



**University of
Zurich**^{UZH}

Social Network Analysis of the VIOLA dataset

*Robin N. Stohler
Zurich, Switzerland
Student ID: 10-296-275*

Supervisor: Andri Lareida
Date of Submission: September 1, 2016

Abstract

Social network analysis is today a widely used tool in online marketing to present customers products they want because of their customer profiles. Originally coming from sociology social network analysis is becoming a more and more important interdisciplinary instrument for researcher to better understand abstract data. This work will construct for each hour in two days 48 different networks using collected raw BitTorrent usage data. Where Autonomous Systems are the nodes and video files the links between them. Multiple hypothesis and the corresponding research questions are stated and discussed during the thesis using various methods from the social network analysis toolbox.

Zusammenfassung

Die Soziale Netzwerk Analyse ist heutzutage ein weit verbreitetes Werkzeug im Online Marketing um Waren anzuzeigen, welche dem Kundenprofil entsprechen. Ursprünglich aus der Soziologie stammend ist sie heute ein immer wichtiger werdendes interdisziplinäres Instrument um mehr Informationen über die zugrundeliegenden Daten zu erhalten. Diese Arbeit konstruiert für einen Zeitraum von zwei Tagen 48 verschiedene Netzwerke aus gesammelten BitTorrent-Daten in denen Autonome Systeme die Akteure spielen und Videofiles die Verbindungen zwischen diesen bilden. Verschiedene Hypothesen und dazugehörige Forschungsfragen werden aufgestellt und mit einer Reihe von Instrumente der Sozialen Netzwerk Analyse beantwortet.

Acknowledgments

I would like to thank Andri Lareida for his support and advise during the whole time of this bachelor project.

Thanks are also due to Professor Stiller who accepted the same bachelor project twice.

Finally, I wish to thank Bernhard Kneubühler for his assistance in statistical concepts and proofreading the thesis.

Contents

Abstract	i
Zusammenfassung	iii
Acknowledgments	v
1 Introduction	1
1.1 Motivation	1
1.2 Description of Work	1
1.3 Thesis Outline	2
2 Background & Related Work	3
2.1 Background	3
2.1.1 Peer-to-Peer	3
2.1.2 BitTorrent	4
2.1.3 Social Network Analysis	5
2.1.4 Autonomous System	6
2.2 Related Literature	7
2.3 Preliminary Work	9

3	Methodologies	11
3.1	Research questions	11
3.2	BitTorrent-Graph	13
3.2.1	Set up of the Network	13
3.2.2	Construction of the Network	14
3.3	Data Processing	16
3.3.1	Data	16
3.3.2	Construction of the Edge List	16
3.3.3	Filtering the Data	18
3.3.4	Analysis	18
3.3.5	Querying Data with Impala	19
3.4	Methods to Answer the Research Questions	19
3.4.1	Correlations	19
3.4.2	Topology Analysis	22
3.4.3	Network over Time	24
3.4.4	Clusters	26
4	Results	29
4.1	Q1: Correlation	29
4.1.1	Selected Node-centric Metrics	29
4.1.2	DataBank Data by Country	30
4.1.3	Raw Data from Autonomous Systems	31
4.1.4	Comparison of Correlation Regions	33
4.1.5	Comparison over the whole Time	37
4.1.6	Conclusions	37
4.2	Q2: Network Topology Analysis	38
4.2.1	Small World	38
4.2.2	Scale-freeness	39

<i>CONTENTS</i>	ix
4.2.3 Comparison of Correlation among the Different Models	42
4.2.4 Q2: Answer	42
4.3 Q3: Network over Time	43
4.3.1 Metrics over Time	43
4.3.2 Important Nodes	48
4.3.3 Animation	49
4.3.4 Q3: Conclusion	49
4.4 Q4: Effects of Data Filtering	50
4.4.1 Effect on the top nodes	52
4.4.2 Conclusion	53
4.5 Q5: Analysis of Autonomous System Behaviour over Time	53
4.5.1 Switzerland	54
4.5.2 Costa Rica	55
4.5.3 Q5: Conclusion	57
4.6 Q6: Cluster Analysis	58
4.6.1 Cluster Building	58
4.6.2 Assortativity	62
4.6.3 Q6: Conclusion	63
5 Summary and Conclusions	65
5.1 Summary	65
5.2 Conclusion	65
5.3 Final Statement and Future Reasearch	68
Bibliography	69
Glossary	73
List of Figures	74
List of Tables	77

A	Futher Material	81
A.1	Correlation	82
A.1.1	Variables overview	82
A.1.2	Pearson	83
A.1.3	Spearman	83
A.1.4	Kendall	85
A.2	Network over Time	86
A.2.1	Important nodes	86
B	Contents of the CD	87

Chapter 1

Introduction

Social Network Analysis (SNA) constructs a network with entities as nodes connected through some kind of relation as edges. These nets might be further analysed using mathematics and statistics to calculate special measures. The measures rank nodes according to various centralization concepts. The ranking helps to identify important players in such nets like a person having many friends, being close to every one or being in the centre of the information flow these are just some examples. Coming from sociology and psychology these nets might also be used in any other field where nodes are not people but proteins, webpages or anything else.

1.1 Motivation

The goal of this thesis is to apply different known concepts of SNA to gain further knowledge about the data. This is done in an explorative research style using different methodologies searching for insights on the data.

1.2 Description of Work

The data where the constructed network is based on was gathered after successfully building a BitTorrent (BT) data collection system [15]. The stored information of the data contains basically who offers what to download when on the BT infrastructure. The who is indicated by the Internet Protocol (IP) or Autonomous System Number (ASN), the when by a time stamp, the where by the continent, country or city and, finally, the what is indicated by an infohash. An infohash is in this case an id of a file available in the BT infrastructure. While one movie in different qualities and uploaded from different people have also different ids. Using this data a network containing Autonomous System (AS) as nodes and infohashes shared among peers in different AS online as edges. In this work the new generated network will be evaluated to understand what kind of network features it possesses. Network measurements are abstract numbers ordering nodes in the network

but together with real world data such metrics can be used to predict real-world properties or chances, therefore correlation analysis will be done using different types of information like the measurements, data related information and also some real-world information will be used. The constructed network changes overtime according to the usage of BT this change in time will be analysed and also the interaction of each measurement. Further the need of filtering, clustering or usage patterns of user will be assessed.

This thesis will apply statistical, SNA, and other methods to the data collected by the Video Consumption in Overlay Networks (VIOLA) BT measurement system. The work is based on the Vertiefungs Arbeit (VA) conducted by the author. However, further literature research on SNA and its applications is necessary, in order to get further knowledge about BT and SNA. From the VIOLA data a graph will be created in which AS are represented as nodes connected by torrents (or files) which represent edges and the number of common peers determines their weight. Specifically, tools and methods shall be found to store a two-mode graph and convert it to a one-mode graph for analysis. Additionally, the following questions shall be answered in the thesis:

- How can the resulting network be characterized? Compare to random, small world, and scale free networks properties. One option is to plot the degree (x-axis) vs. its cumulative frequency (y-axis).
- Are there clusters of peers? Use SNA cluster measures, e.g. coreness or clustering coefficient, to identify clusters. A possibility would be to resolve the geolocation of the IP addresses contained in the cluster and plot those on a map.
- Which are the important peers? Use node centrality metrics to identify important AS. Compare those with with information from the database. Measurement nodes like those used by VIOLA must be filtered.
- How can SNA metrics be interpreted in this graph? Explaining what the used metrics mean in the context of the modelled BT network.

First a set of methods to answer those questions will be defined, secondly applied. The author is encouraged to extend this list of questions and come up with his own, which are driven by the data set under investigation. These methods defined will be applied to the VIOLA data set and the questions will be answered, while comparing the detailed outcomes of those different methods. Finally, all the findings will be documented in a detailed, written, and visualized set of results within this report.

1.3 Thesis Outline

Chapter two describes important topics, current research literature in fields relevant for this thesis and related preliminary work. In Chapter three general work approaches, the underlying data, the evolved research questions and methodologies needed to answer these questions are discussed. Chapter four will cover the results of the used methodologies and will also give some possible explanations. Chapter five will then summarize the conducted research and conclude sensible findings.

Chapter 2

Background & Related Work

For this thesis important topics and background will be introduced and an overview of the literature in similar fields and its methodologies in Related Literature will be given. At the end some preliminary work critical for this thesis is mentioned.

2.1 Background

Some of the important topics for the thesis will be listed and explained. Such as Peer-to-Peer (P2P) a network paradigm with two equal participants. BT a file sharing protocol using the P2P architecture. SNA a methodology coming from psychology and sociology which uses statistics and mathematics to calculate measurements to rank nodes in a network. In the last part AS are explained.

2.1.1 Peer-to-Peer

P2P is a distributed symmetrical network paradigm where usually no server and client exists, but every peer may act as client or as server. P2P had a big time during the early 2000s when Napster came up, a platform to share general music among members. Members registered the music on their hard disk to a central Napster database. If anyone searches music within the Napster system, the database returns a peer holding the wanted data, providing enough information to be able to download the music directly from the returned peer [24]. The amount of decentralisation may vary on the P2P systems. Napster as an example is not fully decentralized because of the main database in which all the locations of the files are saved and where the peers have to query the data they want, to get the download address. But there are also P2P systems that are fully decentralised so there is no central super-node or server that holds data or is responsible for the routing. Examples of such P2P systems are Gnutella, Freenet, FreeHaven etc.,[28].

2.1.2 BitTorrent

BT [34] is a hybrid P2P system, which contains trackers (servers) that keeping "track" of the available files and the users holding them and other user as peers. A tracker keeps information about the swarm. The peers that share their downloaded files are called seeder and peers that download (leech) leecher while as soon as a leecher has one complete file he also might provide data chunks to other leechers. The files are split up in pieces of 256 KB which are again subdivided in sub-pieces of 16 KB [34].

Upload

To publish content on the BT network a static .torrent file has to be created and put on a web server. This file contains the file's length, name, hashing information and the Uniform Resource Locator (URL) of the corresponding tracker. Trackers are responsible that downloading peers and up-loading peers find each other. The communication between tracker and down-loading peers is a simple protocol layered on top of Hypertext Transfer Protocol (HTTP). The down-loading peer shares information about what port it is listening to and information alike, while the tracker answers with a list of peers that are downloading the same file [6].

Download

To download data using the BT technology a desktop client is required. Examples of such clients are uTorrent[43], BT[34], Vuze (previously Azereus)[44]. The .torrent file has to be downloaded from the internet. These files are usually found on torrent pages such as The Pirate Bay, Mininova or isoHunt etc. The .torrent file points to a tracker of the wished data. The client asks the tracker for peers holding the file. The tracker returns a random number of peers holding pieces or the whole file. Then the peers start to share the pieces among each other. A leecher might start seeding after he received the first piece. If the whole file is downloaded a peer stays seeder as long as he is online and keeps the file in the clients database. Various heuristic rules for piece distribution exist these rules assure the fastest possible distribution of pieces among all the peers. The BT [6] protocol defines the following mechanisms :

Rarest Piece First This rule makes sure that the piece that most peers want, but almost no one has get first shared and that pieces which are more common are left for later. So if there are several leechers and only one seeder the pieces that none of the leechers have are shared first such that if said piece is successfully shared more than only one peer has it and, therefore in case the seeder would go offline chances are higher that other peers also might complete the file.

Random Piece First This rule replaces the rarest piece first when a new peer starts to download a file. Because as soon as the first full piece is received this new peer might contribute to the distribution by sharing his completed piece. This makes sense since a new leecher is not as trustworthy as a peer with more pieces, and therefore the scarce pieces are not wasted on new peers.

Tit-For-Tat To maximize download speed a variant of tit-for-tat is used. A peer tries to upload only to a specific number of peers which hold the highest download rate. Uploading in BT is called unchoking while choking is the temporary refusal of uploading to peer. To evaluate the best connections of the peers BT uses a technique called optimistic unchoking. Every 30 seconds a peer gets randomly selected to get unchoked and at the same time the peer with the lowest download rate gets choked. This heuristic makes sure that even if there are good connections better ones are not forgotten trying from time to time other peers and their connections.

Anti-snubbing If a peer "A" with poor download rate gets choked by all other peers it was previously downloading, A will only be able to download if he gets unchoked by other peers. To solve this problem, if a minute passed by and A did not get any piece from any other peer, BT assumes it is "snubbed" by A and stops uploading to it except as an optimistic unchoking. This results in more than one concurrent optimistic unchoking.

Endgame Mode Shortly before downloading all pieces of a file a peer enters the endgame mode. In this mode all missing pieces are requested from all other peers. This avoids delays that would occur if the last view pieces would come from peers with slow transfer rates. For every full piece cancels are sent to the previously requested peers to reduce used bandwidth.

Upload Only If a peer downloaded the whole file the heuristics to which peers should be uploaded changes. It then does not prefer peers with highest download rates, because this can no longer be evaluated, but it starts preferring peers which no one else is uploading to.

2.1.3 Social Network Analysis

SNA is an interdisciplinary methodology of statistics, mathematics, sociology, psychology and more and more computer science. SNA is the construction of a model in which social entities are represented as nodes and their relations are represented as links between the nodes. In this way a graph or a so called network is constructed. This allows the construction of manifold networks such as a network of people working in an enterprise and their email communication as a link between them as described in [3] or people as nodes connected to other peoples they know as described in the famous study of Milgram. The study described a phenomenon in which any person would know any other

person over a median of five intermediate acquaintances [16] also known as six degrees of separation. SNA has a growing popularity in research articles in different fields. This effect is extensively studied in [19] where the growth in selected disciplines of articles using SNA are described, in fact its growth is mostly linear and applied in many different field such as biology, e.g., yeast protein interaction network [13], computer science, e.g., World Wide Web graph or source code [25], economy, e.g., interbank credit network [10], geography etc.

On these networks two different types of measurements might be calculated. Measurements on a network centric level and measurements on a node centric level.

The network centric level describes the network as a whole and makes it possible to compare two different networks with each other. Typical network centric measurements are diameter, the longest shortest path of any two nodes in the network, the average shortest path length of all nodes also called geodesics, density of a network which indicates how densely connected are nodes with each other or similar the cluster coefficient, which informs about the transitivity or connectedness of a node's neighbours.

The node centric level makes it possible to rank the different nodes according to their measurements. Typical node centric measurements are degree, closeness and betweenness they usually measure a node's prominence while a high degree or a high closeness alone might not indicate an important node per say, therefore nodes are looked for having a high rank in most of these measurements [27]. **Degree** describes the number of connections of one node. Depending on the graph if it is a directed graph or an undirected graph a node might have an in-degree the number of incoming links and an out-degree the number of outgoing links whereas in the undirected version would be the sum of the two. **Betweenness** is a measurement that describes bridge building properties, a node that connects two clusters with each other has a higher betweenness than a node in the cluster itself. It actually counts the number of shortest paths from any two nodes where the nodes lies on divided by the total number of shortest paths in the network. **Closeness** describes the position in the network. It is the reciprocal of the sum of all shortest paths to all other nodes also called fareness. The larger the closeness the more central the node. There are also other node centric measurements such as eigenvector centrality, eccentricity, neighbourhood size, coreness and local clustering coefficient but they will be discussed later.

2.1.4 Autonomous System

An AS is a collection of IP addresses and belong usually to one Internet Service Provider (ISP). Every AS has its own unique ASN and every IP has one ASN. The IPs of one AS might belong to multiple countries if the countries are small and the provider big. While one country might have many different AS and one provider also might have multiple ASs. Until 2007 the ASNs were represented as 16-bit integers which allowed 65,536 assignments.

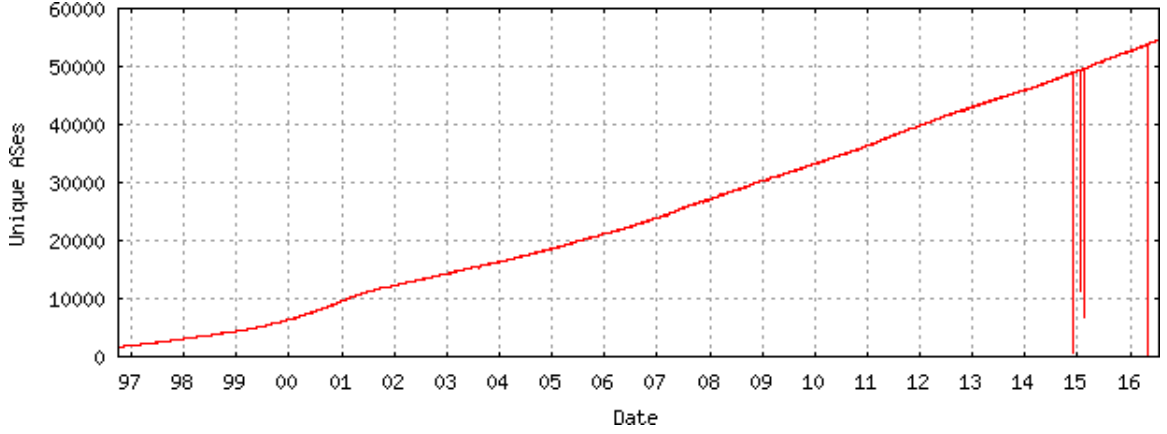


Figure 2.1: Unique ASs from end of September 1996 till mid of July 2016.

Afterwards 32-bit numbers got introduced which allows now roughly around 4,200,000,000 while some ASNs are kept for internal use only [33].

Figure 2.1 shows the number of ASN from the end of September till mid of July 2016. The number of unique autonomous networks in the routing system of the Internet exceeded 5000 in 1999, 30,000 in late 2008, 35,000 in mid-2010, 42,000 in late 2012, and 47,000 in mid-2014 and around 54,602 mid-2016 [38].

2.2 Related Literature

Related literature will be reviewed in fields important for this thesis. Where social network analysis was applied to analyse datasets or special tools such as Correlation, Assortativity combined with SNA was used to get more insights on the data.

Autonomous System

Two works were also centring their research on AS in a similar way compared to this work. Rossi et al. [23] are verifying time zones of ASs using game traffic data. This might also be possible with the network analysis data of this thesis. Pinthong and Lilakiatsakun [20] suggest a different piece selection algorithm based on the closestAS compared to the own AS to be able to overcome long distance delay.

Correlation

Tonelli et al.[25], Abbasi and Altman [1] and Putnik et al. [22] use SNA methodologies together with the statistical correlation to construct a Spearman correlation matrix. This matrix can be used combining the abstract network measurements together with real world data, to learn which network property correlates with the data. This might lead to

assumptions that qualify the network position either as good or bad for some real-world properties.

In [25] a network of the java qualitas corpus is constructed parsing source code and identifying the dependencies among classes, while two networks also have the number of bugs. This knowledge is used evaluating network measurements that have a higher chance provoking bugs than others and thus it helps predicting possible bugs in other corpora.

The research performance is measured in [1] comparing the h-Index to the network measurements of co-authorship network and a Spearman correlation matrix is calculated. The results of the analysis in the paper show that research performance is in fact positively correlated with weighted degree centrality and with the efficiency. Researcher with strong weighted links shows a better research performance than those with weak ones. But also here the authors would have liked more specific demographic data to get an even better image of the researcher and the influence of SNA-metrics.

Putnik et al. [22] measure student success and use extensively correlation analysis with different SNA measurement and asses the importance of each correlation. In a similar manner like [1]. The result is that a student has higher grades compared to the work volume and also higher final grades if the student is better connected to other students, if he connects many pairs of other students, has a high closeness to all other students, is connected to many other students that are themselves well-connected, and has strong ties. Also in this example the authors successfully draw conclusions from the influence of SNA-measurements to the overall performance of the students.

In [27] Valente et al. compare 58 networks of 7 studies and calculate correlations for some of the node related centrality measures. The networks have node count ranging from 45 to 83. The networks in the 7 studies are constructed by asking people names of people they would ask in case of family planing, best friends, friends in voluntary organizations etc. Degree has an average correlation coefficient of 0.69 the symmetric betweenness of 0.53 and a symmetric closeness 0.58 most of the other metrics are also ranging in this area. Unfortunately the calculation method of the correlation is not stated. An interesting approach is, that the correlation coefficients are then again compared with network-centric measurements like density, number of components and others, where the influence of the network-centric measurements to the correlation coefficients are analysed. The highest correlation in the study is between betweenness and eigenvector centrality with an average of $r=0.92$ and the next highest between betweenness and degree with an average of $r=0.85$.

Assortativity

Fricke et al. [10] construct an interbank credit network. Evaluating assortativity, the likelihood of nodes connecting to similar nodes, over time and compare generated Erdős-Rényi (ER) and scale-free random networks with the observed data. Interestingly, the authors found that the density influences the degree of assortativity, with the strongest level of dissasortative mixing for densities around 0.2. Unfortunately only the degree assortativity is examined, while other nominal assortativity examples would be generally

interesting but also hard or even impossible to compare with other networks. It is declared in the paper that assortativity is an important metric to evaluate the scale-freeness of a network where high dissasortative mixing would speak for scale-freeness.

Social Network Analysis

Mislove et al. [17] construct and analyse online social networks by following links to other pages. The networks are partially crawled and metrics of each graph are compared with each other. Giving insights about the topology of the networks and comparing individual features with each other. A similar way goes Wang et al.[29] focusing mainly on multiple small snapshot networks constructed from the real world Gnutella user data. These two papers give a good impression how a SNA based research on a new constructed should be done.

Peer-to-Peer together with Social Network Analysis

W. Wang et al. [30] and Upadrashta et al.[26] analyse the social network structure of P2P-systems to enhance the performance or to overcome specific P2P-system shortcomings. While not focusing on SNA but more on a general notion of a social network as a graph.

2.3 Preliminary Work

Prior to this bachelor thesis following courses and research was done in similar topics:

Business Network Analysis is an introductory course in the field of SNA. The course taught basics about all metrics, basic hands on experience with R, igraph[35] and data visualization with netdraw[37], data crawling, analysing networks and the meaning of measurements for the investigated network.

Seminar in Internet economics was a seminar about the influence of internet based technology and its influence to the real-world or economy. Where the selected seminar topic was the application of SNA to analyse the VIOLA[15] data on a conceptual level with possible enhancement of the system using the knowledge gained from the application of SNA methodologies.

Vertiefungs Arbeit in Big Data Analysis of Social Network Analysis graphs was a work in the field of Big Data Analysis examining different approaches of analysing such data. Examples are SNAP.py[39], Spark[32], Neo4J, and igraph[35] package of R. Where igraph and R were selected as best tools, not performance wise but because of its broad range of already available calculation of SNA metrics where other tools like Spark

were lacking out-of-the-box functions for most of the basic SNA metrics due to its young age.

Chapter 3

Methodologies

First all the research questions are stated and shortly explained why the author was led to them. Then the constructed graph is discussed explaining some graph basics. Afterwards, the data processing is looked at explaining the steps of the MapReduce job and the general work in R to retrieve information from the raw data. Finally, all the methods to answer the research questions are listed and explained. Containing topics like correlation, topology analysis, the network overtime and also cluster building.

3.1 Research questions

Now the research questions that were defined during the start of this thesis will be focused. The research questions will be used in this thesis as a guiding thread. The methodologies chapter and the results chapter will follow a similar sequence according to the research questions.

Q1:Are there correlations in the node centric measurements, with geographical or economical aspects, e.g., country or AS?

The positioning in the network may be used to predict real world properties. The comparison of the correlation of real-world data and network data is convenient because the network data is not clearly explainable or fully understandable. The SNA measurements are a quantification of some network properties. These metrics are highly depending on the magnitude of the network and only allow to construct a ranking of the nodes in the network. Thus they can be used together with real world data to give it a more explicit and comparable meaning.

Q2:What topological features has the constructed Network?

In SNA networks are often compared to each other, in this way networks retrieved from a similar context, share also similar features. Three main features are ER random, scale-freeness [31] and the small-world-problem [16].

Q3:How does the data change over time?

Since there are 48 networks in 48 hours, it is interesting to check different network measurements per time frame to check the network topology not looking at the graphs themselves but only on one specific measurement. This is done mainly for network-centric measurements, or node-centric measurement with normalized metrics like the centrality scores and not normalized metrics like some of the mean centrality measures.

Q4:Is a filtering of the VIOLA-data needed, do seed boxes change the network in a radical ways?

Seed boxes are remote-servers that are mainly used for down and uploading torrent files seeding much more files than an average user because of having much higher down and upload speed. In Figure 3.1 there is an AS with number 559 in the middle of the picture having a link to most other AS. The AS 559 is part of the VIOLA data crawling and thus changed the network in a dramatic way. This image led to this research question.

