CONTENT PROPAGATION

IN ONLINE SOCIAL NETWORKS

CONTENT PROPAGATION

IN ONLINE SOCIAL NETWORKS

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K. C. A. M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op vrijdag 13 juni 2014 om 12.30 uur

door

Norbert BLENN

Diplom Medieninformatiker, Technische Universität Dresden, Duitsland
geboren te Burgstädt, Duitsland.

Dit proefschrift is goedgekeurd door de promotoren:

Prof. dr. ir. P. F. A. Van Mieghem
Copromotor: Dr. C. Doerr

Samenstelling promotiecommissie:

Rector Magnificus,	Voorzitter
Prof. dr. ir. P. F. A. Van Mieghem,	Technische Universiteit Delft, promotor
Dr. C. Doerr,	Technische Universiteit Delft, copromotor
Prof. dr. G. J. Houben	Technische Universiteit Delft
Prof. dr. A. Hanjalic	Technische Universiteit Delft
Prof. dr. ir. N. Baken	Technische Universiteit Delft
Prof. dr. B. Brown	Universiteit van Stockholm
Prof. dr. D. Epema	Technische Universiteit Eindhoven en Technische Universiteit Delft

Dit onderzoek werd gesteund door de Technische Universiteit Delft, de Koninklijke KPN N.V. en TNO in de initiative Trans-sector Research Academy for complex Networks and Services (TRANS).

Title: Content Propagation in Online Social Networks

Front & Back: Twitter friendship relations

ISBN 978-94-6186-324-9

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

To my small social network,
my family and friends.

CONTENTS

1

INTRODUCTION

Societies are nowadays defined as groups of individuals having personal relationships with each other. A society defines itself through cultural norms describing relationships. This denotes that a society can be described by relationships between its members and certain rules or norms of behavior explain how individuals interact with other persons. Therefore an individual as smallest entity of a society has only a limited impact onto the society on it's own but the actions and interactions of many are key.

With the advent of Online Social Networks (i.e. OSNs), researchers became able to analyze social interaction and user behavior within online environments at a larger scale than ever before. It became possible to obtain data not only about attributes of individuals but also their interactions with others, timing and activity information and insights about the quantity of communication on a large scale. This is not to forget about the ability to analyze complete (online) populations of users and the network based on relationships between individuals.

Results of the analysis of online social networks to a large extent mirrors findings from real-world experiments. Among them is the famous finding of Milgram for example, which states that every person in a population is on average connected to any other person by only "six degrees of separation"[1] which was confirmed by Watts [2] in an experiment using e-mails and by Ugander *et al.* [3] analyzing the social graph of Facebook.com as well. Another example may be given by the number of people with which an individual is able to maintain social relations, known as "Dunbar's number". The anthropologist Dunbar found a correlation between the primate brain size and the average size of social groups of primates. Extrapolating this number to the size of human brains led to the assumption that a human is able to maintain approximately 150 relations [4]. A similar number was also found by analyzing conversations within Twitter [5] and Facebook [6]. Another effect called homophily, mening the fact that friends have similar interests and attributes, often called the "birds of a feather flock together" effect, became measurable in large scale online social networks. These examples only state a few points in which sociological findings were repeated in OSNs, typically using larger data sets as before, because in OSN analysis the need for expensive and time consuming personal

interviews diminishes.

The benefits of OSN analysis are therefore manifold. For instance recommendation systems dramatically improve if individuals are not reduced to descriptive numbers specified by the items they bought or their attributes. This indicates a transformation from the classical way of representing users in e-commerce as "you are what you bought" to "you like what your friends like", a change that improves the usage of online marketplaces but raises privacy concerns. Naturally a difficult game arises from this transition as individuals usually want to receive improved recommendations but without revealing too much about themselves. Therefore the strange gut feeling remains, that unknown persons, companies or even your neighbor might know "more" about one than one oneself. Due to homophily it becomes even possible to reconstruct information of OSN users who hide their information based on the information the friends of that person are communicating.

1.1. RESEARCH QUESTIONS

The underlying network of relationships within an OSN describes and limits possible content flows because only connected users may share information with each other. In order to estimate topological factors like the amount, strength and usage of relationships as well as properties of actors in such a network, the approach of interpreting a social network as a graph $G(N, L)$, where individuals are nodes (N) and friendships are represented as links (L) is chosen. The benefit of this interpretation lies in the fact that well established theoretical models from different disciplines, like epidemiological models from medicine and biology but also routing techniques from network science, branching processes and random walks from mathematics and physics or certain concepts from sociology just to name a few, aid in the understanding of content propagation in OSNs and the estimation of importance of users and friendship relations. To which extent these theoretical models and metrics can be verified through measurements of empirical data states one of the main questions in this thesis.

