 2007.
- [9] O. Hein, M. Schwind, W. K"onig: Scale-Free Networks, The Impact of Fat Tailed Degree Distribution on Diffusion and Communication Processes, Wirtschaftsinformatik 48, 2006.
- [10] K. Hornik: What is R?, https://cran.r-project.org/doc/FAQ/R-FAQ.html#What-is-R_003f, version June 2016, last visited in June 2016.
- [11] The igraph core team: igraph, <http://igraph.org/>, last visited May 2016.

- [12] W. Kellerer: Dienstarchitekturen in der Telekommunikation-Evolution, methoden und vergleich. Technical Report TUM-LKN-TR-9801, 1998.
- [13] M. Kryczka, R. Cuevas, A. Cuevas, C. Guerrero, A. Azcorra: Measuring BitTorrent Ecosystem: Techniques, Tips and Tricks. Articulo de investigacion. Universidad Carlos III e Instituto IMEDIA Networks, 2011.
- [14] A. Lareida, S. Schrepfer, T. Bocek, B. Stiller: Overlay Network Measurements with Distribution Evolution and Geographical Visualization.
- [15] M. Li, J. Yu and J. Wu: Free-Riding on BitTorrent-Like Peer-to-Peer File Sharing Systems: Modeling Analysis and Improvement, IEEE Transactions on Parallel and Distributed Systems, July, 2008.
- [16] A. Mondal, I. Trestian, Z. Qin, and A. Kuzmanovic: P2P as a CDN: A New Service Model for File Sharing, the International Journal of Computer and Telecommunications Networking, Vol. 56, No. 14, pp 3233-3246, September 2012.
- [17] M.E.J. Newman: Structure and function of networks, Computer Physics Communications 147, 2002.
- [18] S. Oechsner, F. Lehrieder, T. Hossfeld, F. Metzger, D. Staehle, and K. Pussep: Pushing the Performance of Biased Neighbor Selection through Biased Unchoking, IEEE International Conference on Peer-to-Peer Computing (P2P 2009), Seattle, USA, April 2009.
- [19] J. Pouwelse, P. Garbacki, D. Epema, and H. Sips: The Bittorrent P2P File-Sharing System: Measurements and Analysis, Peer-to-Peer Systems IV: 4th International Workshop, IPTPS 2005, Ithaca, NY, USA, February 24-25, 2005.
- [20] M. Ripeanu, I. Foster, A. Iamnitchi: Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design, arXiv preprint cs/0209028, 2002.
- [21] D. Schoder, K. Fischbach: Peer-to-Peer, Wirtschaftsinformatik 44, 2002.
- [22] R. Schollmeier: A Definition of Peer-to-Peer Networking for the Classification of Peer-to-Peer Architectures and Applications, Proceedings of the First International Conference on Peer-to-Peer computing, IEEE, 2002.
- [23] J. Scott: Social Network Analysis, 3rd edition, Sage, London, 2013.
- [24] C. Sternitzke, A. Bartkowski, R. Schramm: Visualizing patent statistics by means of social network analysis tools, World Patent Information 30/2, 2008.

List of Figures

3.1	Transformation from a two mode to a one mode network.	14
4.1	First 100 IP adresses and their number of torrents.	20
4.2	Degree Distribution of a random network	22
4.3	Degree distribution and power law fitting for the full BitTorrent network. .	23
4.4	Degree distribution and power law fitting for the BitTorrent network ex- cluding IPs with over 1000 torrents.	23
4.5	Degree distribution and power law fitting for the BitTorrent network ex- cluding IPs with over 500 torrents.	24
4.6	Torrents with highest degree, mapped with their popularity rank.	28
4.7	Size of the BitTorrent network over the course of two days.	30
4.8	Popularity of torrents, that make the difference between hours 11 and 12. .	31
4.9	The density and its evelution over two days shown for every network. . . .	32
4.10	Mean degree of nodes in the networks over the course of two days.	33
4.11	Plot showing the mean betweenness of the nodes in the networks.	33
4.12	Mean closeness for the full and reduced networks.	34
4.13	Average path length for all networks.	35
4.14	Plot with the clustering coefficient of the full and reduced networks.	35

List of Tables

3.1	R commands used for calculations.	16
4.1	The average path length for all networks in comparison to a random network of the same size.	21
4.2	The clustering coefficient for all networks in comparison to a random network of the same size.	21
4.3	The number of nodes and edges for the BitTorrent network.	25
4.4	The density of the full network and the two reduced networks.	26
4.5	The mean and median degree of the nodes of the BitTorrent network from 14.03., hour 11.	27
4.6	The mean and median betweenness of nodes of the BitTorrent network from 14.03., hour 11.	27
4.7	The mean and median closeness of nodes of the BitTorrent network from 14.03. hour 11.	28

List of Listings

3.1	Hadoop script used to execute all MapReduce jobs.	15
3.2	R Script used, to calculate SNA measures.	16
4.1	Impala query to extract the most popular torrents from the database . . .	28
4.2	Query to get popularity distribution of torrents making the difference between hours 11 and 12.	30

Appendix A

Contents of the CD

Zusfsg.txt	German summery of the thesis (approximately 100 words)
Abstract.txt	English summery of the thesis (approximately 100 words)
Bechelorarbeit.pdf	PDF file of the complete thesis
Graphics	Folder with all diagrams and graphics that are used in this thesis
Results	Folder with all results that have been used for the visualizations
MapReduce	Folder with the java code containing the Hadoop MapReduce used for the data transformation
MapReducefull.jar	JAR file of MapReduce calculation of the full graph
MapReduceex1000.jar	JAR file of MapReduce calculation for network with excluded IP addresses with over 1000 torrents
MapReduceex500.jar	JAR file of MapReduce calculation for network with excluded IP addresses with over 500 torrents
MapReduceex200.jar	JAR file of MapReduce calculation for network with excluded IP addresses with over 200 torrents
MapReduceex100.jar	JAR file of MapReduce calculation for network with excluded IP addresses with over 100 torrents
hadoopscript.sh	Bash script for execution of Hadoop commands
r-script.r	R script for execution of measure calculations
random-network.r	R script used for the creation and calculation of measures for the random networks
data-visualization.r	R script containing the results and commands used to create the diagrams