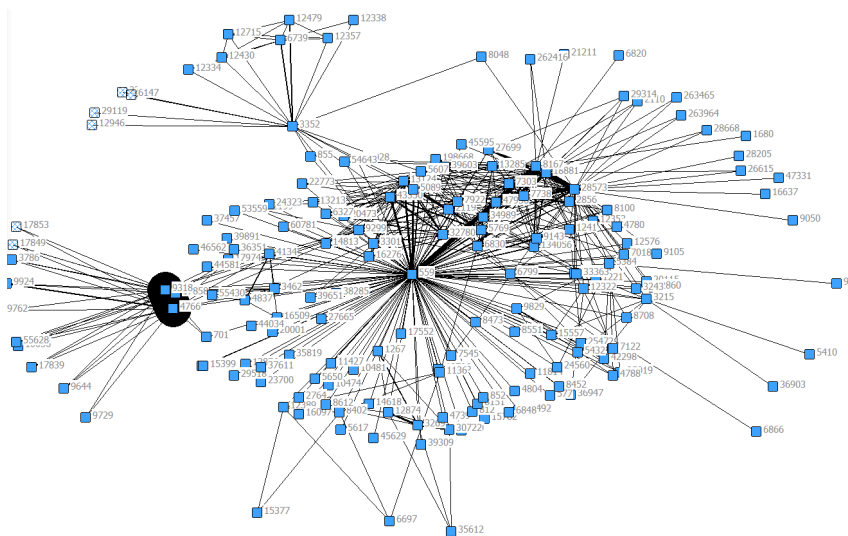


Figure 3.1: Visulaization of a sub graph with edge wheights above 1000 in netdraw

Q5:Are usage patterns visible in the network data?

It would make sense that the centrality of an AS is influenced by the activity of its users. Since BT-users have different activities on different times, like sleeping, working and being

in front of a computer in their leisure time. It is undoubted that a usage pattern would be visible on an IP level but it is not clear if it is also visible on an AS level, in consideration of that many people might let the computer on the whole day or the centrality of an AS could be relatively stable because of the huge number of users.

Q6: Are strongly connected clusters grouped by the country or continent of the Autonomous System

On an early stage of the research a picture was taken of a network with reduced size where only edges and their peers with strength above 1000 were looked at. Figure 3.1 shows some nodes building bridges to other nodes and also some cluster building. A first analysis showed that such clusters belonged to similar regions like Asia or South America.

3.2 BitTorrent-Graph

The constructed graph will be explained and it will be reasoned why it was built like this.

3.2.1 Set up of the Network

A social network shows an entity that has a relation to an other entity. Represented as nodes and edges. In the analysed model a user downloads or shares video data using the BT infrastructure. So there are three entities. The user, the AS of the user and the File. Since this thesis is mainly focused on the IP or for simplicity reason on the AS thus the model has to be reduced to one entity as node and one entity as edge to create a one-mode graph.

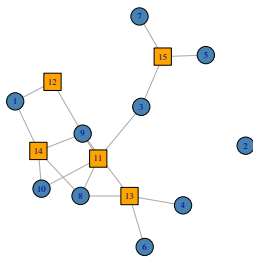


Figure 3.2: Two mode graph

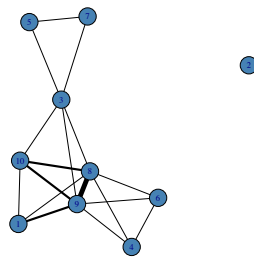


Figure 3.3: First projection

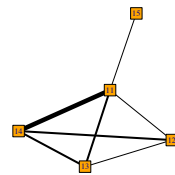


Figure 3.4: Second projection

A one-mode graph is a graph which only contains one sort of node. This is distinguishable from a two node model called two-mode graph. In the case of this VIOLA data the two-mode graph would be peers represented through their IPs that are connected with torrents

they share in their BT-client. An important property of the two-mode graphs is that only different nodes might connect to each other. So in our example of the VIOLA network peers are only connected to other peers by their torrents and not directly. The reason why one-mode graphs are better or more desirable than their two-mode representation is that most of the SNA metrics are not available in the two-mode world and accordingly most of the two-mode graphs are projected in two different one-mode versions. One centring on one kind of node that might be torrents and the other one focusing on the peers. Figure 3.2 shows a bipartite or two-mode graph with the two possible projection Figure 3.3 and Figure 3.4.

To reduce calculation time and simplify the network, the graph was set up containing ASs as nodes and not IPs considering the number of unique IPs is very high (13,003,033) and with a much higher number of edges in between them. An edge between two ASs exists if there are two IPs in the two ASs who are downloading or sharing the same torrent file during one time frame. If such a graph would be created with IPs all IPs that downloaded or shared the same file in one time frame would be fully connected to each other. In other words the constructed networks fully connects the online peers of one swarm according to their AS. This is illustrated in Figure 3.5 and 3.6 that shows one swarm in the two different representations.

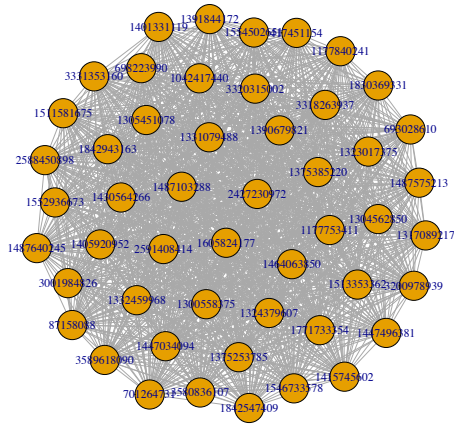


Figure 3.5: Fully connected swarm of one torrent where nodes represent IPs

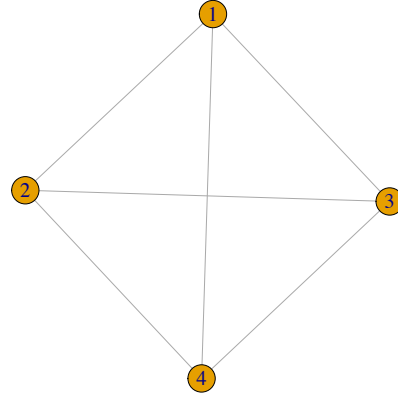


Figure 3.6: Fully connected swarm of one torrent where nodes represent ASs

3.2.2 Construction of the Network

The fundament for SNA is in most cases a simple graph this means there are no loops from one node to itself and there are no two edges between two nodes, which is not true

for directed weighted networks these might have two edges between two nodes in two directions having different weights. For the construction of the graph either an adjacency matrix or an edge list is used. Because adjacency matrices grow in two dimensions it is programmatically much easier to construct edge lists. Table 3.1 shows exemplary how an edge list looks like. The list contains three rows for directed graphs the first row is the node from where the edge is leaving and the second row is the node where the arrow would point at. The last row indicates the weight of the edge. The adjacency matrix for the same graph is shown in Table 3.2. An adjacency matrix for an undirected graph is symmetric where the one of a directed graph is not. A possible representation of the edge list and the adjacency matrix is shown in Figure 3.7. In the way how the data collection of the VIOLA system was constructed the edges between two nodes do not indicate a direction and consequently the graph is undirected in such wise the sequence of the nodes in the edge list is in this case ordered in an arbitrary way.

from	to	weight
1	2	2
1	3	1
2	3	1
.	.	.
.	.	.
.	.	.

Table 3.1: Example of an edge list.

	1	2	3	.	.	.
1		2	1			
2	2		1			
3	1	1				
.						
.						
.						

Table 3.2: Example of an adjacency matrix representing a network.

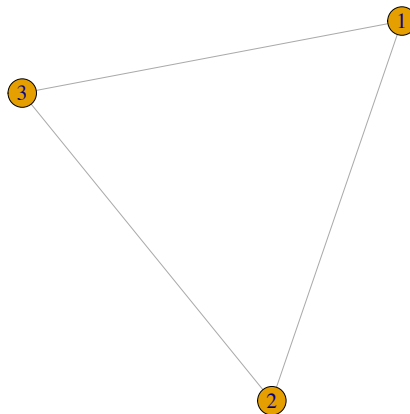


Figure 3.7: Graphical representation of the two examples.

3.3 Data Processing

The steps taken to process the raw data gathered by the VIOLA-system to construct a network are focused on and also the procedure of calculating SNA metrics on it.

3.3.1 Data

The data used and analysed in this thesis is the one gathered by the VIOLA-system [15]. The data contains usual IP information that would be received from the tracker and also some information from the system. The content of the torrents from the collected data are only videos. The main difficulty of the data gathering done in [15] was that for every request the tracker would return a random set of IPs which bares a lot of redundancy because of the mentioned BT protocol. The main advantage of this data set is the time component that indicates which IP was online and able to share a specific infohash at which point in time.

The analysed data contains two days according to Coordinated Universal Time (UTC) from the first hour 13th March 2016 to the very last of the 14th March 2016. The data of these 48 hours contained 7,310,217,062 entries stored in a Hadoop File System (HDFS), where 13,003,033 unique IPs, 19,709 torrents and 21,576 ASs were registered by the VIOLA-system. The biggest swarm has 56,408 peers, the highest number of leecher is 17,487 and the highest number of seeders is 44,866. The data contains information about peers from 243 countries and 60,950 cities.

3.3.2 Construction of the Edge List

The construction of an edge list for SNA may be accomplished in various ways, two of the possible methods are Structured Query Language (SQL) or MapReduce both of them will be explained.

SQL Example

To be able to calculate SNA measurements of the VIOLA data. The data in the .avro files have to be translated either into a matrix or into an edge list presentation. For small networks this can be done using anSQL statement like shown in listing 3.1. The query return for every peer that have one a file in common an edge reducing the list with weights with the group by command. For larger networks the SQL approach is not possible any more because of huge joins. Therefore MapReduce is used.

Listing 3.1: SQL query to form edge list.

```
SELECT p.asnumber , q.asnumber , count(*)
FROM peers AS p, peers AS q
WHERE p.infohash like q.infohash AND p.asnumber < q.asnumber
GROUP BY p.ip , q.ip
```

The Network data is analysed in R with packages like igraph. One of the advantages of the VIOLA dataset is the information about time. This helps splitting up the data and only looking at a certain interval, this results in a smaller network and thus the possibility to calculate network properties and measurements in a more time efficient manner. In this way the main focus of the analysis lies according the UTC on the first hour of the 13th to the last hour of the 14th March 2016. The 13th is a Sunday and the 14th is a Monday. The data was split up in 48 different networks each containing one hour of data.

MapReduce

Because of the huge amount of data the edge list creation was accomplished using a MapReduce job with three different stages. The VIOLA data contains: Timestamps, Infohash, ip(as unsigned int), port, AS number, continent, country, city, announceID. There are many possibilities to create networks with this information. But in the case of this thesis the edge list should contain ASN and connect every AS when there is an IP downloading the file as an IP in another AS. This case would result in a direct link between the two ASs. So the constructed network shows possible data sharing but not real internet traffic.

The real traffic would be hard to measure, considering that most of the communication happens in fact between the peers. This process is mainly depending on the BT protocol and how the pieces and sub-pieces get shared among the peers. Unfortunately this is not related to the tracker and might accordingly not be retrieved. The BT protocol is described in section 2.1.2. Following the three chained MapReduce jobs are explained:

- In the first step of the MapReduce job stores in the mapper for every unique combination of IP and infohash an entry of the IP's ASN and its infohash. Where the results gets reduced by the reducer such that only the unique combinations of infohash and ASN gets saved into another avro file.
- In the second step the mapper reads the previously created avro files and stores it in context. While the reducer creates an edge list of all possible combinations between AS having the same infohash.
- In the last step the mapper stores the previously generated list in context and the reducer reduces all duplicated edges adding a weight to the edge.

Such a job was started using a Virtual Private Network (VPN) connection to the sever and the entering following code listed in listing 3.2. The first line starts the job using the provided jar file and the class containing the entry point. The next two lines are parameters for the job the input avro file and output path both in the HDFS.

Listing 3.2: MapReduce jobs starting code

```
hadoop jar MapReduce.jar org.stohler.ViolaWeightedEdgeListFlat
/user/lareida/database/partitioned/2016/3/14/5
/user/stohler/test/2016-03-14-05_100
```

The first step of the job creates 15 avro files for one time interval where each of these files is around 110 MB. The second step creates 15 text files each being around 50 MB. And the last step returns 15 text files of 5.7 MB size. So the whole output of one time frame is approximately 2.5 GB where only 90 MB are used for further calculations in R.

3.3.3 Filtering the Data

At the beginning of the study a network was constructed containing only edges with weights above 1000 in this net there was one AS that had a much higher degree than all others. It was the AS where the data of the VIOLA data set was gathered. To measure the difference between filtered and unfiltered data IPs having a high infohash count were hard coded and filtered from the net. The impala query shown in listing 3.3 returned IPs having a high infohash count. This was accomplished in several steps where first only IPs having more than 500, then 100 and finally 20 infohashes were filtered. To asses this question a data set of one hour got filtered.

Listing 3.3: Impala query list the count of infohashes per IP if it is above 500

```
select a.ip, count( distinct infohash )
from violapart.announce a
where a.pyear = 2016
and a.pmonth = 3
and a.pday = 13
group by a.ip
having count( distinct infohash)>500
```

3.3.4 Analysis

R is a language and also an environment used for statistical computing and graphics. It provides many statistical instruments and offers many packages which makes it very extensible. It allows easy and fast data analysis. Most of the graph centric part of the code was written using the igraph package [35]. But as soon as all the network related calculations were made the hard part of saving data and re accessing data began. To be able to run calculations on multiple networks simultaneously different sessions had to be started using ssh to login to the R-server. The sessions were detached from the server in this way if the connection of the laptop would break the calculation would not stop and run separately. The main difference of starting a session in R running an R-script and running R in the browser or through the client RStudio is that the global environment this means all the previous defined data sets, all functions etc. are not available. To seize this problem most of the data used for such calculations like all the 48 networks were saved as RDS files. RDS files can easily be accessed in R and data can be well saved to these files. Because most implementations of R functions are not fully able to use a multiple core CPU to its limits, starting many sessions that calculate a part of the whole job makes sense. In such a manner calculations that would take on a single session 144 h could be done in 4 sessions simultaneously in 36 h.

For guidance in R coding and SNA in R stackoverflow [40], stat.ethz.ch [41], Quick R [42] and the book "Statistical analysis of Network Data with R" [14] were consulted.

3.3.5 Querying Data with Impala

The query UI Impala was used for general data analysis. Impala allows to form queries on Apache Hadoop data stored in HDFS. The query language is Cloudera Impala Query Language which is based on HiveQL. This was used to get information about the ASs distribution of IPs among AS or also to retrieve country and continent information.

3.4 Methods to Answer the Research Questions

In a similar manner like the Background section important Methodologies to answer the stated research questions will be explained and also some reasoning why the selected tools were used will be provided.

3.4.1 Correlations

Following part focuses on the different correlations calculation of Spearman, Pearson and Kendall, their differences and their usage.

Pearson, Kendall or Spearman

Correlation is an analysis that measures the association between two variables. The value may vary from -1 to 1. In such wise a value close to 0 means that there is no sign of correlation or no sign of any linearity of the data set. Values from 0.1 to 0.29 represent small association, from 0.3 to 0.49 represent a medium association, where above 0.5 represent a strong associate relationship [5]. The correlation score is not the only important value to asses the correlation of data also the p-value has to be on a significant level. Otherwise it can not be concluded that there is a correlation. If the significance level of $\alpha = 0.05$ is chosen and the p-value is $p > \alpha$ then the null hypothesis can not be rejected and therefore the conclusion is that there is not sufficient evidence that there is a significant relationship between the two examined variables.

In statistics there are three types of correlation calculation.

Pearson's Correltation

Or also called Pearson's product momentum correlation is a standardized measurement of the strength of a relation between two linear related variables [8]. It may answer

questions of the significance of a correlation between two measurements. Assumptions to be able to use the Pearson correlation are that the data should follow a straight line relationship between the two variables also called linearity and that the variables should be normally distributed along the regression line also called homoscedasticity. Pearson is more sensitive towards outliers compared to Spearman rank correlation.

$$r_s = \frac{N \sum xy - \sum (x)(y)}{\sqrt{[N \sum x^2 - \sum (x^2)][N \sum y^2 - \sum (y^2)]}} \quad (3.1)$$

where:

r_s = Pearson r correlation coefficient

N = number of value in each data set

x = scores of x

y = scores of y

Spearman's Rank Correlation

The Spearman rank correlation is a non-parametric test, this means that there are no assumptions made about the distribution of the variables. It may evaluate the degree of association between two variables. For the Spearman rank correlation no assumption of the distribution of data is needed but the data has to be at least ordinal. It is the Pearson correlation coefficient performed on data that has been converted into ranked scores. Spearman is the correlation which is most widely used [11]. It shows monotony rather than linearity.

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)} \quad (3.2)$$

where:

ρ = Spearman rank correlation

d_i = the difference between the ranks of corresponding values x_i and y_i

n = number of value in each data set

and all the ranks are different otherwise a more complicated formula has to be used.

Kendall's Rank Correlation

Kendall rank correlation is also a non-parametric test, thus no assumptions about the distribution are made. It measures the strength of dependence. Concordant means ordered in the same way discordant means ordered differently. Kendall's τ differs in magnitude compared to the ρ of Spearman and it is usually smaller than the latter. For the calculation

the difference of the ranks and not the subtraction like Spearman is used. According to [8] Kendall's τ should be preferred if the data set is small and has a large number of tied ranks. According to [11] Kendall's version of the Rank Correlation was less used in the past sixty years because of the more complicated calculation compared to Spearman, this was already noticeable in the igrph calculations, where Kendall's calculation would take much longer compared to the other two approaches.

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \quad (3.3)$$

where:

n_c = Number of concordant

n_d = Number of discordant

Correlation Matrices

Correlations are evaluated comparing two different data sets to assess their relationship. In case there are more variables that should be cross compared with each other, each of the three possible calculation of correlation may be used to construct a correlation matrix. This matrices are symmetric square matrices although only the lower or upper triangular part is used and the diagonal is always one because the correlation of a data set with it self is always one. The correlation matrix allows to express many different correlation calculations in only one figure.

Statistics of Correlation

Every correlation is a hypothesis test which assess if the null-hypothesis is accepted or rejected. The null-hypothesis says that there is no correlation among the two variables. If the p-value of the correlation is below the chosen significance level of a two tailed hypothesis test the null-hypothesis is rejected. Usual significance levels are either 0.1, 0.05 or 0.01, while 0.05 was chosen for this correlation analysis.

General Notes

Putnik et al. [22] used Spearman Correlation instead of Pearson because of no assumptions about the distribution were taken and also it might be used for variables not measured on an interval scale but also on an ordinal scale. Also Tonelli et al. [25] used the Spearman Correlation as well but with no clear reasoning of the selection. Hauke and Kossowski [11] compared results of correlation analysis using Spearman correlation and Pearson correlation. The analysis showed that the significance of the Spearman correlation can lead to the significance of the Pearson correlation. While the logical reasoning the other way round Pearson significance to Spearman significance is not possible. In the analysis

an example of having a positive Pearson coefficient with a negative Spearman coefficient was found. So the conclusion of the authors is to not over interpret Spearman's rank correlation coefficient as a significant intensity measurement of the relationship between two variables[11].

3.4.2 Topology Analysis

Focusing on the different possible graph topologies the constructed graph could have, the three important examples are mentioned, explained and compared to analyse what distinguishes them from each other.

Erdős-Rényi (ER) Random Graph

To be able to asses the topology of a network it is compared to an ER random graph of the same size regarding vertex and edge count. There are two different types of ER graph models:

- $G(n, m)$
- $G(n, p)$

The $G(n, m)$ takes number of vertices and number of edges as arguments and the $G(n, p)$ takes the number of vertices and the probability of having an edge between any two vertices. In the latter the expected number of edges is equals to $\binom{n}{2}p$. Equations 3.4 and 3.5 show the degree distribution of a random network [7]. While Equation3.4 shows a Poisson distribution.

$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \quad (3.4)$$

$$P(k) = \sum_{i=1}^N \frac{p_i(k)}{N} \quad (3.5)$$

where:

$p_i(k)$ = Probability that vertex i has degree k

N = Nodes

$\langle k \rangle$ = Average degree

Small World

According to Watts and Strogatz [31] a graph may be classified as a small-world network if two structural properties are met. The network has to be compared with an ER random graph of the same size. Then the average local clustering coefficient and the average shortest path length are evaluated in both cases. If the clustering coefficient of the examined graph is much larger and the average shortest path is shorter or about the same than the corresponding values of the random graph, the network may be categorised as small-world.

Scale-freeness

Scale-free networks are containing hub like nodes with much higher degree than the average node. Therefore the vertex connectivities follow a scale-free power-law distribution. This property is shown in equation 3.6. The equation indicates that the probability $P(k)$ of a vertex in the network interacts with k other vertices decays as a power law [2].

$$P(k) \sim k^{-\gamma} \quad (3.6)$$

where:

$$\begin{aligned} P(k) &= \text{probability of connecting to } k \text{ vertices} \\ \gamma &= 2 < \gamma < 3 \end{aligned}$$

To assess the goodness of fit R^2 or R-squared may be calculated. R^2 is a number from 0 to 1, that describes in general how well the model fitted. The construction of scale-free networks was widely studied and many different algorithms exist now to construct such a network. These networks are constructed using two steps.

- Adding or Removing nodes.
- Preferential attachment: The chance of adding a new node to an old one.

These steps assure that the generated network has a power-law distribution.

Many real-world network do have scale-free features such examples are Social networks like the collaboration network, Protein-protein, interaction networks[13] or Interbank credit networks[10].

Graphical Comparison of the three Models

Figure 3.8 shows the three models with their plot and their degree distribution. The ER random graph might be characterized as nodes with equally distributed edges. The small-world graph as many groups having circular connections to each other whereas some members of groups do also have connections to other groups and finally the scale-free graph as having a starting core growing from there in all directions resulting in the typical power-law degree distribution having nodes with a very high degree but having much more nodes with a very low degree.

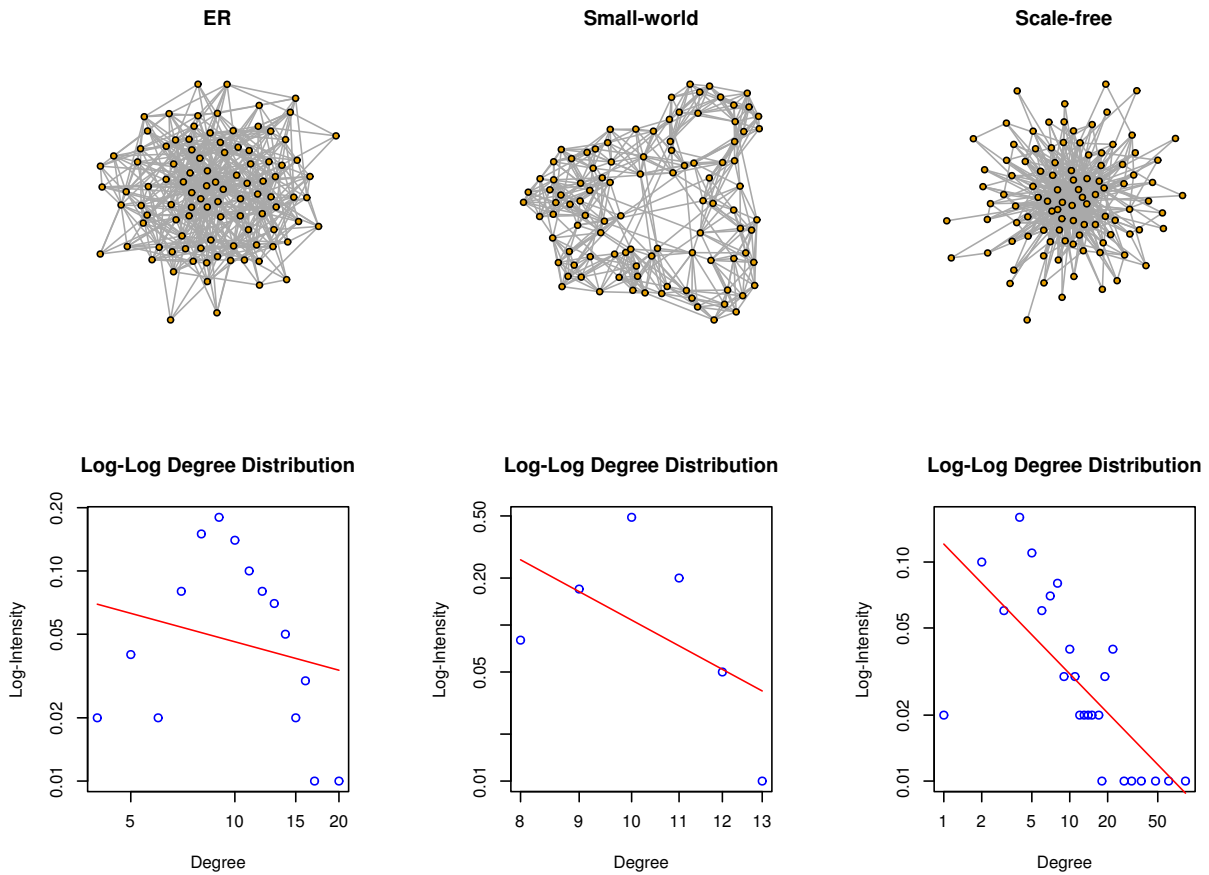


Figure 3.8: Comparison of the three models with a plot and the corresponding degree distribution. Number of nodes is 100 and number of edges is 500.