The strength of relationships between individuals might denote which relations a user will prefer when spreading information. In this respect the method used for the estimation of link-weights, describing how close connected individuals are to each other, states a research question itself. While a common approach is based on the amount of communication traversing a link normalized by the total amount of communication, a more "social" approach would quantify how close two individuals are to each other given the information one obtains from an OSN, therefore incorporating attributes of the users.

Due to the overwhelming size of most OSNs nowadays (Twitter reports 255 million [51] and Facebook, 1.28 billion monthly active users [50]) the process of obtaining data might be quite complicated. If topological data is obtained though crawling it might be infeasible to crawl the whole network due its size. That is why fractions of networks are analyzed for simplicity, which possibly distorts certain metrics because of the used traversal techniques. The quantization of such bias in terms of topological metrics states a research question of utter importance. If such bias occurs, different techniques of traversing a graph would be necessary.

Content that propagates through an online social network is typically given by mes-

sages, images or videos which, once they are created or uploaded to online services are forwarded by registered users to their own peers. This means once a message is written, it is in most cases impossible to prevent the content from distributing. One may even claim that the content behaves like a virus propagating from infected individuals to healthy (non-infected) friends or acquaintances. Extending this line of thought by deeper analysis of the content, one may claim that not only the information propagates, but also opinions and feelings as shown by Christakis and Fowler [7] who found that obesity, smoking and even happiness can be interpreted as viruses propagating via relationships in real-world social networks. The main focus of this thesis lies therefore in the question, how to model content propagation in OSNs and which factors are involved in the process.

1.2. CONTRIBUTIONS

Based on data from a large Dutch OSN, it will be shown how to estimate the similarity of friends to each other and to which extent this similarity can be used to estimate private information, like the age, interests or home town of a user. This analysis described in Chapter 2 shows that current privacy settings in OSNs are not sufficient because of the possibility to reconstruct a users profile from public information of friends. The findings, published in [33], might also depict methods to improve recommendation systems and social search engines because it is shown that, by incorporating information of friends, predictions in terms of interests can be improved. In terms of the desired link-weight distribution the similarity of between users can directly be translated into a measure of strength of the relationship as users with a high similarity are likely to be close friends.

When analyzing social networks in a large scale, one needs to obtain topological information, usually by employing standard methods like breadth- or depth-first searches. Is is shown in Chapter 3.1, that these techniques, while not completely finished, introduce bias towards different network metrics which denotes that only after obtaining a large fraction of the network of an OSN, estimations of the final values of metrics can be drawn. Therefore a more practical way of obtaining data from an OSN called “Mutual Friend Crawling” (MFC) published in [56], is proposed. Mutual Friend Crawling traverses a social graph community-wise, enabling the analysis of sub-graphs, if obtaining the whole network is infeasible.

Based on content traces from Twitter and a complete dataset of an Digg.com, it will be shown in Chapter 3 that friendship relations, inevitably necessary in order to maximize the useability of OSNs are not as useful, in terms of content propagation, as previously thought, because only a limited fraction of friends or followers will actively support a user in spreading content (published in [34, 77]). On the other hand, the existence of influential groups of users is analyzed and it is explained that the detection of such groups is infeasible by only using topological information as also published in [61].

Content being spread through a social network is often referred to as “viral”-spreading. In Chapter 5 the relation between “viral”-spreading and epidemics will be depicted, showing that certain messages in Twitter, may be as infectious as hazardous diseases just after appearing in the online service but limitations exist that hinder content from becoming an Internet meme, i.e. spreading to most registered users in the OSN. Additionally the results from the analysis of empirical data shows that certain models of epi-

demiology cannot directly be facilitated because of non-Markovian properties of content propagation within OSNs as published in [143].

1.3. STRUCTURE OF THIS THESIS

This thesis describes research in online social networks starting with the smallest entity of a social network, an individual or a node, on to ego-centric networks in Chapter 2. These ego-centric networks depict the view of individuals as they only include direct peers enabling studies of individuals attributes and the influence of direct friends onto the central (ego) node. Chapter 3 focuses on large groups of individuals and quantifies relationships in terms of interactions and content propagation. Certain methods to obtain data from OSNs are opposed to each other and the usage of friendship relations in terms of distributing information is analyzed. In chapter 4 the dynamics in terms of structural changes are shown and their influence on content propagation is shown. In Chapter 5 the analysis of content dissemination in concrete examples is described and different models and the implication of user behavior onto these models is explained. The work closes with chapter 6 that gives examples of the analysis of the actual propagating content, followed by a conclusion and possible topics for future work.

2

EGOCENTRIC NETWORK ANALYSIS

In this chapter the most fundamental elements within an online social network, an individual's user account and the relations between peers, will be described. Chapter 2.1 describes individuals and their attributes by analyzing profile information. The general question behind the analysis is, if users of online social networks form a random subset of the population of a country. In Chapter 2.2 the relations between users are introduced and used in Chapter 2.3 to estimate the similarity of pairs of friends. These similarities are further on shown to enable the estimation of private attributes of individuals. Chapter 2.4 exemplifies the analysis of ego-centric networks through the analysis of graphs describing how direct "neighbors" are connected.

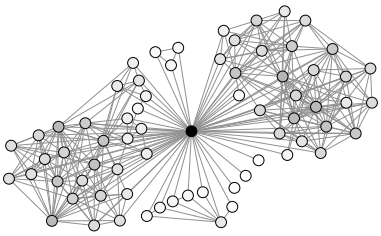


Figure 2.1: Ego network of the author based on Facebook relations. Shading represent the degree of nodes.

Every individual has family members, relatives, friends, acquaintances, colleagues and so on which are linked in some way to the person. The network that is defined through everyone an individual knows is called an egocentric network, because it represents an individual, the "ego", and the directly linked persons("alters" or "neighbors") as nodes. All nodes in the egocentric network have properties like name, gender, age, hometown, education, income, etc., which state the basis for egocentric social network analysis. The relations, modeled as links, can be annotated with attributes as well, like the type of the relationship which denotes if an alter is a family member an acquaintance or a friend for example, or by numerical values defining the strength of a relationship, the amount of communication between peers or other metrics that describe the similarity of a connected pair of people.

Figure 2.1 depicts the egocentric network of the author of this thesis based on relations from the OSN Facebook. The node with the highest degree (black) depicts the "ego"

which is connected to all other nodes ("alters") in the network. This very simplified view does assume that the ego knows which of the alters do know each other and includes therefore the existing links between alters.

2.1. ATTRIBUTES OF AN INDIVIDUAL

Usually, every node in a social graph denotes an individual having certain attributes. In general, the characteristics of a user can be classified into two groups: intrinsic attributes (such as name, age, city and gender) and communities (school, college, university, company, sports club or interests). These attributes are either known to the node, like the age, name, home town etc., or they are estimated based on the observations of the "behavior" of individuals like the time a user was online and the duration or quantity of participation in certain activities, like attending surveys or the number of sent messages to name some examples. This section will show analyses on both of these types of attributes.

In order to store the attributes of a high number of user(-accounts), one usually facilitates a matrix notation as exemplary shown in Table 2.1.

	Name	Age	Gender	Home town	Education	...
Ego	Norbert	32	male	Delft	University	
Alter I	Marcel	28	male	Leipzig	University	
Alter II	Karolina	30	female			
Alter III	Marcel		male	Dresden	MBO	
Alter IV	Daniela		female	Delft	University	
...						

Table 2.1: Excerpt of a matrix describing attributes of nodes (individuals).

The rows of the matrix denote observed individuals, the columns list quantitative or qualitative measured attributes. By comparing rows one may analyze which actors are similar to others, which attributes are more common than others and if attributes are related. This means one may estimate for example the location of a person based on favorite sport clubs, bars or restaurants the person mentioned. Through such an analysis one can on one the hand inform about possible problems concerning privacy or build useful recommendation systems. On the other hand, by inferring attributes of an individual, one may test if provided information is correct or fill missing fields (like the hometown and education of alter II or the age of alter III and IV) in the data set.

Basis for attribute analytics is the availability of a data set of user profiles, that is either complete or states a good sample of all users of interest. The data of Hyves.nl described in A.1.5, used in this thesis contains information about roughly half of all registered users of the OSN obtained though various techniques described in the appendix on page 146. The data from Twitter described in A.1.9 was obtained though listening to the "sample-stream" API interface. Twitter messages received through this API endpoint are

systematically sampled[1], where every 100th out of all sent messages is available. When analyzing the opinion or attributes of users however, one wants to sample users from a certain population of a country. A task which is not as trivial as sampling from sent messages. The following subsection will describe techniques to obtain a sample of individuals from a country.