3.4.3 Network over Time

Some diagrams of network parameters will be discussed over the whole analysed period of the two days. Furthermore an example of an animation was constructed to visualize the network and its behaviour.

Measurements over Time

The data that is already split up by the hours is used to construct one network per hour. On every network, network centric measurements are evaluated and then compared. Three different types of measurements might be distinguished:

Mean of node-centric Measurements In this category mean degree, betweenness, closeness, eigenvector, coreness and eccentricity were selected to examine over time. These measurements are not normalized because the centralization scores are normalized node centric measurements describing the whole graph.

Centralization Scores In this category degree centralization, betweenness centralization, closeness centralization and eigenvector centralization were used to compare the results of each time frame.

Centralization is a method for calculating a graph-level centrality score based on node-level centrality measure[35]. The centralization scores are calculated using the formula shown in equation 3.7 using corresponding centrality measure. The scores are a normalized version of the node centric centrality measures.

$$C(G) = \sum (max(c(w), w) - c(v), v) \quad (3.7)$$

where:

$c(v)$ = is the centrality of vertex v

Degree Centralization measured the number of connections one node has. A high degree means that a node has many connections. In the case of the VIOLA-network a high degree means that an AS has many files in common with many other ASs. Since the VIOLA graph is undirected it exist only one kind of degree.

Betweenness Centralization measures how many shortest paths go through one node. The more shortest paths go through one node the more important is this node. The node is said to be in the centre of information flow. A high betweenness is reached when one node connects two groups while these groups are not otherwise connected. In case of the VIOLA network an AS has a high betweenness if the AS contains peers downloading different content while no other AS has the same two groups of peers.

Closeness Centralization is the inverse of the fareness measurement. Fareness measures the shortest path from one node to all other nodes. The lower the Fareness or the higher the closeness the more central is a node. A node has a higher closeness in the VIOLA graph if the node has a low eccentricity value of 2 and having as much nodes as possible as direct neighbours.

Eigenvector Centralization is a variant of Google's PageRank. The eigenvector centrality scores connections to high scoring nodes higher than the one to low scoring nodes. Therefore a high eigenvector centrality is reached when important nodes connect to other important nodes.

Network-centric Measurements In this category the number of vertices, the number of edges, average shortest path length, density and global clustering coefficient were analysed and compared. While diameter is also a typical measurement also to understand the size of a network. The diameter stayed for all time frames the same, particularly the maximum of the eccentricity which is three.

Animation

Similar to the Figure 3.8 sometimes images or motion pictures might help to get a better understanding of what is happening to the data. Therefore an animation of a reduced set of vertices will be constructed and made available online. The reduced set of vertices is important to get a better readability of the data. It is not advisable to draw the full dataset if it is big, having many nodes being well connected, considering the network could be that well connected that edges would not be visible any more because of the very high number of edges compared to the number of vertices. In this way the network will be reduced to edges with weights bigger than 2000 and another one with edges bigger than 1000.

3.4.4 Clusters

The VIOLA dataset contains information about peers downloading movies. The movie files have a recorded language and in some cases also subtitles, considering this it would be sensible if only peers are downloading movies in languages they also understand. The constructed network models ASs as nodes representing many peers. Since every AS is located in one country on one continent, therefore it can be assumed that a high percentage of peers of one AS would speak the same language. To statistically prove that the country or continent influences the position in the network or the groups of neighbours the assortativity and the cluster building properties have to be analysed.

Cluster building

In loosely coupled networks there might exist multiple smaller network not connected to each other. These are called components. In better connected network there is only one component the whole network. But the network might still be divided into clusters. A cluster is a community of nodes in which contains nodes that are better connected to each other than compared to the rest of the network. Clusters have to be computed, this is accomplished in various ways where some algorithm use the modularity. Modularity is a structural measurement, if the modularity is high it means that there exist many edges in

the cluster but only a few out of the cluster. Most of the cluster building algorithms try to maximize the modularity of each cluster using split and merges. These computations are time consuming[4].

Fast Greedy algorithm by Clauset-Newman-Moore[4] is a hierarchical, bottom-up approach that tries to maximize the overall modularity in a greedy fashion. Where at the beginning every vertex belongs to its own community, the communities get merged with each other to further optimize the modularity. The algorithm stops if the maximum modularity is reached. This algorithm is fast compared to other approaches.

Walktrap algorithm by Pons and Latapy[21] uses random walks on the network. The random walks have a higher chance of ending in the same community. The results of these random walks are used to merge communities in an bottom-up fashion like the fast greedy algorithm. The algorithms runs generally slower compared to the fast greedy algorithm but it is also said to be more precise.

To analyse the network and the underlying structure cluster algorithms will be used to divide the network by its clusters and then a census of the country and continents will be done to asses if country or continent do play a role in the cluster building process or not.

Assortativity

Assortativity or also called homophily describes the tendency of individuals linking together with similar ones. Homophily is often used in online marketing where customers get goods recommended by the system which they actually might like. A network has an assortative mixing if high degree nodes also tend to link to other high degree nodes. A network with a high degree of assortative mixing is resilient to the loss of high degree nodes[18]. Where none assortative or also called dissasortative networks fall apart easily in case of a loss of a high degree node. Real-world networks show either an assortative or a dissasortative mixing where ER random graphs have an assortativity coefficient of zero[18]. Assortativity is nothing else than the Pearson correlation coefficient of node-degree to neighbours-degree [18]. In Equation 3.8 the calculation of the assortativity coefficient is shown. Assortativity is not limited to degree any numeric value may be compared and also nominal properties such as the assortative mixing of one nodes country attribute to the one of its neighbours is possible.

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2} \quad (3.8)$$

where:

- r = assortativity coefficient
- e_{jk} = joint probability distribution
- $\sum_{jk} e_{jk} = 1$
- $\sum_j e_{jk} = q_k$
- q_k = normalized distribution of the remaining degree
- n = number of value in each data set

Chapter 4

Results

The results of the calculations and instruments described in the methodology chapter will be given now. The sequence follows again the order of the research questions formulated in chapter two.

4.1 Q1: Correlation

The correlation evaluation was first only conducted on one time frame because of the manifolds cross-relation possibilities. Therefore one of the 48 time frames was selected, UTC 4 am 14th March 2016. Afterwards the selected time frame got compared with the mean over all time frames.

First the chosen measurements and data will be explained. Then the plots are examined. Each correlation stands for a hypothesis test that might be accepted or rejected according to the result of the correlation matrix where the p-value tells the significance of the test while the significance level was chosen to be five percent of a two tailed test. If the p-value is above the significance level the test leads to accept the null-hypothesis which says there is no correlation among the two variables. For each method Pearson (Appendix A.3, A.6, A.11) the p-values, the correlation coefficient (Appendix A.2, A.4, A.10) and a visualizing diagram (Figures 4.1, 4.2, 4.3) are shown, where in case of Spearman also some more data is listed in the appendix.

4.1.1 Selected Node-centric Metrics

All values checked for correlation on a node-centric level are shown in Table 4.1. The name is indicating the measurement, the letter will be used in tables in the Appendix as reference and a short explanation of the measurement.

Table 4.1: Node-centric measurments checked for correlation

Name	Name in the plot	Explanation
Degree	degree	Number of edges per node.
Weighted Degree	weighteddegree	Summation of all the edges weights of one node.
Betweenness	betweenness	Number of shortest paths going through the node divided by all shortest paths.
Closeness	closeness	Inverse of the sum of the fareness to all nodes.
Eccentricity	eccentricity	Fareness of the most distant node.
Eigenvector	evcent.vetcor	Eigenvector centrality of a node is high if he links to other important nodes.
Coreness	coreness	A vertex u has coreness c if it belongs to a c -core but not to any $(c + 1)$ -core.
Neighbourhood in two hops	n2	Neighbourhood of a node reachable in two hops. Where the neighbourhood of one hop would be the degree plus one, the starting node itself.
Local Clustering Coefficient	localcc	The local clustering coefficient may vary from 0 to 1. It is zero if no neighbour is connected with any other neighbour and it is one if all neighbours are fully connected to each other.

4.1.2 DataBank Data by Country

As an example real world data of countries was added for the correlation calculations. While this is generally a good approach but unfortunately there was not sufficient information available about the different AS holder so data of the corresponding countries was used World DataBank[45]. The selected datasets are from 2014 because the same sets of 2015 were not complete or in 2016 not yet existent. The selection of the datasets were thought to be influential for the constructed net network. The three sets are Gross Domestic Product (GDP) per country in USD (\$) denoted as GDP or **gdp**, growth rate in USD (\$) denoted as growth or **g**, the usage of mobile phones denoted as mobile or **m** and the usage of internet both per 100 inhabitants denoted as internet **i**. Table 4.2 shows an overview of the stated variables.

Table 4.2: Country dependent real-world data

Name	Name in the plot	Explanation
GDP	GDP	GDP per country in USD (\$)
Growth	growth	growth rate in USD (\$)
Mobile	mobile	Usage of mobile phones per 100 inhabitants.
Internet	internet	Usage of internet per 100 inhabitants.

4.1.3 Raw Data from Autonomous Systems

Additionally to the real-world data of the countries Count of IPs per AS (CIP) also denoted as **countips** and Count of infohashes per AS (CI) also denoted as **countinfohash**, were queried on specific time frames using Impala on the original dataset and some extra information of the start and endpoint of the IPs on the AS was added to be able to calculate the theoretical Maximum number of IPs per AS (NIP) denoted as **numips**. The Table 4.3 shows an overview of the stated variables.

Table 4.3: AS dependent data

Name	Name in the plot	Explanation
CIP	countips	Count of IPs per AS
NIP	numips	Number of maximum possible IPs per AS
CI	countinfohash	Number of infohashes per AS

Pearson Correlation Plot

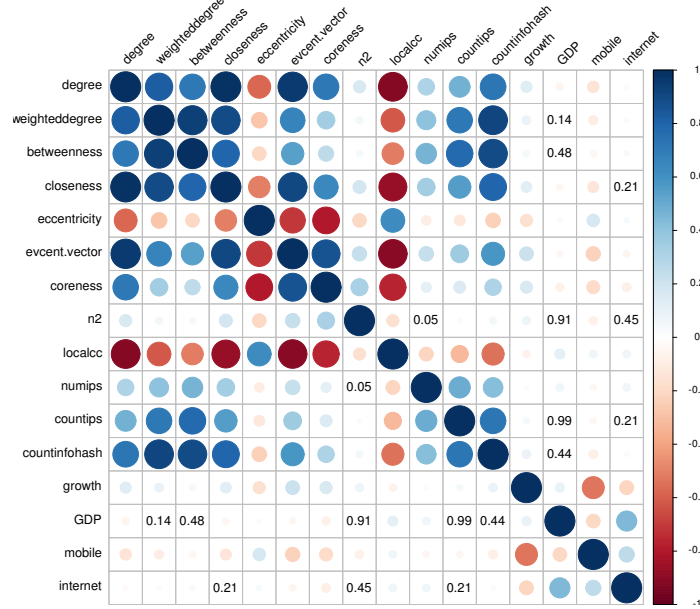


Figure 4.1: Correlation plot of Pearson correlation coefficient. Compared to the other two Pearson shows less uniformity. Interestingly there are regions in the plot where only weak correlation coefficients exist which are high in the other two plots. like NIP, CIP and CI, together with eccentricity, eigenvector centrality, coreness and n2. n2 is a very uniform variable since it has similar values for most of the nodes, this is also why Pearson measures only weak correlation because the linearity of the variable is not met.

Spearman Correlation Plot

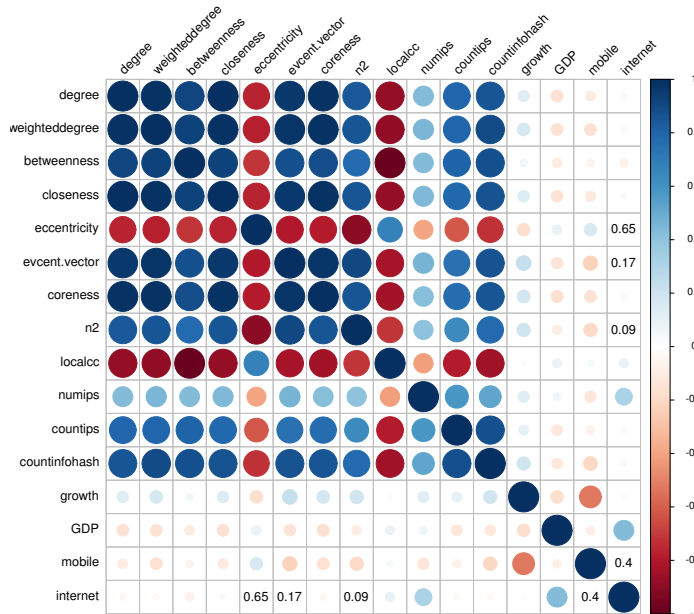


Figure 4.2: Correlation plot of Spearman correlation coefficient.

Kendall Correlation Plot

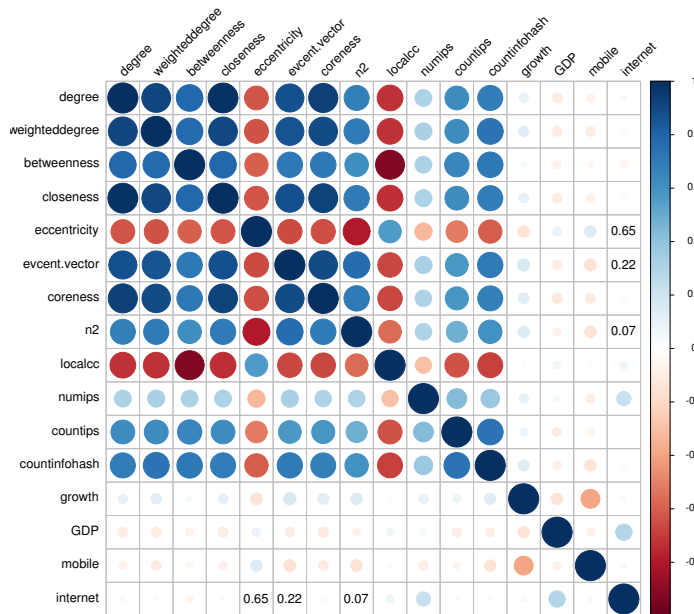


Figure 4.3: Correlation plot of Kendall correlation coefficient. Looks very similar like Spearman but with weaker correlation coefficients.

4.1.4 Comparison of Correlation Regions

In the correlation matrix exists three different types of variables, the network measurements, the impala query results and the World DataBank [45] datasets. For simplicity they will be called, Correlation variables of the SNA metrics (SNAM), Correlation variables of the Word DataBank (WDB) and Correlation variables of the raw AS data (RAW). The correlation matrix can be divided into six different regions. The correlation of SNAM to SNAM, the correlation of SNAM to RAW, the correlation of SNAM to WDB, the correlation of RAW to RAW, the correlation of RAW to WDB and finally the correlation of WDB to WDB. Some of these regions will be further analysed. Where the least interesting region is the correlation between one group of variables e.g. WDB to WDB etc. because of the question if we can derive Information about the real-world through the knowledge of the correlation of the real-world data and the network measurements. To be sure that the null-hypothesis can be rejected and not accepted the p-value of the correlation matrices have to be checked. The p-value has to be between zero and 0.05 because of a two sided test with significance level of five percent or the confidence interval of 95 percent. In this correlation matrix 120 different hypotheses are stated. Table 4.4 shows a census of the correlation per region. Only when WDB is involved very low coefficients are reached.

Table 4.4: Census of the Spearman correlation per region

region	0 to 0.09 none	0.10 to 0.29 weak	0.30 to 0.49 medium	0.50 to 1.0 strong	Σ
Σ	21	30	11	58	120
snam to snam	0	0	0	36	36
snam to raw	0	0	9	18	27
snam to wdb	14	22	0	0	36
raw to raw	0	0	0	3	3
raw to wdb	4	7	1	0	12
wdb to wdb	3	1	1	1	6

The results of Spearman and Kendall are very similar while the correlation coefficient is generally higher in Spearman than in Kendall. Also the p-values do more or less agree only in the correlation of internet and mobile the two do not agree where Kendall is assessing it as significant Spearman disagrees.

On first sight the p-values of Pearson and Spearman or Kendall (Figures 4.1, 4.2, 4.3) do not agree with each other. Spearman advises to accept the null hypothesis in four cases, Kendall in three cases and Pearson would accept it in nine cases.

A coefficient being closer to one or minus one than to zero means in the case of Kendall and Spearman that the two variables have a high degree of monotony. But Pearson says in case of coefficient far away from zero that the relation has a high degree of linearity which is also the reason why the Pearson correlation coefficient is lower than the other two. Pearson has less very strong correlations compared to Spearman. Interestingly Pearson evaluates a degree of linearity between degree, closeness, eigenvector centrality and a strong negative linearity with local clustering coefficient. And one last strong correlation in Pearson is

the CI with weighted degree. This is also plausible because the more infohashes an AS has the higher the weight gets in this case even an AS with one peer might have a high weight in case this peer downloads many of different files.

Because of the assumption that the variables need to be normal distributed and follow linearity the Pearson correlation will not be taken into account for further analysis. Since Spearman was used in [25] and in [22] and Kendall has very similar results to Spearman and was usually not used because of difficult calculation[11] it is valid to limit further analysis to Spearman correlation.

SNAM to SNAM

This region is very impressive since all the evaluated network values correlate strongly with each other. The two values negatively correlated are Eccentricity and the local clustering coefficient. A close up of this region is visible in Figure 4.4.

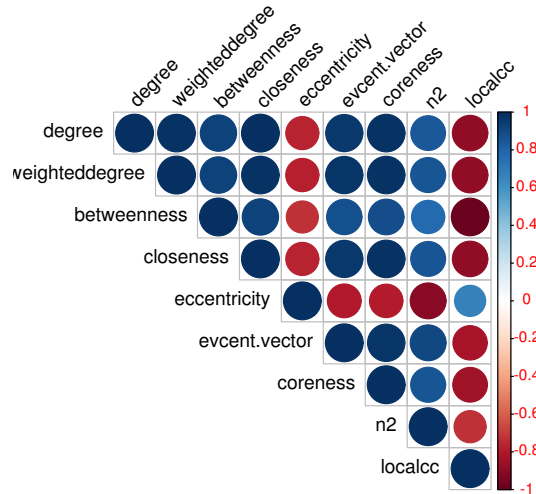


Figure 4.4: Spearman correlation of the region SNAM to SNAM

Eccentricity is a value varying from two to three with a mean of 2.561. Eccentricity is the measure of the longest of the shortest paths to all nodes in the network, therefore it is coherent that most of the measured SNA metrics negatively correlate with high eccentricity because nodes with eccentricity of three are less central compared to betweenness closeness and also degree compared to nodes with eccentricity of two.

Local Clustering Coefficient is the other metric negatively correlating. It measures the connectivity of the neighbours of a node. It varies from zero if no neighbours are connected with each other to one which means all are fully connected. One explication is that a high local clustering coefficient is negatively correlated with most of the metrics because the probability of having a high local clustering coefficient is higher if the node

does not have that many neighbours which also states that degree and local clustering coefficient is negatively correlated. The only positive correlation of local clustering coefficient is with eccentricity. So this means if the neighbours are well connected the chance is higher that the furthest node is reachable in three steps. While nodes with low local clustering coefficient tend to have their furthest away node reachable in two steps. So this correlation tells us that there are two types of nodes or AS in the constructed network the ones situated in a well connected cluster being three steps away from the furthest node and nodes not in a cluster but closer to the furthest node.

SNAM to RAW

The correlation SNAM to RAW, visible in Figure 4.5, is similar to most of the coefficients of the SNAM. NIP has usually the lowest correlation coefficient of the three variables. The correlation of CIP to the SNAM is in the median 64% higher than the NIP and the highest correlation coefficients among the SNAM values has the CI with 12% higher correlation coefficient than CIP and 87% higher than NIP.

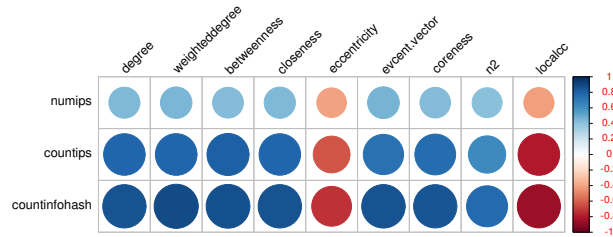


Figure 4.5: Spearman correlation of the region SNAM to RAW

The number of different torrent files has the strongest relation to SNAM. This is reasonable because the network is built in the way that torrent files build the edges. An AS with one IP sharing a high number of torrent files (which is the equivalent for having a high CI) is much better connected than an AS having a high CIP but with low diversity of torrents (infohashes). But it can be stated if there is a high CIP the chance of diversity is also higher and similar if an AS has a high NIP the chance is high that the AS has more customers and therefore also a higher CIP. This proves that the method is generally working.

SNAM to WDB

Compared to the rest of the correlation plot the part where WDB is involved only low coefficient values are reached, visible in Figure 4.6 This is probably because the data is not from the ASs but from their country. Growth presents a similar pattern like most of the metrics being positive for most SNAM metrics but negative for eccentricity and local clustering coefficient. The GDP, mobile usage (m) and internet usage (i), both measured

as how many people per 100 use mobile or internet, have their signs the other way round being negative where the SNAM correlations are positive and vice versa.

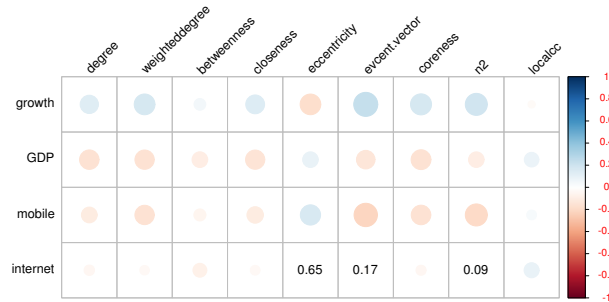


Figure 4.6: Spearman correlation of the region WDB to SNAM

A central position in the network leads to a higher growth but central nodes have usually a lower GDP, mobile or internet usage. This is meaningful if we state that poor countries are more likely to download movies than going to a cinema or simply buying the movie. In this way we would also suspect that internet usage should positively correlate these metrics, because only if people use the internet they are able to download files from BT. But apparently the usage of BT can not be derived from the usage of internet.

Internet usage is the only variable always part of where the p-values leads to accept the null-hypothesis and has overall the lowest correlation coefficients.

RAW to WDB

NIP correlates positive with growth GDP and internet but negatively with mobile. NIP has compared to the other correlation coefficient in this region a high positive coefficient with internet. This makes generally sense that the AS holder allocate enough IPs for future needs in a country where internet usage is already high. These effects are visible in Figure 4.7. NIP correlates negatively with GDP where the sign of GDP is usually positive. This correlation indicates that an AS with a high number of possible IPs has a lower GDP. Maybe poor countries allocate more IPs than rich ones. But the measured effect is on a very small scale and to understand this further analyses would have to be done.

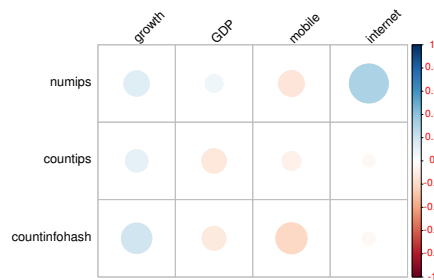


Figure 4.7: Spearman correlation of the region RAW to WDB

4.1.5 Comparison over the whole Time

The mean correlation coefficient over all time frames is shown in Figure 4.8. The comparison of the single time frame correlation and the whole measured time, also the standard deviation and absolute difference is listed in the Appendix A.7, A.8, A.9. The difference between the mean and one time frame is on a very small scale. The biggest difference is visible in the NIP and CIP variables. Both having about the same coefficient. NIP is a good example of a variable that belongs to the AS, which is not measured on a time frame basis and is highly dependent on the structure of the network, while growth, GDP, internet and mobile usage also valid for one year were not as much influenced by the structure of the network. The highest correlations per variable were measured in weighted degree with a mean correlation of 0.680, followed by eigenvector centrality 0.678 and closeness 0.672. While the lowest measured mean correlations were found in growth 0.201, GDP 0.201, mobile 0.175 and internet 0.140. This means weighted degree has over all variables the highest degree of monotony.