2.1.1. DEMOGRAPHICS OF USERS

Everything a person writes or uploads in OSNs is stored and indexed by operators as valuable searchable information. When analyzing to which extent data sampled from Twitter.com or Hyves.nl is representative for the population of the Netherlands, it is possible to analyze the bias, introduced by different sampling techniques using different attributes of individuals:

1. The geographical location,
2. the family and first name
3. the gender and
4. the marital status of a user.

By comparing observed attributes from the two OSNs to reliable data from the Dutch Central Bureau of Statistics[2] and data from Meertens Institute[3], a research institute of the royal Netherlands Academy of Arts and Sciences (KNAW) that studies "the diversity in language and culture in the Netherlands", a qualitative analysis can be performed.

Data used for this experiment is obtained from Twitter and Hyves. The user information from Twitter is obtained through the "sample-stream" API, an interface that continuously delivers a random subset of 1% of all messages written on Twitter. This one percent systematic sample implies that every person or service that connects to the API-endpoint receives precisely the same sample of tweets.

Figure 2.2 shows the frequency of different languages in the messages obtained through the "sample-stream" API of Twitter in the period of the 3 last weeks in April 2013, detected using the library "langdetect" [10]. This list does not follow the most frequently spoken languages in the world which are listed in Table 2.2.

Detecting the language of text is difficult especially if the length of messages is limited to 140 characters. It is therefore possible that some tweets are misclassified. However, the overall distribution shows major differences from the data shown in Table 2.2, which indicates that Twitter users are not randomly distributed across the world's population. The reason lies in the fact that micro-blogging services similar to Twitter exist for example in China "Sina Weibo", a service very similar to Twitter is the prominent platform.

Although the worldwide population is not correctly sampled by Twitter data, publications can be found in which researchers claimed to be successful in predicting certain

[1] **Systematic sampling**: If the size of the population is unknown or continuously growing, a *systematic sampling* process analyzes every $k's$ event out of all produced events within the population. The sample provided by the OSN Twitter through an API endpoint, delivers every 100th message written at the service. The total population in this case is given by the total number of Tweets that have been submitted to the service.

[2] CBS data is available at `www.cbs.nl`

[3] Meertens Institute's data is available at `www.meertens.knaw.nl`

Language	Percentage of world population
Mandarin	14%
Spanish	5.85%
English	5.52%
Hindi	4.46%
Arabic	4.23%
Portuguese	3.08%
Bengali	3.05%
Russian	2.42%
Japanese	1.92%
Punjabi	1.44%
German	1.39%
Javanese	1.25%
Wu	1.20%
Malay/Indonesian	1.16%
Telugu	1.15%
Vietnamese	1.14%
Korean	1.14%
French	1.12%

Table 2.2: Most frequently spoken languages in the world, as listed in the Swedish "Nationalencyklopedin" [9].

events using data from Twitter. In these papers, tweets were interpreted as representative opinions of a population. However, as only a few people in a population are using Twitter in order to broadcast their opinions, obtaining a valid sample of the opinion of all inhabitants of a country is more difficult than just sampling based on messages.

Nonetheless, the question remains whether Twitter users do form a sample of the population of a country like the Netherlands. One may answer this question by sampling data from Twitter in similar ways as described in [11–15], in order to compare attributes of Twitter users to general demographics of inhabitants of the Netherlands. Apart from the already mentioned "sample-stream" interface, another endpoint of Twitter's API was used as well. Though the "filter-stream" API, one may receive up to 50 messages per second matching a filter which can be either a search string matched against the text of messages or a bounding box defining a geographic region. This stream will continuously send tweets that match the provided filter. Therefore, the received messages do not form a random subset of all messages. By defining a set of rectangles that geographically span the Netherlands it is possible to filter all Twitter messages written in the country. This means that every tweet received through this API-endpoint is annotated by a GPS position.

The number of messages captured through the "filter-stream" is smaller compared to the general 1% sample. Whilst the "sample-stream" covers a subset of all messages sent all over the world, the "filter-stream" is limited to cover only the Netherlands. Additionally, only 0.7% of all messages are annotated with GPS information.

In total 61,361,500 tweets were collected through the "sample-stream" and 727,786 through the "filter-stream" in April 2013. Every tweet contains information about the message like the time, the text, hyperlinks in the text, the number of times it was retweeted, and information about the user such as the name, the location, the time and date the user created a Twitter account and the background color of the profile page. All these