Spearman Correlation Plot Mean

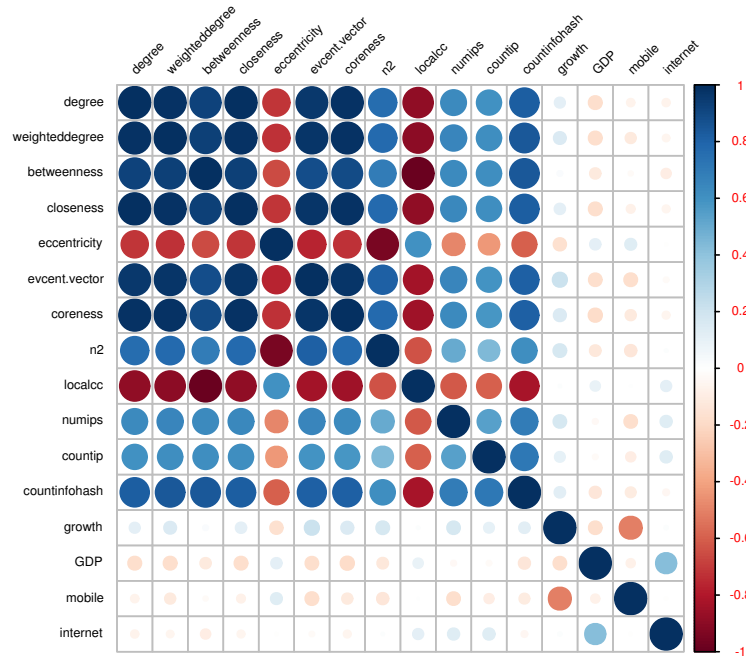


Figure 4.8: Correlation Plot over the 48 measured time frames.

4.1.6 Conclusions

For this study first one time frame was taken into account evaluating the different correlation models. Then Spearman was selected and the time frame got analysed. In the end the single time frame was compared with the mean over the whole period while most

of the variables were about the same only NIP which indicates the maximum amount of possible IPs per AS showed a much higher standard deviation. Interestingly, this is a measurement which is valid for a longer time similar to the variables in WDB and although it was much more influenced on the network than any of the other variables.

This part showed that the correlation of the SNAM to it self and also to RAW is relatively strong and non of the null hypotheses got accepted in these regions. In [27] the authors came to a similar result, that centrality measures are relatively strong correlated, even though the studied networks showed a much smaller correlation than the one evaluated in this study from the VIOLA network. The correlations of the WDB data was generally much weaker than measured in other regions. This might have multiple reasons. The valid time interval of the SNAM, this means the time during which the compared variables were true, is one hour, because the next hour already has different numbers. The valid time frame of the RAW data is mixed where NIP is probably for a longer period valid, CIP and CI was evaluated for one hour. And finally the data of WDB is depending on the country of each AS and was measured in 2014. So the data is not only not very specific for AS but also not in a coherent time frame.

In this way the analysis of correlation is a good methodology used together with SNA metrics but to be able to get further knowledge also "good" real-world data have to be available which was not the case in this example. If the corresponding research question is recapitulated: **Q1:Are there correlations in the node centric measurements, with geographical or economical aspects, e.g., country or AS?** And the answer is yes, in the 120 correlation evaluations, 58 cases of strong correlation, 11 cases of medium correlation, 30 cases of weak correlation and 21 cases of very weak correlation or to high p-values were found. Where the mean correlation measured is 0.505. Correlations of WDB had generally a much weaker coefficient than correlations of SNAM but as already stated above this is due to the quality and also the time wise validity of the data in WDB.

4.2 Q2: Network Topology Analysis

The three models small-world, scale-free and ER random graph will be discussed to answer the research question **What topological features has the constructed Network?**

4.2.1 Small World

To asses the small world properties local clustering coefficient and average shortest path length of the examined graph have to be compared with an ER random graph.

The local clustering coefficient in the constructed VIOLA graph is much higher compared to the ER random graph shown in Figure 4.9). The average path lengths of the ER random graph and the average path lengths of the VIOLA network are about the same length shown in Figure 4.10.

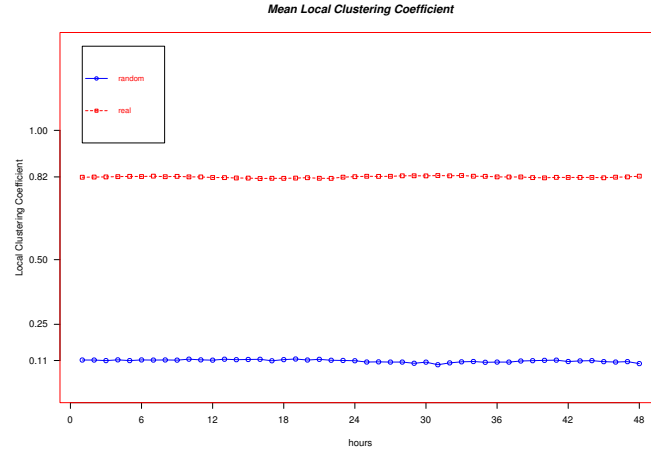


Figure 4.9: Local Clustering Coefficient VIOLA graph compared with ER random graph. Indicating a huge difference of the two networks.

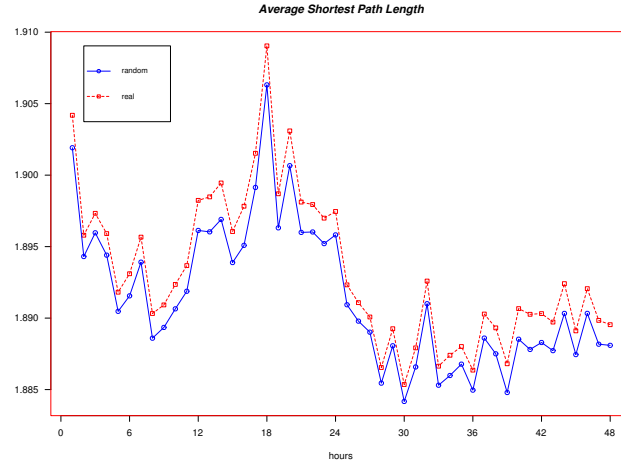


Figure 4.10: Average Shortest Path VIOLA compared with random with ER random graph, while bot networks have about the same average path length.

Considering that the two properties average shortest path length of the random graph is similar to the VIOLA graph and the mean local clustering coefficient is much larger in the VIOLA graph than in the random graph it can be stated that the network fulfils small world properties over the whole examined time period.

4.2.2 Scale-freeness

As an example the power-law fitting of an ER graph shown in Figure 4.11 and a scale-free random graph constructed with igraph is shown in Figure 4.12. The size of the two networks is chosen according to the size of the examined network of the first hour, midnight 12th April 2016. The R^2 value of the scale-free network is 0.6811428 and and the one of the ER-network is 0.00060146, which is indicating a better fitting power-law of the

constructed scale-free network. It is generally not possible to say that a high R^2 justifies the selection of the fitting model, so a high R^2 is in this case no real proof of a scale-freeness, therefore the plot has to be analysed. The fitting line of the VIOLA distribution follows a power-law. Figure 4.13 presents the R^2 value of all the measured networks. The R^2 value can be understood as the steepness of the fitting line. This indicates that the scale-freeness of the networks. Figure 4.14 shows the log-degree-distribution plot of the first measured hour and a steep fitting line.

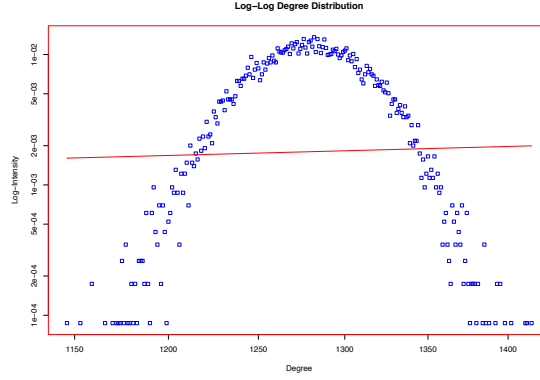


Figure 4.11: Degree distribution of an Erdős-Rényi random graph with 11464 nodes and 7352865 edges

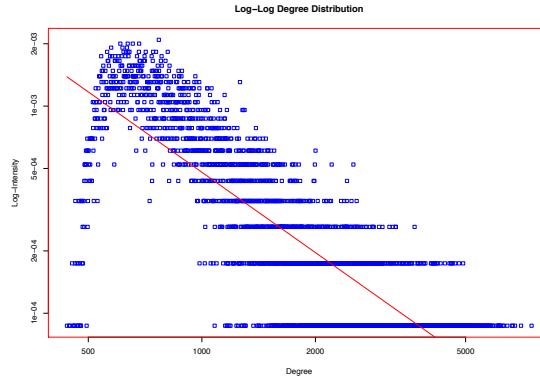


Figure 4.12: Scale-free random graph with 11464 nodes and 7352865 edges. Visible in this plot that the number of nodes with low degree is very small and, therefore the distribution forms an arc which is atypical in case of scale-freeness.

It could be argued that the fitting model is not met since the distribution does not follow a straight line. The intensity is decreasing again when it comes to low degree but since also the constructed scale-free network shows the a similar behaviour it can be stated that the graph has scale-free properties. As a final example a plot of a graph where only edges with weights above 1000 were selected to construct a sub-graph. Figure 4.15 shows the degree distribution of the sub-graph, which has a much better fitting compared to the other two examples with an R^2 score of 0.7496657.

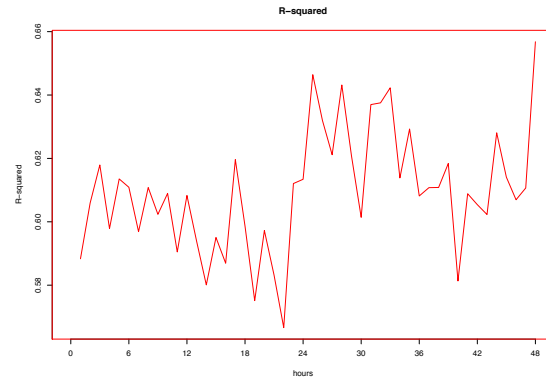


Figure 4.13: Shows the R^2 value of all the networks

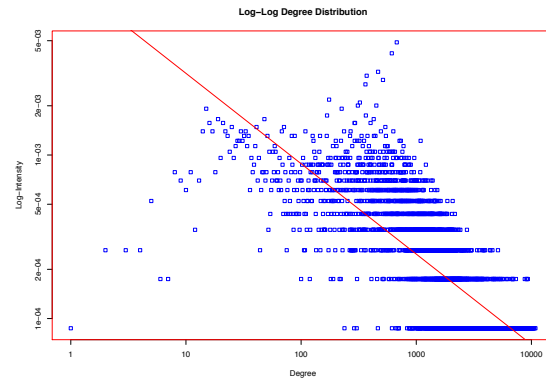


Figure 4.14: Degree distribution of the first analyzed hour April 12, 2016 12 pm with 11464 nodes and 7352865 edges.

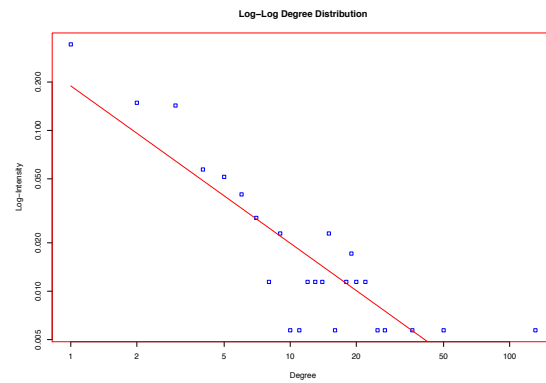


Figure 4.15: The diagram shows the degree distribution of a subgraph of the graph shown in Figure 4.14 where only edges with weights above 1000 were selected.

4.2.3 Comparison of Correlation among the Different Models

To assess the topology model the correlation of different random graphs was compared with the real-world data. Correlation plots of the real world data together with three constructed random graphs one as normal ER random graph, one with small-world features and one with scale-free features are shown in Figure 4.16. Weighted degree in the random graphs is the same as the degree for simplicity, and a question mark indicates that the evaluated variable has a standard deviation of zero which makes it impossible to evaluate a correlation. Interestingly, eccentricity and the size of the neighbourhood in two steps is in most random graphs the same two for eccentricity or 11179 for n^2 . Eccentricity is evaluable in the scale-free random graph because in eight cases the eccentricity is one where in all the other cases its two, this is enough to make it possible to derive correlation in this case.

It is hard to assess which of the models has a similar correlation pattern compared to the real-world data because of the eccentricity is not evaluated in most cases. But from the strength of the correlation the constructed scale-free graph has some similarities.

A feature visible in the small-world graph is that betweenness has a lower correlation coefficient than degree or closeness, this effect is also slightly visible on a smaller scale in the real-world-Spearman plot.

The high correlation coefficient in the scale-free-Spearman plot is explicable because of how a scale-free network is set up. A scale-free network has a growing structure from a central point this is also the reason why some nodes have an eccentricity of 1. This growing structure makes sure that the correlation coefficients are monotonic increasing or decreasing.

This example made it visible that eccentricity is an important metric in the constructed VIOLA graph which should be further analysed.

4.2.4 Q2: Answer

The research question on this topic was: **Q2: What topological features has the constructed Network?**. The topology of the constructed network is not easily distinguishable between scale-free and small-world. An important feature of the scale-freeness is the power-law distribution which is met asymptotically even though there are not many nodes with a small degree and the two properties of a small-world graph short average shortest path and high local clustering coefficient is also true. The constructed graph connects all AS together that have two IPs of the same swarm online but if the swarm is distributed over 10 ASs all these 10 ASs are completely connected to each other. Thus the created graph is well connected with short average shortest path and power-law degree distribution. The comparison of correlation plots from the real-world graph and the constructed random graph models is an interesting approach to see how the different models look as correlation plots and which model has similarities with the real-world graph. This was an experiment because of extensive analysis of correlation in previous part of this

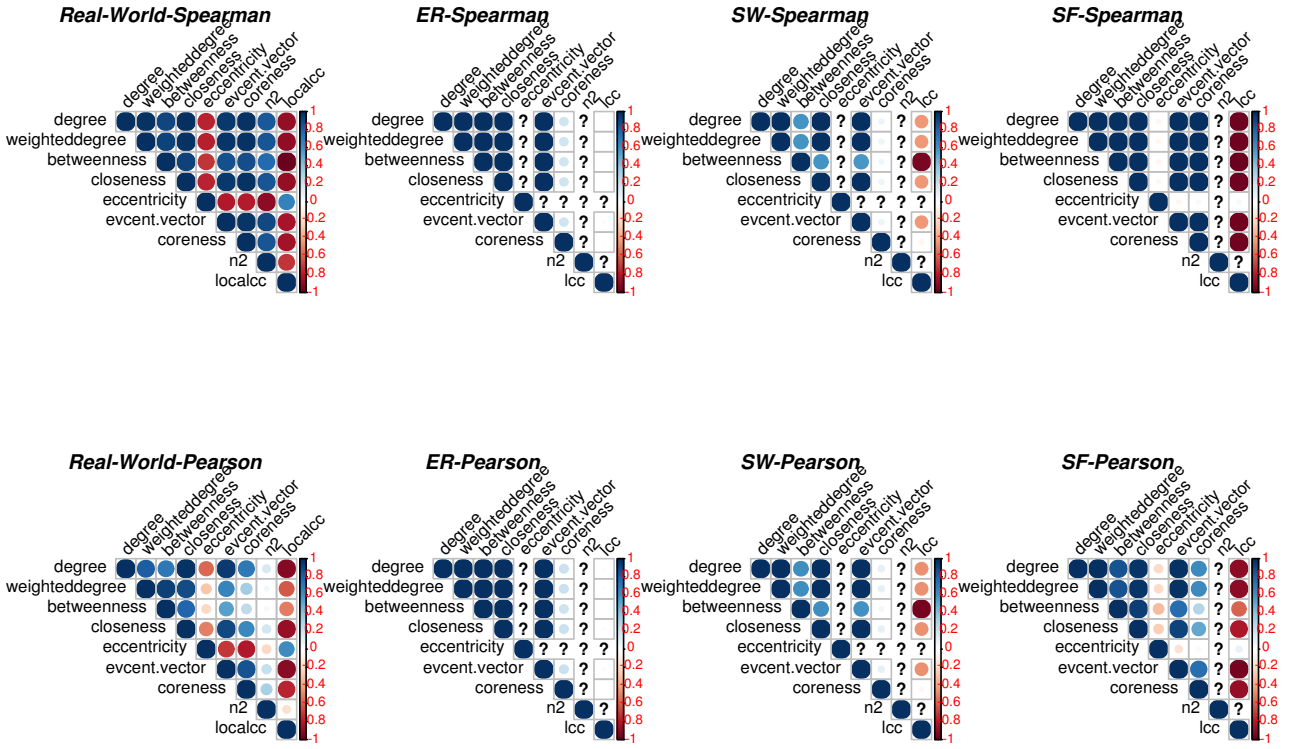


Figure 4.16: Correlation plots of the real-world data and three constructed random models the first according to ER the second with small-world properties and the last with scale-free properties. The first row shows Spearman correlation and the second row Pearson correlation.

thesis and needs refinement and further analysis with other already established research graphs.

4.3 Q3: Network over Time

The change over time of the network will be discussed looking at the metrics and the effect of time during the evaluated period. The subsection Animation will discourse the created animation and the corresponding insights that follow the animation and in the end some top nodes are discussed ranking high in the centrality measurements.

4.3.1 Metrics over Time

The change of the individual metric over time is shown by Figures 4.17, 4.18 and 4.19. Where the minimum value is on the bottom and the maximum value is on the top of

the diagrams. The diagrams are ordered by means of the node centric measurements in Figure 4.17, the centralization scores in Figure 4.18 and finally some network-centric measurements in Figure 4.19.

General Notes Mean degree, mean eigenvector, mean coreness, density and transitivity have their global minimum at hour 18 where also the global maximum of the eigenvector centralization and of the average shortest path is. This important point is also visible in vertex or edge count having a local or global maximum. The eigenvector centrality score and mean eigenvector centrality have their minimum in case of the mean and maximum in case of the score at the same point in time.

Node-centric Measurements over Time

Mean degree, mean eigenvector and mean coreness show a very similar behaviour of the curve. Having most local and global minima or maxima at the same position. A similar relation is visible in the mean betweenness and the mean closeness where closeness is the horizontally flipped version of the betweenness curve. The mean closeness has its global and local minima where the maxima of betweenness are and the same is true for the maxima of closeness with the minima of betweenness.

The curve of eccentricity does not match any other evaluated curves. Astonishing is that the Maximum is followed by a steep decay reaching almost the minimum eccentricity. This means that around UTC 0 of the second day an event happens that reduces the high eccentricity of most nodes, but during this time also the number of edges gets reduced. The eccentricity could be reduced if some of the nodes that are hard to reach for most of the other nodes in the network get better connected, so maybe a bridge building torrent file was released. Eccentricity has its maximum at a similar position where also the maximum of degree centralization and closeness centralization lies.

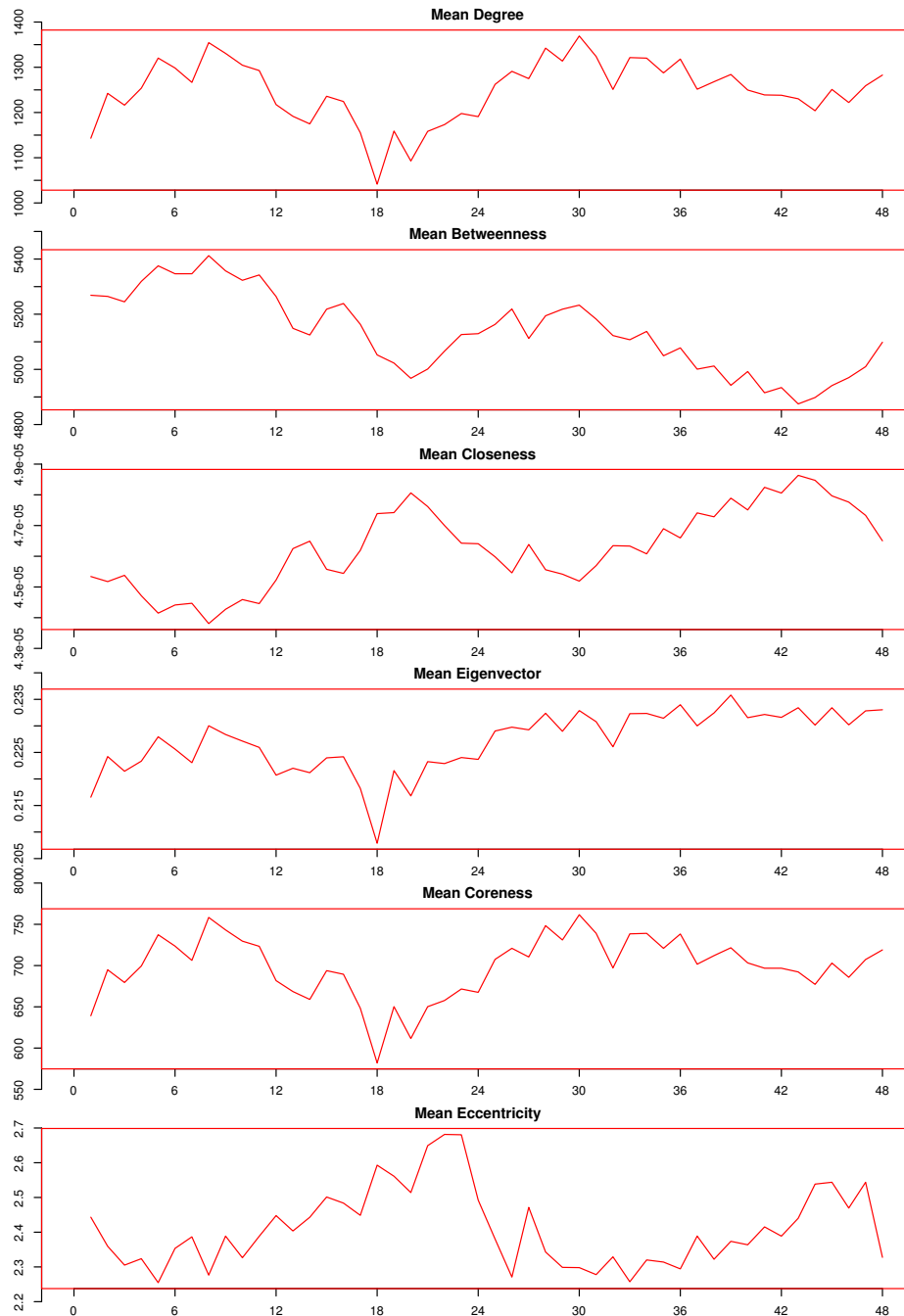


Figure 4.17: Mean of most node-centric measurements over time. The title indicates the y axis and the x axis is the time

Centralization Scores over Time

Closeness score and degree score are closely linked together (Fig. 4.18). This is probably because of the very well connected network having very short average path lengths and also low mean eccentricity which is explainable in the way the graph was constructed connecting AS with the same infohash completely together. It is impressive that a similar curve

behaviour in degree centralization and closeness centralization is not visible in Figure 4.17 for mean degree, betweenness or closeness.

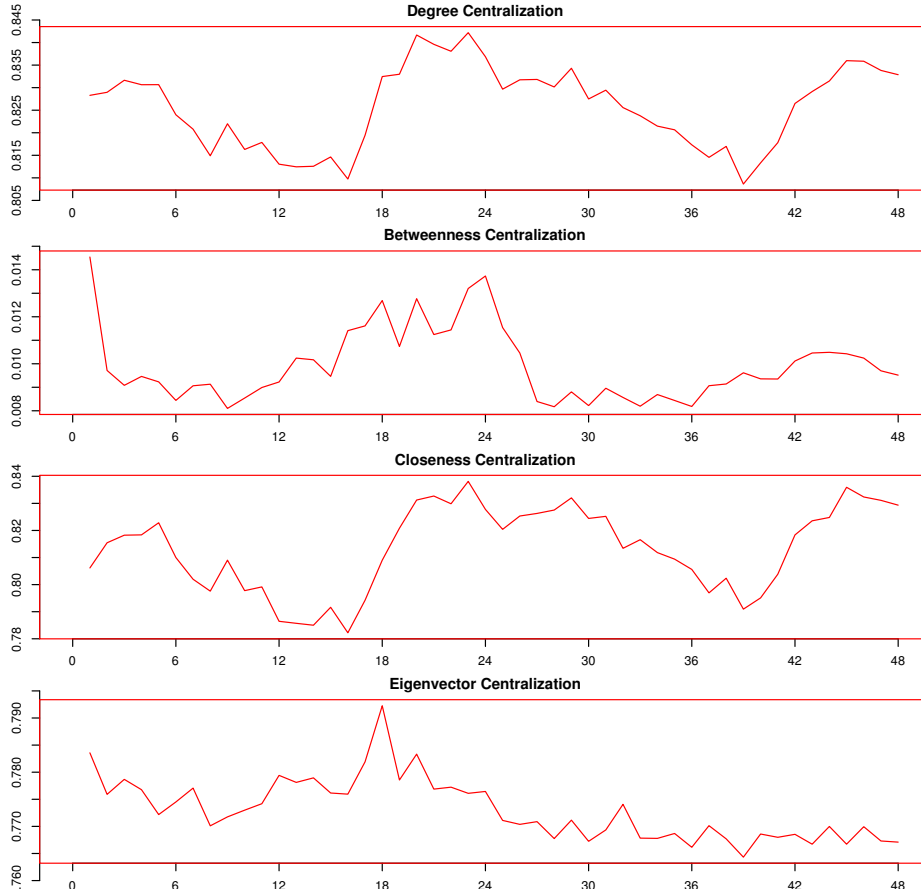


Figure 4.18: Centralization scores over time. The title indicates the y axis and the x axis is the time

Net-centric Measurements over Time

Edge and vertex count have a similar curve. Also this linked to the fact how the graph was constructed. The number of vertices shows has two minima where 24 hours lie in between them, this indicates that at UTC 18 and 42 more people are contributing to BT. Interestingly this is the prime time in eastern Europe and middle east. At this time the network is expanding.

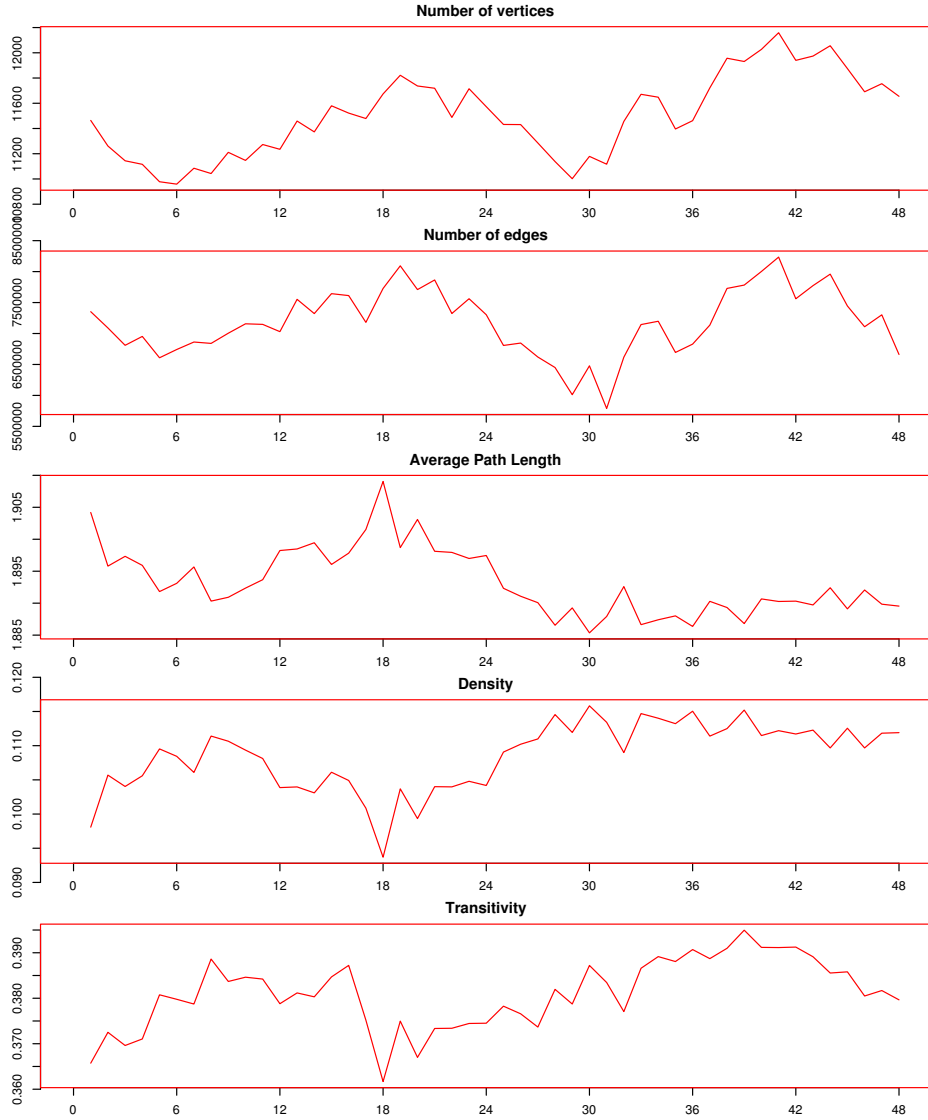


Figure 4.19: Net-centric measurements over time. The title indicates the y axis and the x axis is the time

Transitivity (the global clustering coefficient), density and the average path length have a similar curve behaviour where the average shortest path is horizontally flipped compared to the other two. Transitivity measures the connectivity of neighbours, where density calculates the general connectivity in the graph and average shortest path length calculates the average of all shortest paths in the network. In such wise it makes sense that these three network-centric measurements are tightly linked together while the network gets bigger. At UTC 18 the density and the transitivity have their minimum where average shortest path length has its maximum. So the connectivity in the network gets weaker while the shortest paths have to get longer because of weaker connectivity where now some path have to take a detour to reach their target destination.

4.3.2 Important Nodes

To better understand the network important nodes are looked at. These nodes are in the top ten of the four measured centrality measures degree, closeness, betweenness and eigenvector centrality. The evaluation was made over all time frames counting afterwards the occurrences by each centrality measure. Remarkably, 15 AS take part in all centrality top tens. Table 4.5 lists the AS's ASN, its name and also the country. It is remarkable that so many nodes occur in all rankings but this is again due to the way the network was built. These nodes contain the highest diversity of infohashes CI among the AS and therefore, they are central in the net regarding all the metrics. This also explains the strong correlation of all the centrality metrics among each other.

AS with higher infohash diversity are more important, to reach a higher diversity different languages spoken in one country might help. An important role plays also that English should be understood. So it makes sense that Philippines has an AS in this ranking since many languages are spoken and compared to other Asian countries people speak English but also the legal situation could play a role or the wealth of the country, but another fact of Philippines is that it is highly influenced by America, this also might be important. It is interesting that Austria has an AS in the top ranking. Austria is a relatively small country that speaks German.

Neither number of unique IPs alone nor number of unique infohashes alone lead to a high ranking but also the swarm sizes of the infohashes. This is visible in AS 36351 having only few unique IPs and also not the highest number of infohashes.

Two of these ASs occurring in all the top ranks are also part of the top ranks in every measured hour in the top 10.

The next step would be to identify important infohashes that link all these AS together. This is beyond the scope of this thesis but generally spoken infohashes that have huge swarms and link together other big, important swarms, where the ASs in between two big swarms has a higher centrality.

ASN	Name	Country	countIp	countInfoh.	top d	top b	top c	top e
0	-Reserved -	ZZ	335282	17374		x		
7922	Comct Cable Communications	US	262519	16720	x	x	x	x
5089	NTL Virgin Media Limited	GB	157572	15247	x	x	x	
6830	LGI-UPC Liberty Global Operations B.V.	AT	174196	15174	x	x	x	x
7018	ATT-INTERNET4 - AT&T Services	US	113728	14687	x		x	x
1221	N-TELSTRA Telstra Pty Ltd	AU	175024	14040				
2856	BT-UK- British Telecommunications PLC	GB	100097	13417				
5384	EMIRATES-INTERNET ETC	AE	121083	12679	x		x	x
6327	SHAW - Shaw Communications Inc.	CA	64815	12495				
9299	IPG-AP PLDTC	PH	173582	12304	x		x	
28573	CLARO S.A.	BR	196452	12028	x	x	x	
3352	TELEFONICA DE ESPANA	ES	60043	11795				
36351	SOFTLAYER - SoftLayer Technologies Inc.	US	5493	10849				
2860	NOS COMUNICACOES	PT	53411	9841				
18881	Global Village Telecom	BR	181033	9666				
3243	MEO-RESIDENCIAL MEO	PT	49243	9351				
			138973.3125	12979.1875	mean top			
			85769.09123	2437.433328	sd			
			602.634	255.049	mean all			

Table 4.5: List of all the top nodes occurring in all centrality rankings. Continent wise most of the important nodes belong to Europe with six followed by North America with four followed by Asia and South America with two and finally Oceania with one. An x in "top d" means in the top ten of the degree in all 48 networks, "top b" stands for betweenness, "top c" closeness and "top e" eigenvector centrality. On the bottom also mean of the top ranks with cooresponding standard deviation and global mean is noted.

4.3.3 Animation

To visualize the appearance of the network over time an animation was created. The animation was created using a subset of all possible nodes and edges because huge networks are less readable than smaller ones. A snapshot of the animation where the size of the node represents the degree and the colour the continent of the AS is shown in Figure 4.20 . The code to construct the animation was based on the approach described in this Esteban Moro's R-blog[36]. An early approach to make two different animations was not successfully since in the network with weights higher than 2000 there was not much movement. The animation is very jiggly this is probably because of the many connections each node has already from the beginning. Most of the shown nodes do not change much during the animation this is also coherent with the finding that 30 percent of the AS in the network stay in the network the whole time, which is probably even higher considering that only edges with a high weight were selected. And as shown in the previous section some of the top nodes are holding their network position during the whole two days.

[Link to the animation](#)

4.3.4 Q3: Conclusion

The comparison of the line diagrams of the different SNA metrics showed that some metrics have exactly the same line behaviour and others have the exact same but flipped horizontally. In the 18th time frame most of the averaged metrics but also some of the network centric measurements and eigenvector centralization have either a minimum or maximum. Interestingly the peak is only visible in eigenvector centrality where in the other centralization scores a transition is visible from low to high. The vertex and edge count have at this point a maximum which is again reached 24 hours later. The edge

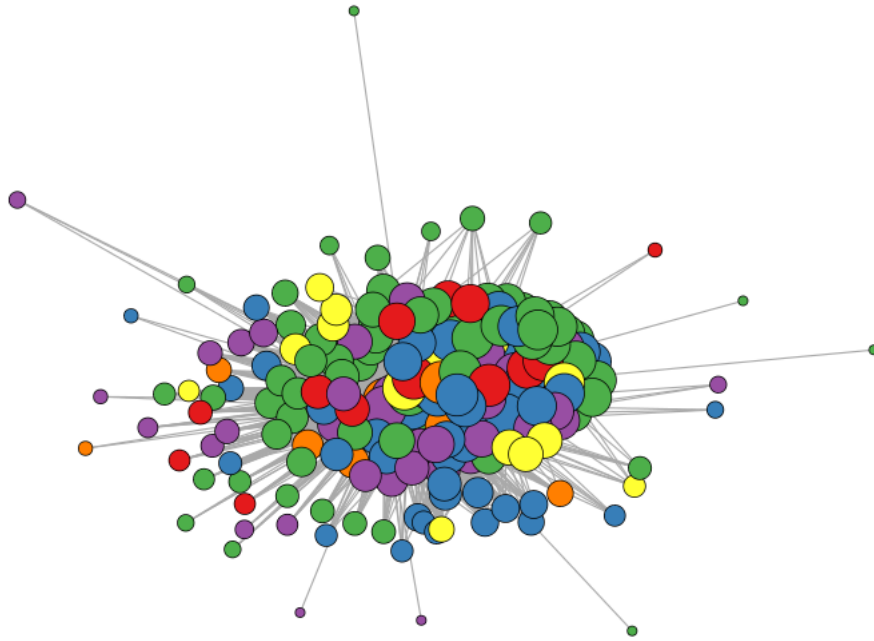


Figure 4.20: Example image from the animation of the network only containing edges with weights bigger than 1000

and vertex wise peak is reached around the prime time of Europe and middle east, which might indicate a higher influence of Europe and parts of Asia as expected. The network is less dense and has therefore a longer average path length and lower transitivity. But from such analyses it is hard to conclude changes in the network or for the real world because this analysis is from a bird's-eye view but the influence for an AS or even an IP is more or less guesswork. A problem might also be that only two days were measured. Interestingly two of the top nodes ranking in the top ten of the four centrality measures were in the ranking the whole time. This is also visible in the animation showing that there is a stable core not changing.

4.4 Q4: Effects of Data Filtering

Filtered data with unfiltered data is compared in different sets and checked how the filtering influences the network. The hypothesis is that seed boxes do not have a very big influence on the network. In all the datasets the VIOLA data crawling was filtered away. The different metrics are visualized in Figures 4.21 and 4.22.

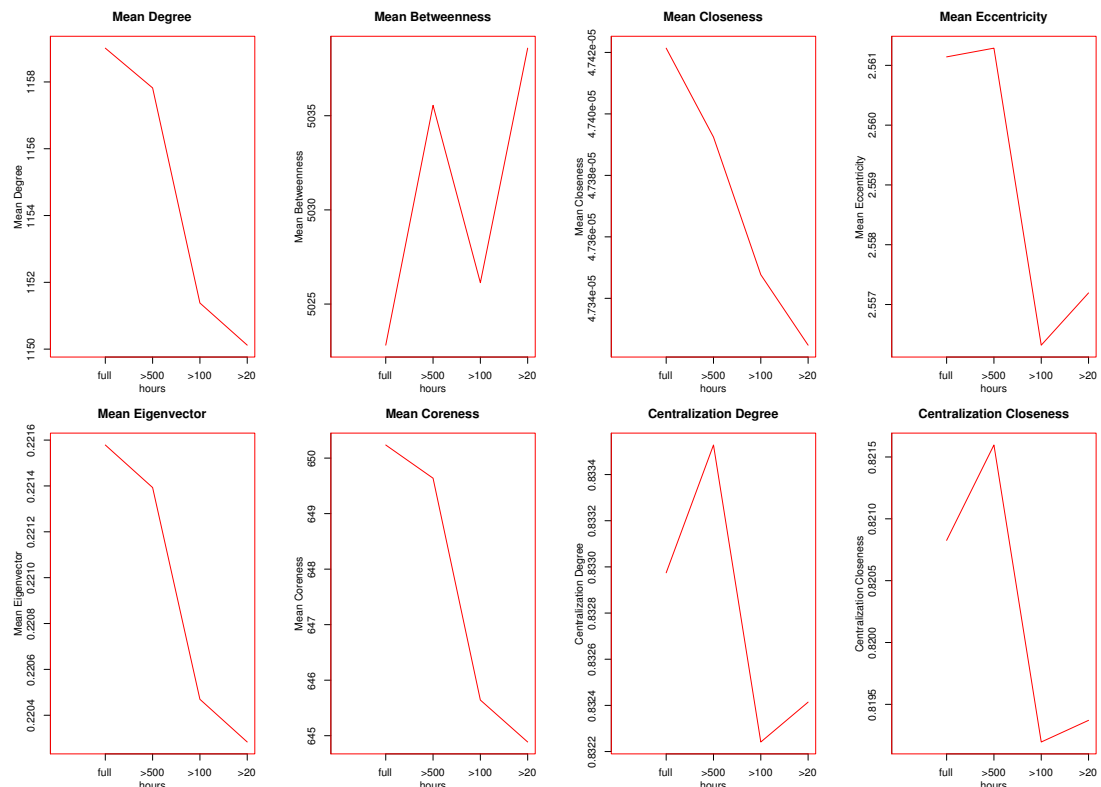


Figure 4.21: Influence of the filtering on the dataset 5 am 14th March 2016, first part of metrics

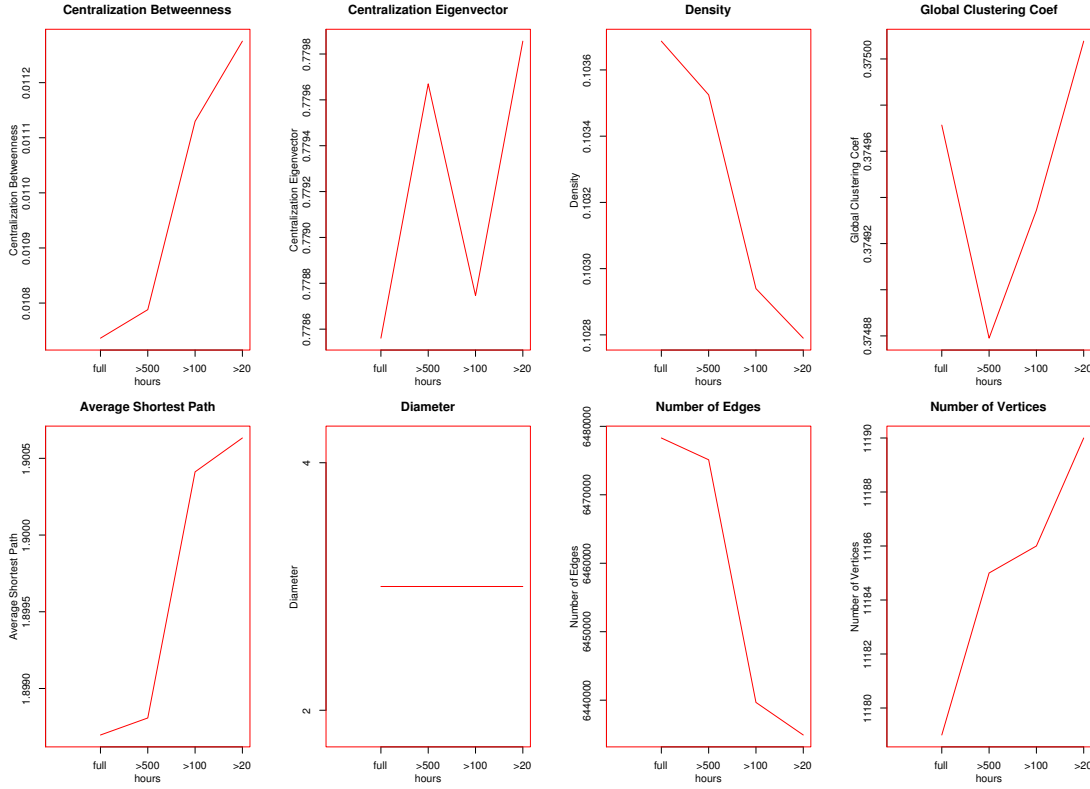


Figure 4.22: Influence of the filtering on the dataset 5 am 14th March 2016, second part of metrics

The mean degree, mean closeness, mean eigenvector, mean density and number of edges show a similar curve being highest at the beginning and then decreasing gradually. This makes sense because these metrics are generally influenced by the number of edges which were gradually reduced in numbers to the end.

Similar curves but the other way round are shown in average shortest path and number of vertices. The shortest path is increasing with the lower number of edges. Not explainable is that the number of vertices is increasing by eight if the data gets filtered. Since this line is strictly monotonic, it is assumed that there is no error in the computation of the filtering.

Interestingly, mean betweenness and eigenvector centrality show a similar curve being low, going up very high, going steeply down again and going up to the maximum in the last dataset. Having a local maximum at the filtering of above 500 (>500).

Degree and closeness centralization and mean eccentricity also have a similar curve where the global maximum of the dataset is at >500 and the global minimum at >100.

4.4.1 Effect on the top nodes

To check whether the filtering would also have an effect on the top nodes. The filtered dataset got checked for AS being in all the top 10 rankings. Following ASs were part of

all of them 7922, 7018, 6830, 9299, 28573, 5089 and 5384. These ASs were also found in the top nodes section. In this way these nodes do not have a top rank because of similar system like VIOLA but because of real usage.

4.4.2 Conclusion

There are differences in the measurement of each datasets but the influence of the filtering is on a very small scale. This analysis was similar to the analysis over time done from an aerial perspective. It is not plausible that some special network feature got dramatically changed because of the filtering of some, compared to all others, few IPs. It has not a big impact on the network since this network is not centred on IPs but on ASs. The filtering also had no measurable effect on the identified top nodes these ASs were also found in the set of the strongest filtering. Considering this the datasets were not filtered because of the mentioned reduced influence on the AS network constructed in this thesis.

For further analysis it would be interesting if cluster building algorithms form more homogeneous clusters the more the network is filtered.

4.5 Q5: Analysis of Autonomous System Behaviour over Time

After watching the network from far distance some AS in different countries will be further analysed for patterns like peaks and lows on specific times. The assumption is that people are more likely to be online during the evening when they are at home. But it is questionable if it is possible to see such activity only on the AS level. In the examined networks of the 48 hours exist 19,196 unique ASN where 5,990 of these ASs are part of all the 48 networks. So about 31% of the AS are always part of the network activities measured in this study.

An attempt to visualize this possible behaviour not on a network level but on the data level is shown in Figure 4.23. This figure shows the number of all the ASs for each country in the data. The effect of highs and lows in regular 24 hour intervals is visible. The two blue lines and the dashed red line have a minimum at similar positions in time 6 am UTC and between 24 and 30 where the two highest lines dash-dotted red and orange have a minimum around 12 and one between 24 and 30. So event to number of AS shows the activity during the day with peaks and lows.

The measured time frame of this Thesis is from Sunday morning 12 pm to Tuesday morning 12 pm UTC. Figure 4.24 shows some SNA metrics over the evaluated period for some AS selected for Switzerland and Figure 4.25 shows the same information but about Costa Rica as another example to prove that user patterns exist among the metrics.

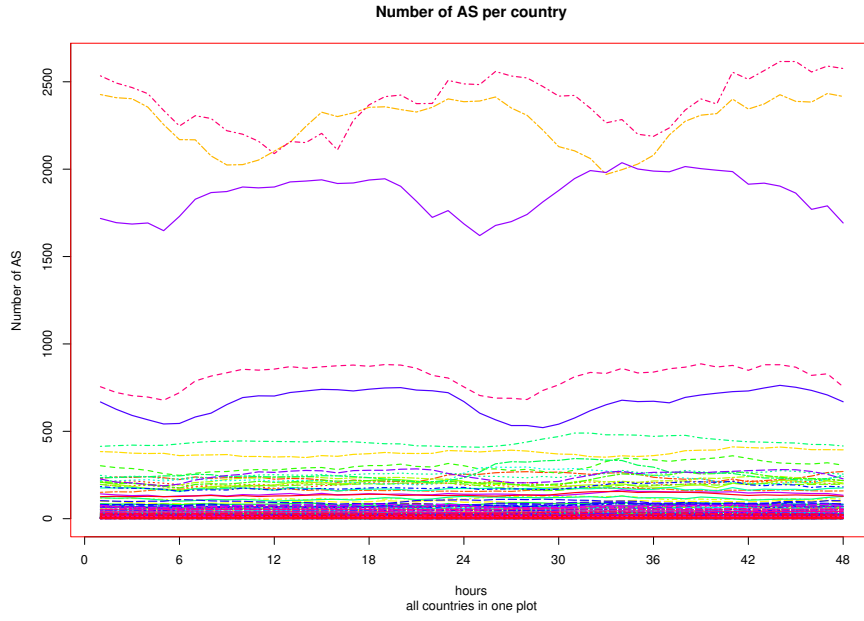


Figure 4.23: Number of ASs per country per time frame this is only an overview

4.5.1 Switzerland

The difference between the labelled UTC and the Swiss local time is UTC plus two hours. So the visible time frame in Swiss local time is from 2 am 13th March till 2am 15th March.

The curve of the degree of the Swisscom AS has a local minimum on the Sunday morning at 6 am local Swiss time. Then the degree is gradually increasing till the peak of about 6 pm then again decreasing till 6 am on the 14th Monday morning having a spike at 8 am also visible in the dataset of green. The dataset has twice a minimum at 6 am so during the day more and more people start the BT client where after the prime time, the time when the most important TV programs are on, more and more people start to shut down the client. A steep curve is visible at 2 am where many people might go to bed. Then on Monday morning at 8 am again a spike indicating that probably many people start their computer to check e-mails and BT clients usually have an auto start set up.

The degree of the Switch AS has an interesting curve since it is known that the first half of the curve is on the Sunday where not much activity is happening, this is because most students are at home and not at university. Then from 8 am on the curve is increasing rapidly till 2 pm where the curve is raising again till 5 pm when probably most of the students are about to go home which is indicated in the decaying degree.

The degree curve of other companies AS is in existent on the Sunday because no one is in the office but then it shows some small activities during the office time which is ending after the office hours because the people left the AS switching to their home ISP.

Most of the metrics do have a similar curve behaviour where some regions might be stretched and others might be squeezed. Betweenness shows an amount of shortest path crossing a corresponding node where Swisscom is on many more shortest paths than

all others. The curve is comparable to the degree curve but it scales with the starting betweenness, so an AS with an already high betweenness gets higher betweenness the higher the degree gets. Closeness also looks similar like degree but starting in the middle of the range. Eigenvector squeezes high values where low values get stretched but also here the curve is generally similar to the degree curve. Coreness shows an upper bound which is the maximum number of cores. Switch has a high variance of the k-cores it belongs to. Eccentricity might only be either two or three only the two telecommunication ASs have all the time an eccentricity of two.

4.5.2 Costa Rica

To prove that also in other countries the same patterns are visible Costa Rica was selected because of its similar country size compared to Switzerland and relatively big time difference. The time difference from the labelled UTC in the diagram to the local Costa Rican time is minus six hours. The dataset of Cable Tica with the red line shows the typical usage pattern of a telecommunication company similar to Swisscom being high from 6 pm to 10 pm with a declining curve until 2 am where it slowly starts to rise again. The American Data Networks degree has its minimum also around the same time but the curve is less stable having highs and lows on short intervals. In the Radiographica Costaricense dataset one minimum is at a similar position (6 am, label 36) compared to the other two ASs, while from 0-33 UTC the line is going up and down on short intervals. Also in this data set nodes with already mediocre betweenness get amplified by a higher degree to reach very high betweenness. The lines of closeness in this example are much closer together then compared with the Swiss dataset and Swisscom.