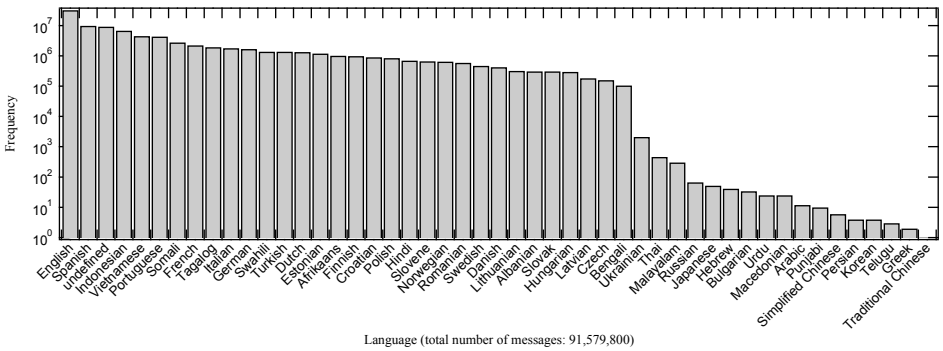


Figure 2.2: Distribution of languages detected in the messages obtained through the sample-stream API of Twitter, detected by the library "langdetect" [10].

tweets were written by a number of 1,645,526 distinct users.

In the second measurement, GPS-tagged messages within the Netherlands from the "filter-stream" API, between January and November 2013 were collected, through which 30,296,659 messages and 844,180 distinct user profiles were received.

The following methods summarize the possibilities of sampling messages from Twitter written by users living in the Netherlands:

1. Filtering messages received through the "sample-stream" by their language, assuming that citizens of the Netherlands are writing in Dutch.

2. Comparing the self-reported location provided in a user's profile to existing locations in the Netherlands.

3. Obtaining messages written in the Netherlands through the "filter-stream" API of Twitter delivers only those messages, annotated by a GPS position.

These techniques of filtering and sampling data from Twitter have advantages and disadvantages. For all techniques, the general question whether a user is a Dutch citizen or if someone just writes in Dutch or defined his location in the profile to be resident of the Netherlands, or has been in the country for vacation, remains. It is difficult to determine the real home town or name of a Twitter user as Twitter does not force its users to specify a correct location nor their real name.

Mislove *et al.* [16] analyzed messages and users in Twitter through the self-reported location in their profiles. By using the Google Maps API, the location string provided by 75% of the users, was converted into GPS positions. From these positions it was estimated that 8.8% of all users lived in the US. Mislove *et al.* found, by comparing their data to the U.S. census data, that Twitter users possess a "highly non-uniform distribution"[16, p. 2] in terms of their geographical location. The larger counties are found to be over-represented and in return, smaller regions were underrepresented by an order of magnitude. The sex, estimated by comparing the most frequent 1,000 first

names in the census data to names specified in Twitter's user profiles was strongly biased towards male users.

Even though the sampling of possible voters in a population is skewed, certain publications claim that predicting the outcome of elections is possible.

O'Connor *et al.* [17] describe that Twitter data can be used in order to estimate public opinions. By estimating the sentiment of used words in tweets, polarity values were created for concepts. These polarity values, when being multiplied to the frequency of messages containing topics of interest, were then compared to data obtained through polls about political opinions and consumer confidence. Using sentiment analysis, the number of positive and negative messages about different topics were counted per day and compared to the result of telephone surveys. The results show that the sentiment and magnitude of tweets follow the results of surveys with a Pearson correlation of 79%.

Tumasjan *et al.* [11] describe that counting messages mentioning political parties or their candidates reflected the outcome of German elections in 2009. Their data set was based on tweets collected one week before the election, in which the name of a political party or selected politicians appears in the text. The amount of Twitter traffic created by messages for the 6 main parties in the German election compared to the actual result had an average prediction error of only 1.65% and achieved therefore equally good results as classical (survey based) prediction methods.

Jungherr *et al.* [18] replied to Tumasjan *et al.* claiming exactly the opposite, that predicting elections based on word frequencies in tweets is not possible. By repeating the measurements of Tumasjan *et al.* they found very different results and show that the number of mentions of political parties does not reflect the political sentiment, nor future election outcomes. Jungherr *et al.* describe that the reason for the differences lies in the fact that the process of obtaining data from Twitter and the choice of political parties was not well described by Tumasjan *et al.*

Sang and Bos [12] reported that predicting the outcome of the elections in 2011 for the Dutch senate based on Twitter messages were possible. Their data set contained messages written in Dutch, acquired through a filter using high-frequent Dutch words. Using this technique they estimated to have sampled 37% of all Dutch tweets. The authors mentioned that Twitter is quite popular among Dutch teens which are not allowed to vote but they could not account for this fact because estimating the age of a Twitter user is a complicated task. The problem that people possibly write multiple messages about a political party was solved by just keeping the first message of every user. The tweets were then analyzed using manual sentiment analysis and defining polarity scores per party. By multiplying these "weights" to the number of tweets mentioning a particular party, equally good results as the ones obtained by Tumasjan *et al.* were achieved with an average prediction error of 1.45%.