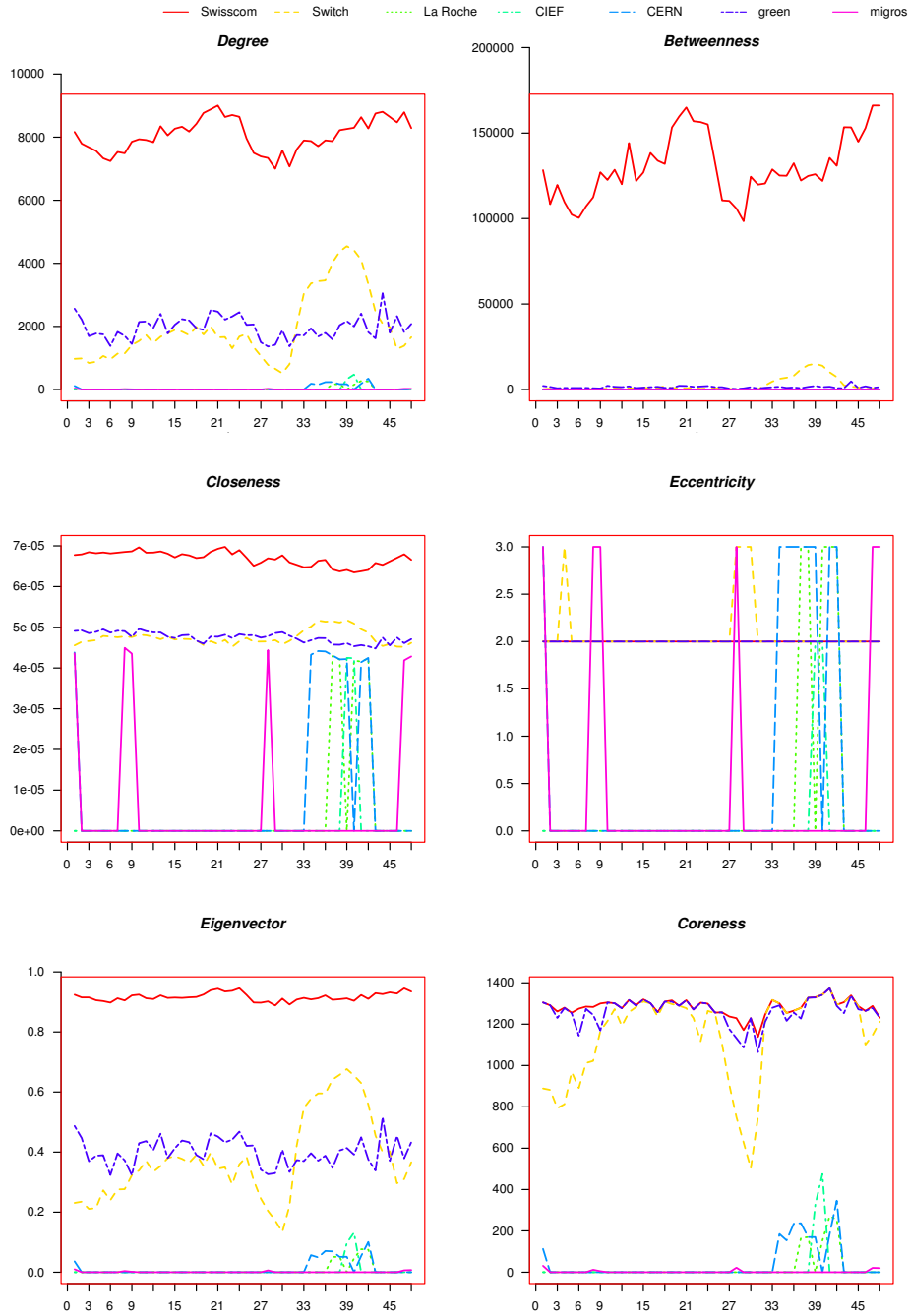


Figure 4.24: SNA metrics over the measured period from AS located in switzerland, where green (AS1836) and swisscom (AS3303) are telecommunication companies, switch (AS559) is a network service provider mostly specialised on univervities and some companies like Migros(AS8803) a retailer, La Roche a pharma company (AS2047), the famous CERN (AS513) and gouvernemental Centre Informatique Etat de Fribourg as CIEF (AS25021)

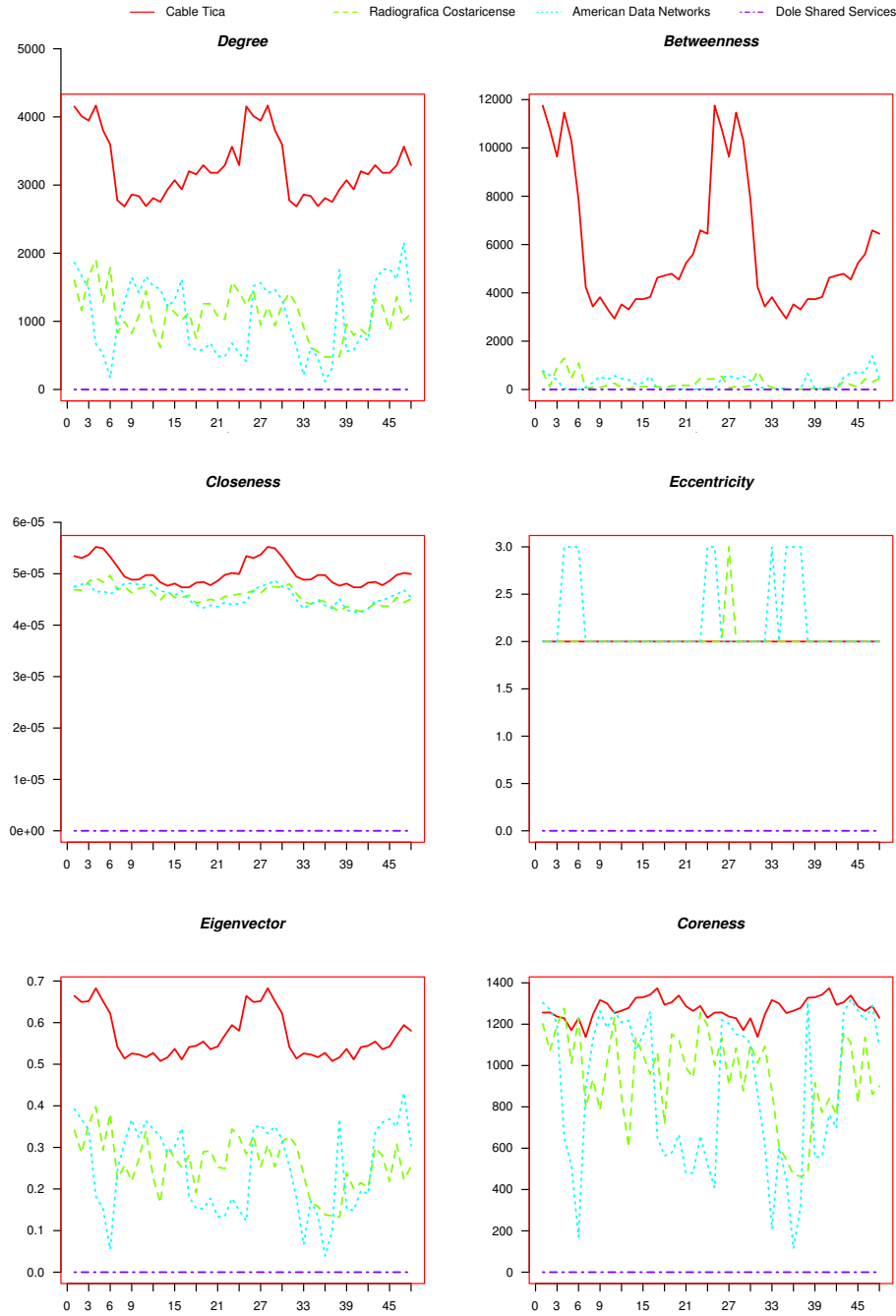


Figure 4.25: SNA metrics over the measured period from AS located in costa rica with datasets of Cable Tica (AS52228), Radiografica Costaricense (AS3790), American Data Networks (AS27876) and Dole Shared Services (AS61479). Dole was added to scale the diagrams similar to the swiss ones because the dataset only contains zeros.

4.5.3 Q5: Conclusion

The examined few datasets of ASs from Switzerland showed that there are plausible explanations for the curve behaviour. Degree showed a general good first basis for the

analysis of the user pattern. But this part needs further analysis over a longer period of time. The example with Swisscom a usual telecommunication company and switch a telecommunication service provider for universities and also the companies showed that the download behaviour is completely dependent from the users of the AS and the more homogeneous the users of one AS are the easier it is to derive specific patterns. The evaluation of these patterns could be used to categorize the different types of ASs like universities, telecommunication company or any other company. Similar to [23] it might also be possible to evaluate the time zone of an AS according to the studied usage pattern. To check if such behaviour would also be visible in other countries also some ASs from Costa Rica were analysed indicating similar patterns but shifted in time.

4.6 Q6: Cluster Analysis

Focusing on the answer the Research question Q6, two cluster building algorithms will be used and the result will be analysed in the first part. While the second part looks at the assortative mixing of the country or continent of the nodes.

4.6.1 Cluster Building

The results of the two selected algorithms for cluster building will be examined and will assess whether the cluster division makes sense or not. The fast greedy algorithm returned for every time frame two or three big main communities having about the same size. The walk trap algorithm returned mostly three once two bigger communities while also returning several communities of size 1, while the big clusters also had huge differences in the number of AS.

The analysis of clusters over time is impossible because the algorithm might return different communities. Figure 4.26 shows the modularity value over time. While the fast greedy approach had a mean modularity of 0.1896611 and the walk trap approach had a mean of 0.1659124. The result of both of the algorithms is relatively low. It might well be that an other algorithm with fine-tuned input parameter could return a better result.

Figure 4.27 shows the different AS of the network from the first measured hour. Europe in green is divided among all the clusters while South America belongs the biggest part to the first cluster and also Africa has a bigger part in the second cluster. Figure 4.28 shows the same time frame but using the walk trap algorithm. Since the fast greedy algorithm showed in general better results, also having a higher modularity, and returning bigger groups of clusters further analysis will be done with the result of the fast greedy algorithm. A possible reason why these algorithms performed not that well is because how the graph is constructed. The network can be understood as extremely well connected, therefore the network has a very low shortest paths and also the eccentricity is generally very low between two and three. Considering this it is reasonable that the modularity, a measure that compares the inter cluster connectivity with the connections leaving the cluster, is low in case of well connected nodes. An interesting fact is that the difference in modularity

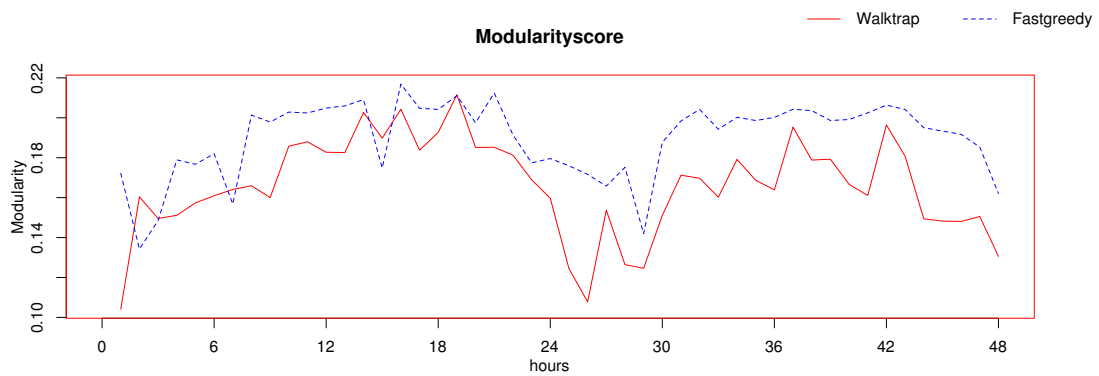


Figure 4.26: Modularity of the two different algorithms over time. Where fast greedy shows better performance in terms of modularity then walktrap.

between the fast greedy and the walktrap algorithm is not that big even though walktrap has not a uniform splitting approach having few clusters with huge differences in size and even then a comparable modularity to fast greedy is reached.

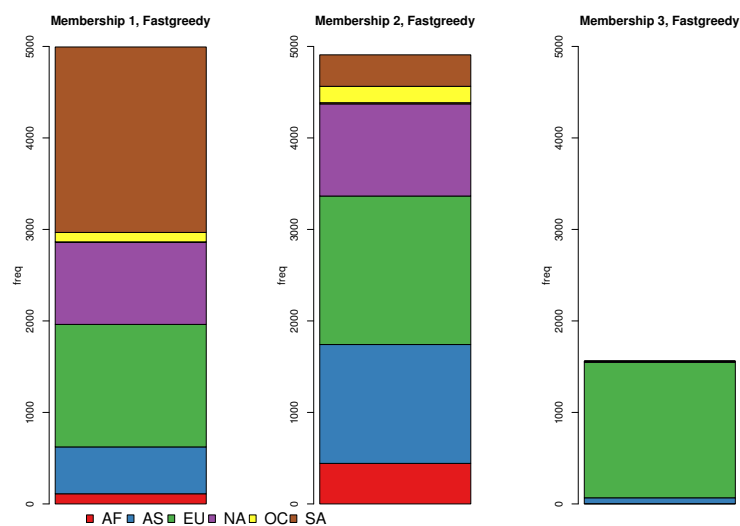


Figure 4.27: This figure shows the division from the fast greedy algorithm of the AS in the different cluster groups. Membership means the cluster the AS was assigned. Showing the continent of the AS and its frequency. Clearly visible the algorithm did not divide the cluster continent wise.

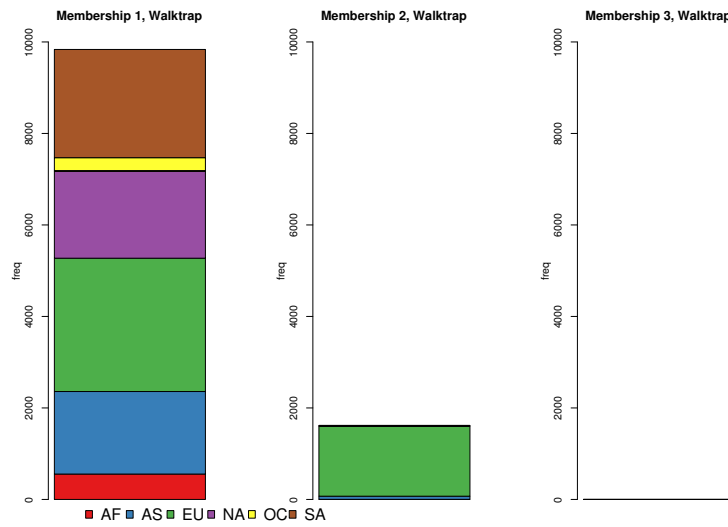


Figure 4.28: This figure shows the division from the walk trap algorithm of the AS in the different cluster groups. Membership means the cluster the AS was assigned. Showing the continent of the AS and its frequency.

To further analyse the clusters only the ASs of Europe will be examined. Figure 4.29 shows the number of ASs belonging to Europe and to which cluster they belong. Figure 4.30 shows the normalized number of the European ASs.

The cluster three in Figure 4.30 has the best match of some kind of similarity of countries. Russia, White Russia Moldavia, Latvia and Lithuania all belonged once to the USSR and they speak a Slavic language or even Russian. With the other two clusters there is no clear pattern visible. Interestingly many very small countries belong to cluster two like, Andorra, Ilse of Man, Luxembourg etc. For the rest most of the big countries are split among the two clusters while cluster two has generally a larger part of it.

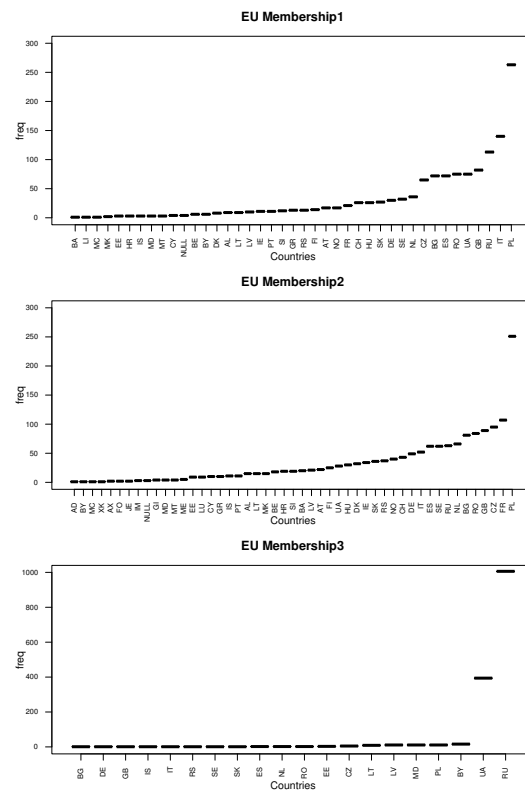


Figure 4.29: AS from Europe belonging to the different clusters by frequency.

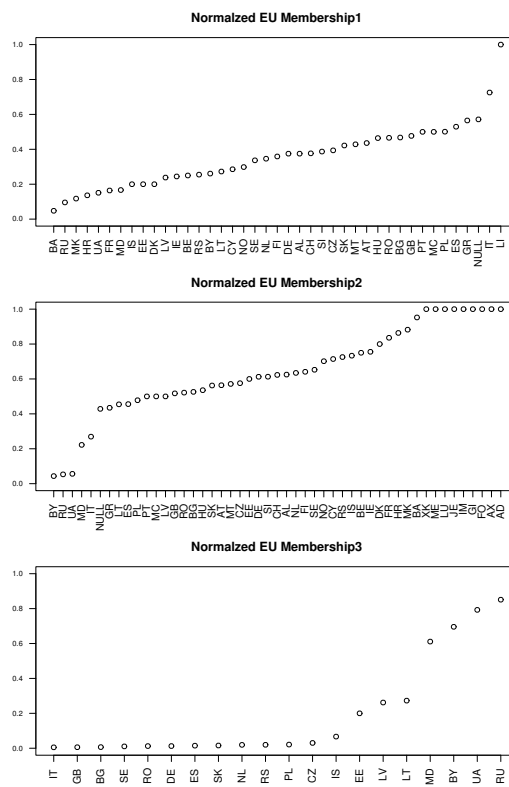


Figure 4.30: AS from europa belonging to the different clusters by normalized frequency.

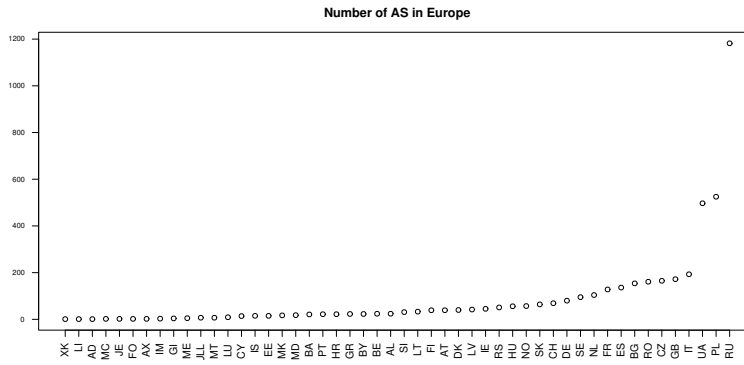


Figure 4.31: AS from europa belonging to the different clusters by frequency.

In Figure 4.31 the distribution of the AS per European country is illustrated. Uncommonly, the Figure shows clearly that Russia, Poland and Ukraine have a much higher ASs count. For Russia this is explainable because all AS belonging to Russia were added to Europe, while probably about the half of them would naturally belong to Asia. But generally it is astonishing that three of the countries in the count of AS ranking are from a similar region. Besides interesting is that at the measured time frame it was 3 am in Moscow and 2 am in Warsaw. Which is for sure not the prime time in this region. Maybe BT is more important in this region of the world than in Europe. Another hypothesis is that in this region less big telecommunication service provider exist but many different smaller ones or another possibility could be that it is more common as a company to have their own AS. And one last possibility might be that the VIOLA data collection system focused on this part of the world collecting data while leaving out some other parts. These hypotheses would need further analysis and also comparison with all other continents and their AS distribution of the countries per cluster.

4.6.2 Assortativity

The assortativity coefficient for each of the 48 hours is shown in Figure 4.32. A very similar curve is visible in the assortativity coefficient for country and continent and also the degree assortativity diagram shows similar behaviour having local maxima at corresponding positions like the other two.

The assortativity of the country lies between 0.07094 and 0.09123 and the assortativity of the continent between 0.112 and 0.1412. So in case of the attribute country and continent of the AS the results shows that there is a higher likelihood of connecting to ASs with the same continent or country than linking with any other node not having the same country or continent. Where the chances of linking with the same continent are higher than linking with the same country which generally makes sense because there are more ASs having the same continent than ASs having the same country.

The degree assortativity shows dissasortative behaviour with a very low negative coefficient from -0.38 to -0.4002. This means that high degree nodes tend to link together with low degree nodes. As an example the mean of a sufficient number of evaluations of

the degree assortativity of a Erdős-Rényi random graph should be zero [18]. In this way the degree disassortative behaviour would speak for the scale-freeness of the graphs this was also concluded for graphs with similar features in the paper of Fricke et al [10]. The assortative degree mixing of the constructed scale-free random graph at the beginning of the thesis in Figure 4.12 shows also a low disassortative coefficient of -0.1013899.

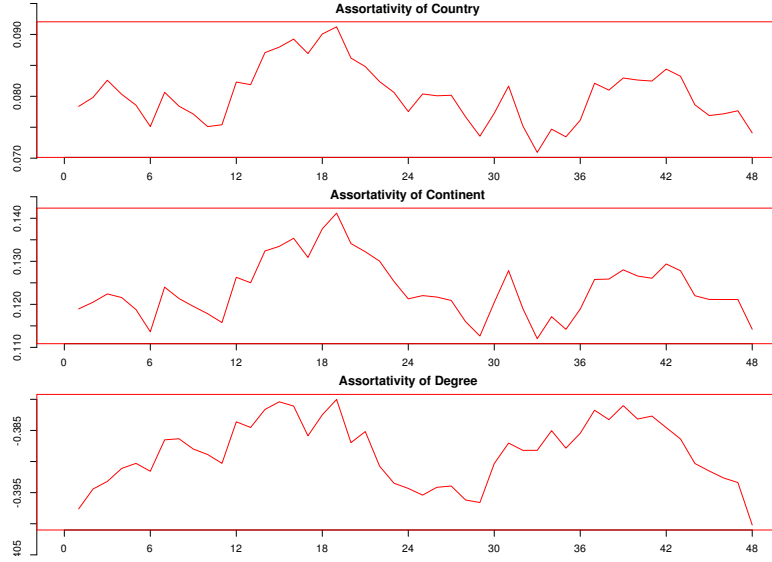


Figure 4.32: Assortativity of country, continent and degree.

In the end the assortative mixing of CIP, the count of distinct IPs per AS from the raw data, CI, the count of distinct infohashes per AS from the raw data, and NIP, the maximum number of possible IPs per AS, were checked for one time frame. The highest negative assortative mixing was encountered for CI with an assortativity coefficient of -0.1641209 this means ASs with lower diversity of infohashes tend to link to ASs with higher diversity. This is plausible because with more infohashes the chances are bigger that the two ASs have some files in common which results in a link. The value for CIP is -0.05412892, which means that ASs with less IPs tend to link to ASs with more different IPs but this is on a smaller scale compared to CI considering many IPs do not guarantee high diversity of infohashes and for NIP the effect is even weaker with -0.02233759. These results are very similar to the ones found in the correlations section. Such analysis would also be possible for the collected data from the World DataBank[45] but since the correlation in this region was rather weak the assortativity would probably also be rather weak.

4.6.3 Q6: Conclusion

In the first part of this section the two cluster building algorithms fast greedy and walk trap were studied and also applied on the first measured hour to check the influence on the country continent division of these algorithms. Fast greedy returned better results, therefore it was used to analyse the distribution of the AS in Europe among the three

clusters. While one cluster had similar countries of former USSR the other two clusters were quite mixed each containing half of most of the big European countries. So possible hypotheses like cluster would contain one region or one language cloud not be proven.

The second part of this section discussed the assortative measurement which means the linking to similar nodes. The degree assortativity is negative this indicates again that the graph is scale-free because it is typical for scale-free graphs that high degree nodes link together with low degree nodes. But the assortativity of country and the one of continent are positive where country is lower than continent. So the chances of linking two nodes with same country or same continent are generally higher than links to a randomly selected node, which would be indicated by assortativity of zero.

To answer the question **Q6:Are strongly connected clusters grouped by the country or continent of the Autonomous System** the answer is that one example was found in Europe where it actually was the truth one cluster containing most of a region and also smaller countries belonged in the examined case usually fully to one cluster. But as already stated is could not be proven and should be further analysed.

Despite of the clustering the country and continent does play an important role in the linking between nodes where same country and same continent have a higher linking chance than same degree to same degree linking.

Chapter 5

Summary and Conclusions

A short summary of what was accomplished in this thesis is stated and then the most important conclusions from Chapter four are drawn again to answer the research questions. In the end a final statement with possible future research is given.

5.1 Summary

The second Chapter looks at important topics for the thesis then discusses related literature ordered by their topics. The third Chapter starts with the research questions at the beginning which should be answered till the end of the thesis. Then the construction of the graph is discussed and also a short overview over the data processing work is given. Then one by one the important methodologies and concepts that are used to answer the research questions are discoursed. The introduced topics are correlation, topology analysis, the network over time and cluster building. The fourth Chapter shows the results of the analysis and calculations that were chosen in Chapter three.

5.2 Conclusion

Every research question will be listed and shortly answered according to the results from Chapter four.

Q1: Are there correlations in the node centric measurements, with geographical or economical aspects, e.g., country or AS?

Yes, there are correlations among the various types of variables. The region with the strongest correlation was where the node centric measurements were correlated with themselves. This is also nothing new since [27] came as well to this result investigating in several

studies with 58 smaller networks. The real-world data from [45] showed very weak correlation but there are several reasons that might result in such low correlation. First the data was unrelated in means of time since the calculated network was about one hour, and the data collected from [45] was about one year and measured in 2014. The second reason why the data showed low correlation was because the data was not directly about the corresponding AS but about the country of the AS. This approach was exploratory research to see what is doable. The data queried from the VIOLA data bank showed much better correlation which is also reasonable because the collected data also influenced the construction of the network. The variables with the strongest correlation were weighted degree followed by eigenvector centrality.

For future research in the field of correlation more information about real-world data AS would be appreciated.

Q2: What topological features has the constructed Network?

The constructed graph is analysed using an ER random graph which is a usual method to evaluate whether or not a graph has small-world features or not. The two features having a higher local clustering coefficient and having an equal or lower average shortest path length were met over the whole measured time. The second step was to assess the scale-free properties. Scale-free properties are mostly a power-law distribution. Which indicates that there are much more nodes having a low degree while only some have a high degree. R^2 shows a good fit, while the curve is not as straight as it is expected from a power-law distribution. The comparison with a scale-free random graph reveals a similar curve. This is also explainable by the way the graph was built making it almost impossible to generate nodes with low degree of 1. Then a comparison of the correlation plot comparing the three different models with the real-world data is done. The correlation plot of the real-world data resembled the most to the scale-free random graph but this approach needs further refinement but would be an interesting field of further analysis. The dissasortative mixing measured in another part of the work indicates also scale-freeness.

Q3: How does the data change over time?

To evaluate this question, different types of metrics were watched over time, while some showed interesting behaviour like having exactly the same curves or curves flipped horizontally. An example: if transitivity and density are decreasing the average path and the eigenvector centrality are increasing. These analyses were done from an aerial perspective and it is rather hard to see the influence from the mean value behaviour on the real network or on actual nodes. To better understand this shortcoming an animation was created to visualize the network. This animation also displays that many nodes are always part of the network, building a strong core. This is also visible in the top ranking of the four centrality measures. Most of the top ten nodes in all of the four evaluated centrality measures were in all rankings, while some of the nodes were in all rankings on every time frame. Which is very special since these centrality are measuring very different

concepts and are meant to categorize different groups of nodes. But in this constructed network most of the measurements are tightly coupled together.

Q4: Is a filtering of the VIOLA-data needed, do seed boxes change the network in a radical ways?

The filtering had generally no big effect on the net or the metrics. This is plausible because the filtering is based on the IP level and not on the AS. Interestingly the previously mentioned top nodes ranking high in the top ten of the centrality measures take also part in the dataset which was filtered from IPs holding more than 20 different infohashes. But also here further analysis would be interesting to check if there is an influence on the clustering behaviour, if a filtered set would return "better" clusters, that are better understandable or easier to pigeon-hole because of language, region or same interest of the cluster.

Q5: Are usage patterns visible in the network data?

To analyse this questions some ASs were selected from Switzerland and from Costa Rica. Important for the AS selection was the diversity of the AS. AS might be divided into different categories, like companies, telecommunication providers or universities. Each of which has its own pattern. The analysed data is from a Sunday and a Monday, the data expressed that there is no activity during the weekend, for universities but on Monday the metrics increased immensely, or telecommunication provider having highs and lows every 24 hours while slowly increasing, the decay is steeper. So it is suggested that the time zones from the AS could possibly be derived from the pattern or also the type of AS should be derivable from such behaviour like university, telecommunication company or normal company.

Q6: Are strongly connected clusters grouped by the country or continent of the Autonomous System

To asses this question two approaches were followed. The first approach was using two different algorithms to split the network into communities. Then the communities were compared en behalf of their continent, the modularity measure and also the fact that the walktrap algorithm returned mostly one big, one very small and some clusters only containing one AS the fast greedy algorithm was selected for further analysis. While the analysis was reduced to the three clusters of countries only belonging to Europe. In one case the community contained mostly countries from the former USSR and the other two clusters contained each half of most of the bigger countries. None of the continents was completely in one cluster and also only a few very small countries in Europe belonged completely to one cluster but interestingly they all belonged to the same one.

The second approach was to test the network for assortative mixing. Which indicates if same nodes according to some node attribute are linking together. So the assortativity of degree, CIP, CI, NIP, country and continent was evaluated. The degree assortativity is negative this indicates that high degree nodes have a higher potential to link with lower degree nodes.

In the group of assortativities with the raw VIOLA data the highest negative coefficient was measured in CI while the other two were also negative but closer to zero, so this points out that AS with low diversity, in terms of different infohashes, are linking together with diverse ASs. This is reasonable because chances are higher that an AS with high diversity also has files in common as the AS with only a few infohashes.

The country and continent assortative mixing is positive so chances that nodes link with same country or same continent is higher, the continent assortativity was a bit higher then the country but this is also plausible because there are much more nodes having the same continent as nodes having the same country.

5.3 Final Statement and Future Research

This thesis looked at some possibilities to use SNA to get more information from the raw VIOLA data. Most of the findings in this thesis are linked with the fact how the graph is built. Not showing real data exchange but more possible links of data exchange. A next step could be to generate a network while only linking some random selected nodes similar how the BT protocol works, returning only about 35 peers when a file is requested, and constructing a graph that might have similar features compared the real-world data exchange.

While not answering all research questions completely some new research fields were explored. Correlation studies were done extensively but leaving out comparing network centric measurements correlated with correlation coefficients, e.g. "Is the correlation of weighted degree and betweenness linked to the average path length of the network?". Another new topic found was the influence of the correlation of node centric measurement on the model of the network. Temporal network analysis[12] or also called dynamic network analysis is a relatively new topic examining networks over time. Such research was done but is yet still not very explainable and needs further analysis. The effect of data filtering was tested but a possible influence on cluster building is still missing and could return better results then the constructed clusters with the unfiltered data. In this work usage pattern in ASs could easily be distinguished this could be used to either evaluate the time zones or categorize ASs as universities, telecommunication companies, or any other companies holding an AS. Cluster analysis over time would be an interesting research field since it give the opportunity to investigate only part of a network. Unfortunately, cluster building algorithms tend to return different results making it hard to follow one cluster over time.

Bibliography

- [1] Abbasi and Altmann: *On the Correlation between Research Performance and Social Network Analysis Measures Applied to Research Collaboration Networks*, 44th Hawaii International Conference on System Sciences (HICSS 11), Hawaii, USA, January 2011, pp. 1-10.
- [2] Barabási, Réka: *Emergence of Scaling in Random Networks.*, Science, Volume 286, No 5439, 1999, pp. 509–512
- [3] Burger, Hock, Scholtes, Hoffeld, Garcia, Seufert: *Social Network Analysis in the Enterprise: Challenges and Opportunities*. Socioinformatics-The Social Impact of Interactions between Humans and IT. Springer International Publishing, 2014. 95-120.
- [4] Clauset, Newman, Moore: *Finding community structure in very large networks* Phys. Rev. E 70, 066111 (2004)
- [5] Cohen: *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum Associates. 1988
- [6] Cohen: *Incentives Build Robustness in BitTorrent*, Workshop on Economics of Peer-to-Peer systems (Vol. 6), pp. 68-72, May 2003.
- [7] Erdős-Rényi: *On Random Graphs. I* Publicationes Mathematicae. 6: 290-297, 1959
- [8] Field: *Discovering Statistics Using SPSS* London, England : SAGE. 2005, 2nd Edition
- [9] Freeman: *The development of social network analysis: A Study in the Sociology of Science*. Empirical Press, 2004.
- [10] Fricke, Finger, Lux: *On Assortative and Disassortative Mixing in Scale-Free Networks: The Case of Interbank Credit Networks* Kiel Working Paper No. 1830, February 2013.
- [11] Hauke, Kossowski: *Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data* Quaestiones Geographicae 30(2), 2011
- [12] Holme, Saramäki *Temporal networks* Physics Reports, Volume 519, Issue 3, p. 97-125, October 2012.

- [13] Jiang, Nariai, Steffen, Kasif, Kolaczyk: *Integration of relational and hierarchical network information for protein function prediction*. BMC Bio Med Central Bioinform. 9, 350 2008.
- [14] Kolaczyk, Csárdi: *Statistical Analysis of Network Data with R* Springer, 2014
- [15] Lareida, Schrepfer, Bocek, Stiller: *Overlay Network Measurements with Distribution Evolution and Geographical Visualization*, 2016 IEEE/IFIP Network Operations and Management Symposium (NOMS 16), Istanbul,Turkey, April 2016, pp. 1-9.
- [16] Milgram: *The small-world problem* Psychol. Today 2, 60 1967.
- [17] Mislove, Marcon, Gummadi, Druschel, Bhattacharjee: *Measurement and Analysis of Online Social Networks*, 7th ACM SIGCOMM conference on Internet measurement Pages 29-42 (IMC 07) San Diego, California, USA, April 2016, pp. 29-42.
- [18] Newman: *Assortative mixing in networks* Phys. Rev. Lett. 89, 208701, May 2002
- [19] Otte, Rousseau: *Social network analysis: a powerful strategy, also for the information sciences.*, Journal of information Science 28 (No. 6), 2002, pp.
- [20] Pinthong, Lilakiatsakun: *Performance of bittorrent-like p2p file sharing systems inspired by ALTO*. TENCON 2013 - 2013 IEEE Region 10 Conference (31194), Xi'an, China, Oct 2013, pp. 1-5.
- [21] Pons, Latapy: *Computing communities in large networks using random walks (long version)*., eprint arXiv:physics/0512106; 12/2005.
- [22] Putnik, Costa , Alves, Castro, Varela, Shah: *Analysing the correlation between social network analysis measures and performance of students in social network-based engineering education* International Journal of Technology and Design Education, Volume 26, Issue 3, pp 413-437, August 2016.
- [23] Rossi, Branch, Armitage: *Inferring the Time-zones of Prefixes and Autonomous Systems by Monitoring Game Server Discovery Traffic*, 21st International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV 11), Vancouver, British Columbia, Canada, 2011, pp. 69-74.
- [24] Tannenbaum: *Computer Networks*, 2002 4th Edition
- [25] Tonelli, Concas, Marchesi, Murgia: *An Analysis of SNA Metrics on the Java Qualitas Corpus*, 4th India Software Engineering Conference (ISEC 11), Thiruvananthapuram, Kerala, India, 2011, pp. 205-213.
- [26] Upadrashta, Vassileva, Grassmann: *Social networks in peer-to-peer systems.*, 38th Annual Hawaii International Conference on System Sciences (HICSS 05), Hawaii, USA, January 2005, pp. 200.
- [27] Valente, Coronges, Lakon, Costenbader: *How Correlated Are Network Centrality Measures?.* Connections (Toronto, Ont) (Vol. 28, No. 1), pp. 16-26, 2008.

- [28] Vu, Lupu, Ooi : *Peer-to-Peer Computing* Principles and Applications, Springer Berlin Heidelberg, 2010
- [29] Wang, Moreno, Sun: *Structure of peer-to-peer social networks.*, Phys. Rev. E, 73:036123, March 2006.
- [30] Wang, Zhao, Yuan: *Improving cooperation in peer-to-peer systems using social networks* 20th IEEE International Parallel Distributed Processing Symposium, Rhodes Island, Greek, April 2006, 8 pp.-.
- [31] Watts, and Strogatz: *Collective Dynamics of 'Small-World' Networks*, Nature, Volume 393, p. 440, 1998.
- [32] Apache Spark, online, <http://spark.apache.org/>, last accessed July 2016
- [33] Apnic, online <http://icons.apnic.net/pages/viewpage.action?pageId=491564> last accessed August 2016.
- [34] BitTorrent, online, <http://www.bittorrent.com/> , last accessed July 2016.
- [35] igraph, online, <http://igraph.org/> , last accessed July 2016.
- [36] Implicit None: Esteban Moroś Blog on Complexity, Networks, Marketing, Simulation, and more, online <http://estebanmoro.org/2015/12/temporal-networks-with-r-and-igraph-updated/>, last accessed July 2016.
- [37] netdraw, online, <https://sites.google.com/site/netdrawsoftware/download>, last accessed July 2016.
- [38] CIDR REPORT, online, <http://www.cidr-report.org/as2.0/>, last accessed July 2016.
- [39] SNAP: *Stanford Network Analysis Project*, online, <http://snap.stanford.edu/>, last accessed July 2016.
- [40] stackoverflow, online, <http://stackoverflow.com/>, last accessed July 2016.
- [41] stat.ethz.ch, online <https://stat.ethz.ch/R-manual/R-devel/library/>, last accessed July 2016.
- [42] statmethods, online, <http://www.statmethods.net/> , last accessed July 2016.
- [43] uTorrent, online, <http://www.utorrent.com/>, last accessed July 2016.
- [44] Vuze, online, <http://www.vuze.com/>, last accessed July 2016.
- [45] World Development Indicators World DataBank, online, <http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators&Type=TABLE&preview=on>, last accessed July 2016.

Abbreviations

AS Autonomous System. 1, 2, 3, 6, 7, 12, 13, 14, 16, 17, 18, 19, 25, 26, 29, 30, 31, 33, 34, 35, 36, 37, 38, 42, 45, 47, 48, 49, 52, 53, 54, 55, 57, 58, 59, 60, 61, 62, 63, 65, 66, 67, 68, 75, 76, 77, 79, 81

ASN Autonomous System Number. 1, 6, 17, 47, 53

BT BitTorrent. 1, 2, 3, 4, 5, 12, 13, 16, 17, 36, 46, 54, 62, 68

CI Count of infohashes per AS. 30, 31, 33, 35, 38, 47, 63, 67, 68, 75

CIP Count of IPs per AS. 30, 31, 35, 36, 38, 63, 67, 75

ER Erdős-Rényi. 8, 11, 22, 23, 27, 38, 39, 41, 42, 66, 75, 76

GDP Gross Domestic Product. 29, 30, 35, 36, 81

HDFS Hadoop File System. 16, 17, 19

HTTP Hypertext Transfer Protocol. 4

IP Internet Protocol. 1, 2, 6, 12, 13, 14, 16, 17, 18, 19, 30, 31, 35, 36, 37, 42, 48, 49, 53, 63, 67, 81

ISP Internet Service Provider. 6, 54

NIP Maximum number of IPs per AS. 30, 31, 35, 36, 37, 38, 63, 67, 75

P2P Peer-to-Peer. 3, 9

RAW Correlation variables of the raw AS data. 32, 35, 38

SNA Social Network Analysis. 1, 2, 3, 5, 7, 8, 9, 11, 13, 14, 15, 16, 18, 38, 49, 53, 55, 68, 76, 77

SNAM Correlation variables of the SNA metrics. 32, 35, 38

SQL Structured Query Language. 16

URL Uniform Resource Locator. 4

UTC Coordinated Universal Time. 16, 44, 46, 47, 53, 55

VA Vertiefungs Arbeit. 2

VIOLA Video Consumption in Overlay Networks. 2, 9, 12, 13, 14, 15, 16, 17, 18, 25, 26, 38, 39, 42, 50, 52, 62, 65, 68, 75, 76

VPN Virtual Private Network. 17

WDB Correlation variables of the Word DataBank. 32, 35, 37, 38

List of Figures

2.1	Unique ASs from end of September 1996 till mid of July 2016.	7
3.1	Visulaization of a sub graph with edge wheights above 1000 in netdraw . .	12
3.2	Two mode graph	13
3.3	First projection	13
3.4	Second projection	13
3.5	Fully connected swarm of one torrent where nodes represent IPs	14
3.6	Fully connected swarm of one torrent where nodes represent ASs	14
3.7	Graphical representation of the two examples.	15
3.8	Comparison of the three models with a plot and the corresponing degree distribution. Number of nodes is 100 and number of edges is 500.	24
4.1	Correlation plot of Pearson correlation coefficient. Compared to the other two Pearson shows less uniformity. Interestingly there are regions in the plot where only weak correlation coefficients exist which are high in the other two plots. like NIP, CIP and CI, together with eccentricity, eigenvector centrality, coreness and n2. n2 is a very uniform variable since it has similar values for most of the nodes, this is also why Pearson measures only weak correlation because the linearity of the variable is not met. . . .	31
4.2	Correlation plot of Spearman correlation coefficient.	32
4.3	Correlation plot of Kendall correlation coefficient. Looks very similar like Spearman but with weaker correlation coefficients.	32
4.4	Spearman correlation of the region SNAM to SNAM	34
4.5	Spearman correlation of the region SNAM to RAW	35
4.6	Spearman correlation of the region WDB to SNAM	36
4.7	Spearman correlation of the region RAW to WDB	36

4.8	Correlation Plot over the 48 measured time frames.	37
4.9	Local Clustering Coefficient VIOLA graph compared with ER random graph. Indicating a huge difference of the two networks.	39
4.10	Average Shortest Path VIOLA compared with random with ER random graph, while bot networks have about the same average path length.	39
4.11	Degree distribution of an Erdős-Rényi random graph with 11464 nodes and 7352865 edges	40
4.12	Scale-free random graph with 11464 nodes and 7352865 edges. Visible in this plot that the number of nodes with low degree is very small and, therefore the distribution forms an arc which is atypical in case of scale- freeness.	40
4.13	Shows the R^2 value of all the networks	41
4.14	Degree distribution of the first analyzed hour April 12, 2016 12 pm with 11464 nodes and 7352865 edges.	41
4.15	The diagram shows the degree distribution of a subgraph of the graph shown in Figure 4.14 where only edges with weights above 1000 were selected.	41
4.16	Correlation plots of the real-world data and three constructed random mod- els the first according to ER the second with small-world properties and the last with scale-free properties. The first row shows Spearman correlation and the second row Pearson correlation.	43
4.17	Mean of most node-centric measurments over time. The title indicates the y axis and the x axis is the time	45
4.18	Centralization scores over time. The title indicates the y axis and the x axis is the time	46
4.19	Net-centric measurements over time. The title indicates the y axis and the x axis is the time	47
4.20	Example image from the animation of the network only containing edges with weights bigger than 1000	50
4.21	Influence of the filtering on the dataset 5 am 14th March 2016, first part of metrics	51
4.22	Influence of the filtering on the dataset 5 am 14th March 2016, second part of metrics	52
4.23	Number of ASs per country per time frame this is only an overview	54

4.24	SNA metrics over the measured period from AS located in switzerland, where green (AS1836) and swisscom (AS3303) are telecommunication companies, switch (AS559) is a network service provider mostly specialised on univerities and some companies like Migros(AS8803) a retailer, La Roche a pharma company (AS2047), the famous CERN (AS513) and gouvermental Centre Informatique Etat de Fribourg as CIEF (AS25021)	56
4.25	SNA metrics over the measured period from AS located in costa rica with datasets of Cable Tica (AS52228), Radiografica Costaricense (AS3790), American Data Networks (AS27876) and Dole Shared Services (AS61479). Dole was added to scale the diagrams similar to the swiss ones becasue the dataset only contains zeros.	57
4.26	Modularity of the two different algorithms over time. Where fast greedy shows better performance in terms of modularity then walktrap.	59
4.27	This figure shows the division from the fast greedy algorithm of the AS in the different cluster groups. Membership means the cluster the AS was assigned. Showing the continent of the AS and its frequency. Clearly visible the algoritm did not divide the cluster continent wise.	59
4.28	This figure shows the division from the walk trap algorithm of the AS in the different cluster groups. Membership means the cluster the AS was assigned. Showing the continent of the AS and its frequency.	60
4.29	AS from Europe belonging to the different clusters by frequency.	61
4.30	AS from europa belonging to the different clusters by normalized frequency.	61
4.31	AS from europa belonging to the different clusters by frequency.	62
4.32	Assortativity of country, continent and degree.	63

List of Tables

3.1	Example of an edge list.	15
3.2	Example of an adjacency matrix representing a network.	15
4.1	Node-centric measurements checked for correlation	30
4.2	Country dependent real-world data	30
4.3	AS dependent data	31
4.4	Census of the Spearman correlation per region	33
4.5	List of all the top nodes occurring in all centrality rankings. Continent wise most of the important nodes belong to Europe with six followed by North America with four followed by Asia and South America with two and finally Oceania with one. An x in "top d" means in the top ten of the degree in all 48 networks, "top b" stands for betweenness, "top c" closeness and "top e" eigenvector centrality. On the bottom also mean of the top ranks with corresponding standard deviation and global mean is noted.	49
A.1	Variables overview	82
A.2	p-Values of the Pearson correlation matrix 14th March 5 am UTC	83
A.3	p-Values of the Pearson correlation matrix 14th March 5 am UTC	83
A.4	Correlation coefficient of the Spearman correlation matrix 14th March 5 am UTC	83
A.5	R-squared or coefficient of Determination of the Spearman correlation matrix 14th March 5 am UTC	84
A.6	p-Values of the Spearman correlation matrix 14th March 5 am UTC	84
A.7	Mean of the correlation coefficient of the 48 matrices. The mean absolute correlation is 0.5048278 and the mean correlation is 0.2161199.	84

A.8	Standart deviation of all time frames over the measured time. Interestingly nip the count of maximum possible IPs has the highest standard deviation of all the variables usually around 0.2. The mean standard deviation is 0.05480553.	85
A.9	Mean absolute difference of one time frame to the mean of all time frames. The mean diffrenece is 0.05545162	85
A.10	Correlation coefficient of the Kendall correlation matrix 14th March 5 am UTC	85
A.11	p-Values of the Kendall correlation matrix 14th March 5 am UTC	86
A.12	Shows a statistics of the four top ten lists evaluated for all 48 hours counted together.	86

Appendix A

Futher Material

A.1 Correlation

A.1.1 Variables overview

Table A.1: Variables overview

Name	Letter	Explanation
Degree	d	Number of links per node.
Weighted Degree	wd	Summation of all the links weight of one node.
Betweenness	b	Number of shortest paths going through the node divided by all shortest paths.
Closeness	c	Inverse of the sum of the fareness to all nodes.
Eccentricity	e	Fareness of the most distant node.
Eigenvector	ev	Eigenvector centrality of a node is high if he links to other important links.
Neighbourhood in two hops	n2	Neighbourhood of a node reachable in two hops. Where the neighbourhood of one hop would be the degree plus one, the starting node itself.
Coreness	cr	A vertex u has coreness c if it belongs to a c -core but not to any $(c + 1)$ -core.
Local Clustering Coefficient	lcc	The local clustering coefficient may vary from 0 to 1. It is zero if no neighbour is connected with any other neighbour and it is one if all neighbours are completely connected to each other.
GDP	gdp	GDP per country in USD (\$)
Growth	g	growth rate in USD (\$)
Mobile	m	Usage of mobile phones per 100 inhabitants.
Internet	i	Usage of internet per 100 inhabitants.
Countips	cip	Count of IPs per AS
Numips	nip	Number of maximum possible IPs per AS
Countinfohash	ci	Number of infohashes per AS