Larsson and Moe [13] studied Twitter users during the 2010 Swedish election.They found that activity on Twitter correlates with mainstream media and that the most active users are part of the political sphere, using Twitter as a broadcast media. Due to this fact, one cannot truly answer the question whether messages on OSNs are reflecting the opinion of inhabitants of a country or are used by a few to try to manipulate the overall opinion. Therefore, Larsson and Moe conclude that political success cannot be predicted through data collected only from Twitter.

Gayo-Avello *et al.* draw a similar conclusion in [19], namely that data from Twitter "did no better than chance" for the elections for US congress in 2010. They tested the "predictive power of Twitter metrics against several races of the" US Congressional elections. In exactly half of the tests (for different states of the US), an approach using sentiment analysis was able to predict the outcome and in the other half, analyzing the number of tweets was "correct". Gayo-Avello *et al.* explain that the reason lies in the fact that the demographics of users involved in discussions about elections are nearly unknown and difficult to estimate.

Data from Twitter is lately also used to estimate more global systems like the stock market or box-office revenues to name two examples.

Bollen *et al.* [14] predicted the behavior of the stock market by estimating the mood of Twitter users. Based on a random sample of tweets they estimated the sentiment of every message in 6 different dimensions (Calm, Alert, Sure, Vital, Kind, and Happy), out of which the dimension describing how "calm" a message is, seems to correlate with the stock market. Also Asur and Huberman [15] claim to have succeeded in predicting the Hollywood Stock Exchange and box-office revenues based on a Twitter data set created by searching for messages containing movie titles. Their prediction was also based on the number of unique users writing about movies and a sentiment weight of messages achieving results with a very low prediction error.

As it is rather hard to find negative results about predicting certain events or the state of different systems, a general conclusion about the quality and applicability of predictions based on Twitter data cannot be drawn. On the one hand, as most of the previously mentioned techniques only claim after the events that predicting outcomes would have been possible and on the other hand do not answer the question about demographics of Twitter users, as done in this thesis.

The information from Twitter and Hyves was compared with two trustful sources of information.

1. Data from the Meertens Institute [20] containing a dictionary of family names linking to their geographic distribution and a list of first names with the number of men and women in the Netherlands having this first name.

2. A map listing data of districts and neighborhoods in the Netherlands from 2012, published by the Central Bureau for Statistics in the Netherlands (CBS) [21]. For every municipality in the Netherlands the map contains: the number of inhabitants, the number of male & female inhabitants, item the age distribution, the percentage of married/divorced/widowed inhabitants, the population density, the number of foreigners, the number of registered cars and motorbikes.

As the Twitter user profiles do not contain information about the age or the marital status one can only compare family names, first names, sex, estimated by the first name, and the location of users to general demographics. User profiles in Hyves do contain information about age and marital status additional to the name, the location and other interests. Using a technique as described by Nguyen *et al.* [22] may enable researchers to estimate the age of Twitter users. Nguyen *et al.* describe that machine learning techniques can estimate the age of a Twitter user as words and grammar used in messages

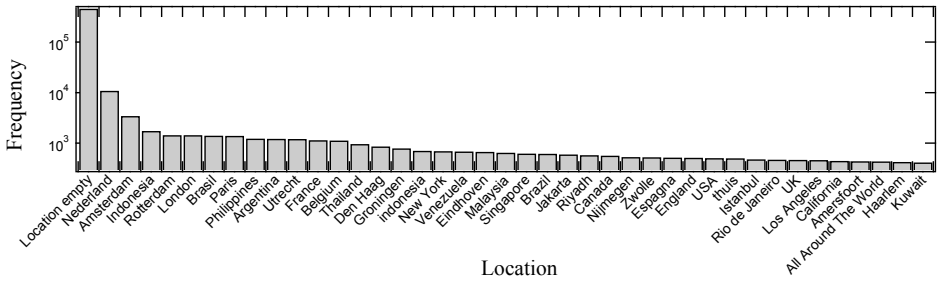


Figure 2.3: 40 most frequent locations provided in the "sample-stream" data set filtered by Dutch language.

of a person do change for different age groups. However, such techniques are computational expensive, need a lot of messages written by every user and training the detection system is not a trivial task.