A.1.2 Pearson

Correlation Coefficient

	d	wd	b	c	ec	ev	cr	n2	lcc	nip	cip	ci	g	gdp	m	i
d	1	0.83	0.71	0.99	-0.57	0.95	0.72	0.17	-0.91	0.3	0.48	0.73	0.13	-0.06	-0.14	-0.03
wd	0.83	1	0.93	0.9	-0.27	0.67	0.34	0.06	-0.62	0.4	0.71	0.92	0.09	-0.01	-0.09	0.03
b	0.71	0.93	1	0.8	-0.2	0.54	0.25	0.04	-0.51	0.46	0.77	0.9	0.05	-0.01	-0.05	0.03
c	0.99	0.9	0.8	1	-0.51	0.91	0.64	0.19	-0.86	0.34	0.56	0.8	0.13	-0.05	-0.13	-0.01
ec	-0.57	-0.27	-0.2	-0.51	1	-0.7	-0.79	-0.2	0.62	-0.09	-0.12	-0.24	-0.16	0.02	0.17	0.03
ev	0.95	0.67	0.54	0.91	-0.7	1	0.86	0.23	-0.9	0.23	0.35	0.59	0.21	-0.04	-0.23	-0.05
cr	0.72	0.34	0.25	0.64	-0.79	0.86	1	0.32	-0.77	0.11	0.16	0.3	0.16	-0.07	-0.19	-0.08
n2	0.17	0.06	0.04	0.19	-0.2	0.23	0.32	1	-0.16	0.02	0.03	0.05	0.07	0	-0.08	-0.01
lc	-0.91	-0.62	-0.51	-0.86	0.62	-0.9	-0.77	-0.16	1	-0.21	-0.33	-0.55	-0.06	0.11	0.06	0.07
nip	0.3	0.4	0.46	0.34	-0.09	0.23	0.11	0.02	-0.21	1	0.52	0.41	0.02	0.07	-0.05	0.06
cip	0.48	0.71	0.77	0.56	-0.12	0.35	0.16	0.03	-0.33	0.52	1	0.72	0.04	0	-0.05	0.01
ci	0.73	0.92	0.9	0.8	-0.24	0.59	0.3	0.05	-0.55	0.41	0.72	1	0.08	-0.01	-0.09	0.03
g	0.13	0.09	0.05	0.13	-0.16	0.21	0.16	0.07	-0.06	0.02	0.04	0.08	1	0.09	-0.53	-0.22
gdp	-0.06	-0.01	-0.01	-0.05	0.02	-0.04	-0.07	0	0.11	0.07	0	-0.01	0.09	1	-0.21	0.45
m	-0.14	-0.09	-0.05	-0.13	0.17	-0.23	-0.19	-0.08	0.06	-0.05	-0.05	-0.09	-0.53	-0.21	1	0.26
i	-0.03	0.03	0.03	-0.01	0.03	-0.05	-0.08	-0.01	0.07	0.06	0.01	0.03	-0.22	0.45	0.26	1

Table A.2: p-Values of the Pearson correlation matrix 14th March 5 am UTC

p-Value

	d	wd	b	c	ec	ev	cr	n2	lcc	nip	cip	ci	g	gdp	m	i
d	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
wd	0	NA	0	0	0	0	0	0	0	0	0	0	0	0.14	0	0.01
b	0	0	NA	0	0	0	0	0	0	0	0	0	0	0.48	0	0
c	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	0.21
ec	0	0	0	0	NA	0	0	0	0	0	0	0	0	0.03	0	0
ev	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0	0
cr	0	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0
n2	0	0	0	0	0	0	0	NA	0	0.05	0.01	0	0	0.91	0	0.45
lc	0	0	0	0	0	0	0	0	NA	0	0	0	0	0	0	0
nip	0	0	0	0	0	0	0	0.05	0	NA	0	0	0.02	0	0	0
cip	0	0	0	0	0	0	0	0.01	0	0	NA	0	0	0.99	0	0.21
ci	0	0	0	0	0	0	0	0	0	0	0	NA	0	0.44	0	0
g	0	0	0	0	0	0	0	0	0.02	0	0	0	NA	0	0	0
gdp	0	0.14	0.48	0	0.03	0	0	0.91	0	0	0.99	0.44	0	NA	0	0
m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	0
i	0	0.01	0	0.21	0	0	0	0.45	0	0	0.21	0	0	0	0	NA

Table A.3: p-Values of the Pearson correlation matrix 14th March 5 am UTC

A.1.3 Spearman

Correlation Coefficient

	d	wd	b	c	ec	ev	cr	n2	lcc	nip	cip	ci	g	gdp	m	i
d	1	0.99	0.92	1	-0.77	0.97	0.99	0.85	-0.88	0.44	0.79	0.86	0.13	-0.15	-0.1	-0.04
wd	0.99	1	0.93	0.99	-0.77	0.97	0.98	0.86	-0.89	0.45	0.8	0.89	0.17	-0.15	-0.15	-0.04
b	0.92	0.93	1	0.92	-0.72	0.88	0.89	0.78	-0.99	0.44	0.81	0.88	0.06	-0.1	-0.06	-0.08
c	1	0.99	0.92	1	-0.77	0.97	0.99	0.85	-0.88	0.45	0.79	0.86	0.14	-0.15	-0.11	-0.04
ec	-0.77	-0.77	-0.72	-0.77	1	-0.8	-0.78	-0.9	0.68	-0.4	-0.62	-0.73	-0.17	0.1	0.16	0
ev	0.97	0.97	0.88	0.97	-0.8	1	0.98	0.9	-0.83	0.46	0.75	0.87	0.23	-0.14	-0.22	-0.01
cr	0.99	0.98	0.89	0.99	-0.78	0.98	1	0.86	-0.84	0.43	0.76	0.86	0.18	-0.16	-0.15	-0.04
n2	0.85	0.86	0.78	0.85	-0.9	0.9	0.86	1	-0.72	0.41	0.64	0.77	0.2	-0.1	-0.2	0.02
lc	-0.88	-0.89	-0.99	-0.88	0.68	-0.83	-0.84	-0.72	1	-0.42	-0.8	-0.85	-0.02	0.09	0.04	0.09
nip	0.44	0.45	0.44	0.45	-0.4	0.46	0.43	0.41	-0.42	1	0.59	0.53	0.13	0.07	-0.14	0.32
cip	0.79	0.8	0.81	0.79	-0.62	0.75	0.76	0.64	-0.8	0.59	1	0.88	0.11	-0.13	-0.08	-0.03
ci	0.86	0.89	0.88	0.86	-0.73	0.87	0.86	0.77	-0.85	0.53	0.88	1	0.19	-0.12	-0.2	-0.04
g	0.13	0.17	0.06	0.14	-0.17	0.23	0.18	0.2	-0.02	0.13	0.11	0.19	1	-0.18	-0.53	-0.03
gdp	-0.15	-0.15	-0.1	-0.15	0.1	-0.14	-0.16	-0.1	0.09	0.07	-0.13	-0.12	-0.18	1	-0.08	0.44
m	-0.1	-0.15	-0.06	-0.11	0.16	-0.22	-0.15	-0.2	0.04	-0.14	-0.08	-0.2	-0.53	-0.08	1	0.01
i	-0.04	-0.04	-0.08	-0.04	0	-0.01	-0.04	0.02	0.09	0.32	-0.03	-0.04	-0.03	0.44	0.01	1

Table A.4: Correlation coefficient of the Spearman correlation matrix 14th March 5 am UTC

Correlation Determination Coefficient

	d	wd	b	c	ec	ev	cr	n2	lcc	nip	cip	ci	g	gdp	m	i
d	1	0.98	0.85	1	0.59	0.93	0.98	0.71	0.78	0.2	0.63	0.74	0.02	0.02	0.01	0
wd	0.98	1	0.86	0.97	0.6	0.94	0.97	0.73	0.79	0.21	0.64	0.8	0.03	0.02	0.02	0
b	0.85	0.86	1	0.85	0.52	0.77	0.79	0.6	0.97	0.19	0.66	0.77	0	0.01	0	0.01
c	1	0.97	0.85	1	0.59	0.94	0.97	0.73	0.78	0.2	0.63	0.75	0.02	0.02	0.01	0
ec	0.59	0.6	0.52	0.59	1	0.64	0.61	0.81	0.46	0.16	0.39	0.53	0.03	0.01	0.03	0
ev	0.93	0.94	0.77	0.94	0.64	1	0.95	0.81	0.68	0.21	0.56	0.75	0.05	0.02	0.05	0
cr	0.98	0.97	0.79	0.97	0.61	0.95	1	0.74	0.71	0.19	0.58	0.74	0.03	0.03	0.02	0
n2	0.71	0.73	0.6	0.73	0.81	0.81	0.74	1	0.52	0.17	0.41	0.6	0.04	0.01	0.04	0
lc	0.78	0.79	0.97	0.78	0.46	0.68	0.71	0.52	1	0.17	0.63	0.73	0	0.01	0	0.01
nip	0.2	0.21	0.19	0.2	0.16	0.21	0.19	0.17	0.17	1	0.34	0.28	0.02	0	0.02	0.1
cip	0.63	0.64	0.66	0.63	0.39	0.56	0.58	0.41	0.63	0.34	1	0.77	0.01	0.02	0.01	0
ci	0.74	0.8	0.77	0.75	0.53	0.75	0.74	0.6	0.73	0.28	0.77	1	0.04	0.01	0.04	0
g	0.02	0.03	0	0.02	0.03	0.05	0.03	0.04	0	0.02	0.01	0.04	1	0.03	0.28	0
gdp	0.02	0.02	0.01	0.02	0.01	0.02	0.03	0.01	0.01	0	0.02	0.01	0.03	1	0.01	0.19
m	0.01	0.02	0	0.01	0.03	0.05	0.02	0.04	0	0.02	0.01	0.04	0.28	0.01	1	0
i	0	0	0.01	0	0	0	0	0	0.01	0.1	0	0	0	0.19	0	1

Table A.5: R-squared or coefficient of Determination of the Spearman correlation matrix 14th March 5 am UTC

p-Value

	d	wd	b	c	ec	ev	cr	n2	lcc	nip	cip	ci	g	gdp	m	i
d	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
wd	0	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	0
ec	0	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	0.65
ev	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0	0.17
cr	0	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0
n2	0	0	0	0	0	0	0	NA	0	0	0	0	0	0	0	0.09
lc	0	0	0	0	0	0	0	0	NA	0	0	0	0.03	0	0	0
nip	0	0	0	0	0	0	0	0	0	NA	0	0	0	0	0	0
cip	0	0	0	0	0	0	0	0	0	0	NA	0	0	0	0	0
ci	0	0	0	0	0	0	0	0	0	0	0	NA	0	0	0	0
g	0	0	0	0	0	0	0	0	0.03	0	0	0	NA	0	0	0
gdp	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	0	0
m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	0.4
i	0	0	0	0	0.65	0.17	0	0.09	0	0	0	0	0	0	0.4	NA

Table A.6: p-Values of the Spearman correlation matrix 14th March 5 am UTC

Mean Correlation Coefficient over all Time Frames

	d	wd	b	c	ec	ev	cr	n2	lcc	nip	cip	ci	g	gdp	m	i
d	1	0.99	0.93	1	-0.71	0.97	0.99	0.76	-0.89	0.64	0.61	0.83	0.11	-0.18	-0.06	-0.06
wd	0.99	1	0.93	0.99	-0.73	0.97	0.98	0.78	-0.89	0.66	0.62	0.84	0.15	-0.18	-0.11	-0.05
b	0.93	0.93	1	0.93	-0.66	0.89	0.9	0.69	-0.99	0.64	0.61	0.84	0.03	-0.12	-0.02	-0.1
c	1	0.99	0.93	1	-0.72	0.97	0.99	0.77	-0.89	0.64	0.61	0.83	0.12	-0.18	-0.07	-0.06
ec	-0.71	-0.73	-0.66	-0.72	1	-0.76	-0.72	-0.95	0.61	-0.49	-0.43	-0.6	-0.16	0.12	0.13	-0.01
ev	0.97	0.97	0.89	0.97	-0.76	1	0.98	0.82	-0.83	0.66	0.6	0.81	0.21	-0.17	-0.18	-0.03
cr	0.99	0.98	0.9	0.99	-0.72	0.98	1	0.77	-0.85	0.63	0.59	0.82	0.16	-0.19	-0.12	-0.06
n2	0.76	0.78	0.69	0.77	-0.95	0.82	0.77	1	-0.64	0.5	0.44	0.61	0.17	-0.12	-0.14	0.01
lc	-0.89	-0.89	-0.99	-0.89	0.61	-0.83	-0.85	-0.64	1	-0.61	-0.59	-0.82	0.01	0.1	0	0.11
nip	0.64	0.66	0.64	0.64	-0.49	0.66	0.63	0.5	-0.61	1	0.54	0.69	0.17	-0.03	-0.17	0.14
cip	0.61	0.62	0.61	0.61	-0.43	0.6	0.59	0.44	-0.59	0.54	1	0.72	0.11	-0.03	-0.09	0.14
ci	0.83	0.84	0.84	0.83	-0.6	0.81	0.82	0.61	-0.82	0.69	0.72	1	0.12	-0.14	-0.11	-0.04
g	0.11	0.15	0.03	0.12	-0.16	0.21	0.16	0.17	0.01	0.17	0.11	0.12	1	-0.17	-0.5	0.02
gdp	-0.18	-0.18	-0.12	-0.18	0.12	-0.17	-0.19	-0.12	0.1	-0.03	-0.03	-0.14	-0.17	1	-0.08	0.42
m	-0.06	-0.11	-0.02	-0.07	0.13	-0.18	-0.12	-0.14	0	-0.17	-0.09	-0.11	-0.5	-0.08	1	-0.01
i	-0.06	-0.05	-0.1	-0.06	-0.01	-0.03	-0.06	0.01	0.11	0.14	0.14	-0.04	0.02	0.42	-0.01	1

Table A.7: Mean of the correlation coefficient of the 48 matrices. The mean absolute correlation is 0.5048278 and the mean correlation is 0.2161199.

Standard deviation over all Time Frames

	d	wd	b	c	ec	ev	cr	n2	lcc	nip	cip	ci	g	gdp	m	i
d	0	0.000945352	0.007913588	0.000375432	0.042735956	0.003312386	0.001309238	0.063688667	0.011804499	0.215817758	0.191056816	0.039008578	0.028815825	0.028527442	0.029994826	0.013599893
wd	0.000945352	0	0.00577693	0.000866401	0.038033107	0.002579121	0.001495753	0.059233713	0.009156745	0.224288435	0.185996496	0.052103889	0.025305111	0.025532461	0.02612591	0.012417678
b	0.007913588	0.00577693	0	0.007510621	0.050334815	0.009955822	0.009372529	0.065182212	0.001573262	0.231059478	0.202076432	0.031921953	0.022346744	0.020531859	0.02180684	0.010451586
c	0.000375432	0.000866401	0.007510621	0	0.041875483	0.00282866	0.001401325	0.062849424	0.011354549	0.214364845	0.187651558	0.039599925	0.029610708	0.028011212	0.030351997	0.013559368
ec	0.042735956	0.038033107	0.050334815	0.041875483	0	0.033676561	0.042389166	0.038986128	0.05664297	0.174244204	0.104344243	0.068275344	0.038111947	0.018816619	0.046321685	0.018742796
ev	0.003312386	0.002579121	0.009955822	0.00282866	0.033676561	0	0.003064934	0.046270496	0.013847084	0.207408531	0.157867981	0.065473185	0.031412206	0.027492304	0.031444573	0.01376018
cr	0.001309238	0.001495753	0.009372529	0.00282866	0.033676561	0.003064934	0	0.06145691	0.013598875	0.221082732	0.182511757	0.05143543	0.030521629	0.030960877	0.030750576	0.013687493
n2	0.063688667	0.059233713	0.065182212	0.062849424	0.038986128	0.046270496	0.06145691	0	0.067964932	0.187605081	0.109960603	0.075023748	0.032427849	0.017404913	0.036797657	0.019220347
lcc	0.011804499	0.009156745	0.001573262	0.011354549	0.05664297	0.013847084	0.013598875	0.067964932	0	0.22839131	0.202758479	0.027953714	0.021887407	0.018581098	0.019153637	0.010994886
nip	0.215817758	0.224288435	0.231059478	0.214364845	0.174244204	0.207408531	0.221082732	0.187605081	0.22839131	0	0.025758272	0.148059797	0.047369952	0.050901011	0.01105277	0.018662458
cip	0.191056816	0.185996496	0.202076432	0.187651558	0.104344243	0.157867981	0.182511757	0.109960603	0.202758479	0.025758272	0	0.148059797	0	0.050901011	0.01105277	0.018662458
ci	0.039008578	0.052103889	0.025305111	0.022346744	0.029610708	0.038111947	0.031412206	0.030521629	0.027953714	0.021887407	0.017960364	0.047369952	0.050901011	0	0.01105277	0.018662458
g	0.028815825	0.025532461	0.020531859	0.028011212	0.030351997	0.031444573	0.030750576	0.036797657	0.019183637	0.01839779	0.06524387	0.071208899	0.011406249	0.011412062	0	0.018662458
gdp	0.028527442	0.025532461	0.020531859	0.028011212	0.030351997	0.031444573	0.030750576	0.036797657	0.019183637	0.01839779	0.06524387	0.071208899	0.011406249	0.011412062	0	0.018662458
m	0.029994826	0.02612591	0.02180684	0.030351997	0.046321685	0.031444573	0.030750576	0.036797657	0.019183637	0.01839779	0.06524387	0.071208899	0.011406249	0.011412062	0	0.018662458
i	0.013599893	0.012417678	0.010451586	0.013559368	0.018742796	0.01376018	0.013687493	0.019220347	0.010994886	0.176175063	0.177211083	0.01882495	0.030141538	0.013815651	0.018662458	0

Table A.8: Standart deviation of all time frames over the measured time. Interestingly nip the count of maximum possible IPs has the highest standard deviation of all the variables usually around 0.2. The mean standard deviation is 0.05480553.

Mean absolute difference

	d	wd	b	c	ec	ev	cr	n2	lcc	nip	cip	ci	g	gdp	m	i
d	0	0	0.01	0	0.05	0	0	0.08	0.01	0.2	0.18	0.03	0.02	0.03	0.04	0.02
wd	0	0	0.01	0	0.05	0	0	0.08	0	0.21	0.18	0.05	0.02	0.03	0.04	0.02
b	0.01	0.01	0	0.01	0.06	0.01	0.01	0.08	0	0.2	0.2	0.03	0.03	0.02	0.04	0.02
c	0	0	0.01	0	0.05	0	0	0.08	0.01	0.19	0.18	0.04	0.02	0.03	0.04	0.02
ec	0.05	0.05	0.06	0.05	0	0.04	0.06	0.05	0.07	0.09	0.19	0.13	0.01	0.02	0.03	0
ev	0	0	0.01	0	0.04	0	0	0.08	0.01	0.19	0.15	0.06	0.02	0.03	0.04	0.01
cr	0	0	0.01	0	0.06	0	0	0.09	0.01	0.2	0.17	0.04	0.02	0.03	0.04	0.02
n2	0.08	0.08	0.08	0.08	0.05	0.08	0.09	0	0.09	0.09	0.2	0.16	0.02	0.03	0.06	0.01
lcc	0.01	0	0	0.01	0.07	0.01	0.01	0.09	0	0.2	0.2	0.03	0.03	0.01	0.03	0.02
nip	0.2	0.21	0.2	0.19	0.09	0.19	0.2	0.09	0.2	0	0.05	0.17	0.04	0.1	0.03	0.18
cip	0.18	0.18	0.2	0.18	0.19	0.15	0.17	0.2	0.2	0.05	0	0.16	0	0.1	0.02	0.17
ci	0.03	0.05	0.03	0.04	0.13	0.06	0.04	0.16	0.03	0.17	0.16	0	0.07	0.02	0.09	0.01
g	0.02	0.02	0.03	0.02	0.01	0.02	0.02	0.02	0.03	0.04	0	0	0	0	0.02	0.05
gdp	0.03	0.03	0.02	0.03	0.02	0.03	0.03	0.03	0.01	0.1	0.1	0.02	0	0	0	0.01
m	0.04	0.04	0.04	0.04	0.03	0.04	0.04	0.06	0.03	0.03	0.02	0.09	0.02	0	0	0.02
i	0.02	0.02	0.02	0.02	0	0.01	0.02	0.01	0.02	0.18	0.17	0.01	0.05	0.01	0.02	0

Table A.9: Mean absolute difference of one time frame to the mean of all time frames. The mean diffrence is 0.05545162

A.1.4 Kendall

Correlation Coefficient

	d	wd	b	c	ec	ev	cr	n2	lcc	nip	cip	ci	g	gdp	m	i
d	1	0.92	0.78	0.98	-0.63	0.87	0.93	0.69	-0.73	0.31	0.63	0.69	0.1	-0.11	-0.07	-0.03
wd	0.92	1	0.77	0.91	-0.63	0.87	0.9	0.7	-0.73	0.32	0.63	0.74	0.12	-0.11	-0.1	-0.03
b	0.78	0.77	1	0.78	-0.6	0.71	0.72	0.61	-0.92	0.31	0.65	0.72	0.04	-0.07	-0.04	-0.06
c	0.98	0.91	0.78	1	-0.63	0.88	0.92	0.71	-0.73	0.32	0.62	0.7	0.1	-0.1	-0.08	-0.03
ec	-0.63	-0.63	-0.6	-0.63	1	-0.65	-0.65	-0.8	0.56	-0.33	-0.52	-0.6	-0.15	0.08	0.14	0
ev	0.87	0.87	0.71	0.88	-0.65	1	0.89	0.77	-0.66	0.33	0.57	0.71	0.16	-0.1	-0.15	-0.01
cr	0.93	0.9	0.72	0.92	-0.65	0.89	1	0.71	-0.66	0.3	0.59	0.69	0.13	-0.11	-0.11	-0.03
n2	0.69	0.7	0.61	0.71	-0.8	0.77	0.71	1	-0.56	0.31	0.49	0.61	0.15	-0.07	-0.15	0.01
lcc	-0.73	-0.73	-0.92	-0.73	0.56	-0.66	-0.66	-0.56	1	-0.3	-0.63	-0.69	-0.02	0.06	0.03	0.07
nip	0.31	0.32	0.31	0.32	-0.33	0.33	0.3	0.31	-0.3	1	0.43	0.38	0.1	0.05	-0.1	0.22
cip	0.63	0.63	0.65	0.62	-0.52	0.57	0.59	0.49	-0.63	0.43	1	0.73	0.08	-0.09	-0.05	-0.02
ci	0.69	0.74	0.72	0.7	-0.6	0.71	0.69	0.61	-0.69	0.38	0.73	1	0.14	-0.08	-0.14	-0.03
g	0.1	0.12	0.04	0.1	-0.15	0.16	0.13	0.15	-0.02	0.1	0.08	0.14	1	-0.14	-0.39	0.04
gdp	-0.11	-0.11	-0.07	-0.1	0.08	-0.1	-0.11	-0.07	0.06	0.05	-0.09	-0.08	-0.14	1	-0.07	0.29
m	-0.07	-0.1	-0.04	-0.08	0.14	-0.15	-0.11	-0.15	0.03	-0.1	-0.05	-0.14	-0.39	-0.07	1	0.03
i	-0.03	-0.03	-0.06	-0.03	0	-0.01	-0.03	0.01	0.07	0.22	-0.02	-0.03	0.04	0.29	0.03	1

Table A.10: Correlation coefficient of the Kendall correlation matrix 14th March 5 am UTC

p-Value

	d	wd	b	c	ec	ev	cr	n2	lcc	nip	cip	ci	g	gdp	m	i
d	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
wd	0	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	0
ec	0	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	0.65
ev	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0	0.22
cr	0	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0
n2	0	0	0	0	0	0	0	NA	0	0	0	0	0	0	0	0.07
lc	0	0	0	0	0	0	0	0	NA	0	0	0	0.02	0	0	0
nip	0	0	0	0	0	0	0	0	0	NA	0	0	0	0	0	0
cip	0	0	0	0	0	0	0	0	0	0	NA	0	0	0	0	0
ci	0	0	0	0	0	0	0	0	0	0	0	NA	0	0	0	0
g	0	0	0	0	0	0	0	0	0.02	0	0	0	NA	0	0	0
gdp	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	0	0
m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	0
i	0	0	0	0	0.65	0.22	0	0.07	0	0	0	0	0	0	0	NA

Table A.11: p-Values of the Kendall correlation matrix 14th March 5 am UTC

A.2 Network over Time**A.2.1 Important nodes**

degree	freq degree	betweenness	freq betweenness	closeness	freq closeness	evcent	freq evcent
28573	48	14618	48	28573	48	5384	48
5089	48	28573	48	5089	48	6830	48
5384	48	5089	48	5384	48	7018	48
6830	48	6830	48	6830	48	7922	48
7018	48	7922	48	7018	48	5089	47
7922	48	18881	47	7922	48	9299	37
9299	34	36351	36	9299	34	8551	35
1221	24	5384	32	1221	24	28573	30
3352	21	7018	32	3352	21	3352	21
8551	17	9299	23	8551	17	1221	20
36351	14	3352	8	36351	14	4134	16
4788	13	2860	6	4788	13	4788	14
6327	9	1221	3	6327	9	6327	13
2856	8	3243	2	2856	8	2856	11
2860	7	27699	1	2860	7	5607	8
5607	6	2856	1	5607	6	6799	4
18881	3	6327	1	18881	3	2860	3
3243	3			3243	3	36351	3
25019	2			25019	2	8708	3
4134	2			4134	2	3243	2
						17974	1
						18881	1
						25019	1
						852	1

Table A.12: Shows a statistics of the four top ten lists evaluated for all 48 hours counted together.

Appendix B

Contents of the CD

- Abstr.txt
- Bachelorarbeit.pdf
- BA_Dokumente
- Bilder_und_Tables
- Code
- Jars
- Latex_Dateien
- Literature
- Midterm_Präsentation
- R_scripts
- Text_files
- Zusfs.txt