GEOGRAPHIC LOCATION

The location of a Twitter user is specified in the user profile as a string that can be chosen freely or left empty. In Hyves the hometown of a user is always an existing place which was selected on a map if not left empty. As mentioned, "listening" to the sample-stream of Twitter and applying a filter that estimates the language of the tweet constitutes a prominent way of sampling users of a certain country. However, this technique has two major drawbacks. 1. Estimating the language of a message is a difficult task and there is no exact tool achieving this task. 2. When trying to sample inhabitants of a certain country, filtering messages by their language only works for languages spoken only in one country which means it is not applicable for tweets written in English, Spanish, French etc.

"Sample-Stream" Filtered by Language After applying a language detector [10] on the corpus of 91,579,800 Twitter messages, 1,340,963 tweets written in Dutch by 1,005,526 distinct users were found. Out of all profiles of these users, 487,156 list a value for the location, whereas all others left the location field empty. Comparing the provided locations to a list of municipalities and cities from CBS showed that 42,591 (8.7%) existed as regions or cities in the Netherlands and 10,849 users specified their location to be "the Netherlands" or "Holland". Figure 2.3 shows the 40 most frequently provided locations.

The frequency plot shown in Figure 2.3 contains Dutch cities like Amsterdam, Rotterdam, Utrecht and others but also names of different countries, cities in other countries as well as the string "thuis" (Dutch for "at home" on position 31) and "All Around The World" (position 38). When comparing the percentage of users who provided an existing position to the actual percentage of users living in this municipality (Figure 2.4) one observes that bigger cities in the Netherlands like Amsterdam and Rotterdam are underrepresented and 32 municipalities were not represented at all (white). This means no Twitter user in the data set set his location to be in these municipalities. The average absolute sampling error, defined as the sum of all absolute sampling errors divided by

the number of municipalities (415 in the Netherlands) was 7.92%.

One needs to memorize that this kind of sampling is affected by the quality of the employed language detection as well as the fact that most messages in Twitter are written in English. This means that even Dutch people, when using Twitter, are possibly writing English messages as followers of them might not understand Dutch especially as they are possibly distributed all over the world.

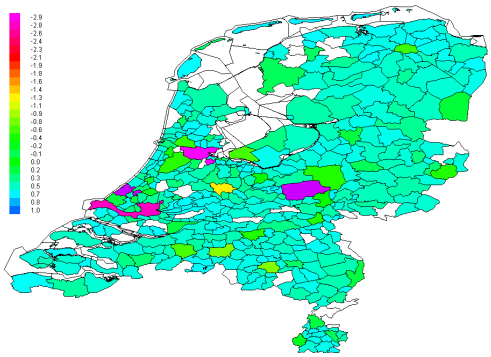


Figure 2.4: Municipalities in the Netherlands based on Twitter profiles. Colors indicate under- or oversampling in %, white indicates that no users from these municipalities were found.

Filtered by GPS Positions The self reported location in the Twitter user profile is not accurate for a high number of Twitter users as shown earlier. However, if a person tweets using a smart-phone, the current GPS location of the user can be attached to the tweet. By using the filter stream endpoint of Twitter's API one may filter geographical regions by specifying minimum and maximum GPS coordinates in order to receive all messages written within the area. Using the geographical filter, 30,296,659 messages were obtained, sent from the country of the Netherlands by 844,180 distinct users. Accounts of users that sent less than 10 GPS annotated messages was removed from this data, because not everyone who sends a message from a certain country is automatically an inhabitant of this country. For all other users, the most frequently found location of tweets, within an area of 2×2km, was estimated and interpreted as the home place of the person. The GPS location matched the location named in the profile for 30% of users (384,589 in the GPS-data set). This does not imply that the method of estimating the home town of an individual is not correct, but in most cases it did not correspond, bogus locations were defined in the profiles, as exemplified in Table 2.3.

Number of Users	Specified location
46	"onder je bed" (below your bed)
118	"earth"
301	"ergens" (anywhere)
317	"home"
452	"overal en nergens' (everywhere and nowhere)
1,224	"thuis" (at home)
140,257	did not specify their location

Table 2.3: Bogus locations specified in user's profiles.

Additionally, 16,401 users specified "the Netherlands" or "Holland" as their location in the data set filtered by GPS positions. Figure 2.5 compares the distribution of the number of inhabitants reported by CBS (green), the number of inhabitants estimated through the GPS-position attached to Twitter messages (blue) and the self-reported location in the user profile (red).

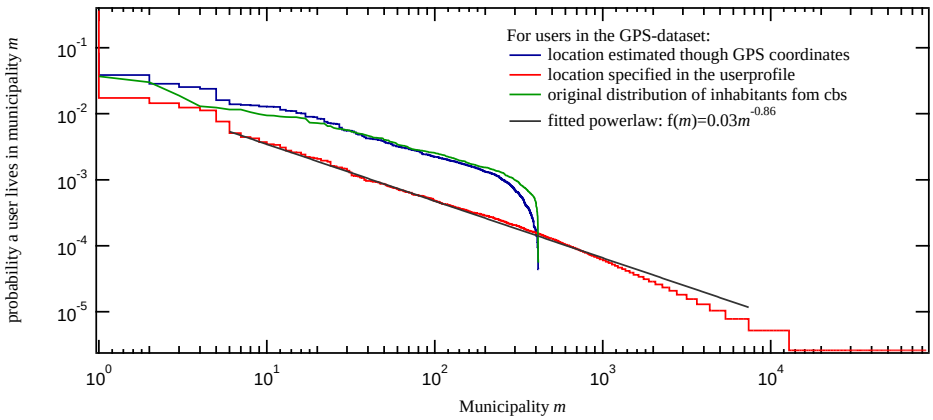


Figure 2.5: Comparison between the distribution of the number of inhabitants reported by CBS to estimated positions from GPS-tagged messages and the self-reported location in user's profiles.

Clearly, the self-reported locations do not match the actual distribution of Dutch inhabitants, whereas the locations estimated using GPS filtered data are closer to the real distribution.

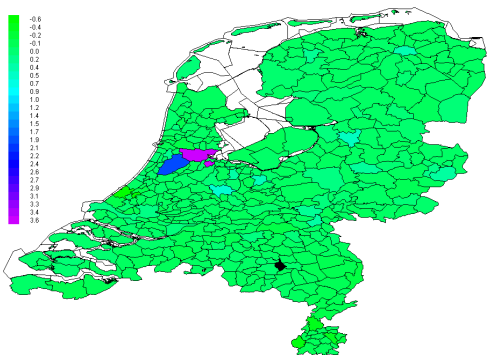


Figure 2.6: Municipalities in the Netherlands based on Twitter profiles. Colors indicate under- or oversampling in %.

Comparing the number of inhabitants of dutch cities to the estimated number of users showed that 6.2% of all inhabitants of the Netherlands are sampled with a coefficient of determination (R^2, which equals the square of the Pearson correlation coefficient between the real number of inhabitants and the Twitter users that reported to live in the municipality) of 0.76.

Figure 2.6 shows that the technique of using GPS filtered data to estimate the home place of a person, over-sampled the largest city, Amsterdam and the surroundings. The oversampling is likely due to the capital being an attractive place with lots of museums, restaurants or other places from which users tend to tweet. The biggest airport of the Netherlands (Shiphol) is located in the municipality to the southwest of Amsterdam which is slightly overrepresented using this technique.
The average sampling error when using GPS filtered data and the home place estimation is 0.094%, and therefore significantly (ca. 100 times) smaller than the one reported for data from the "sample-stream" using language detection which was 7.9%. For locations

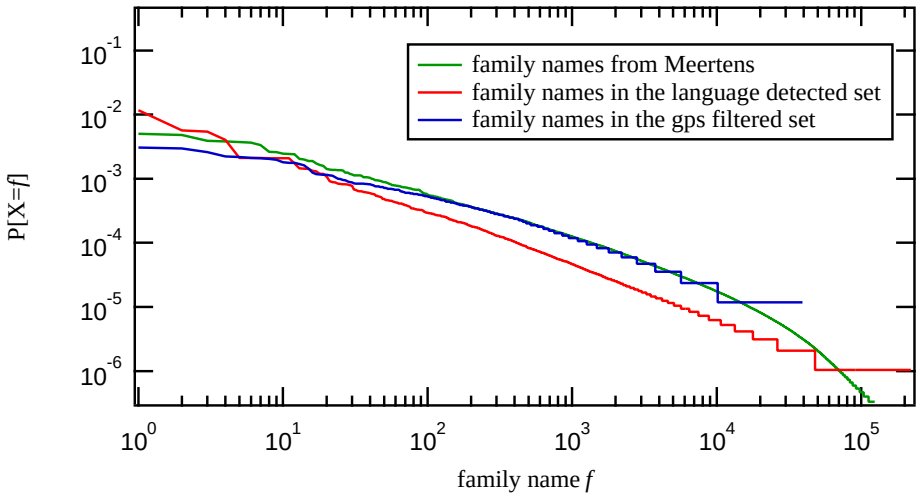


Figure 2.7: Distribution of family names in the data set of the Meertens Institute(green), Twitter sample #1(red) and Twitter sample #2 (blue).

reported in Hyves profiles, the average sampling error is 1.34% for 922,963 users that reported a location out of the 2.7 million in the data set.

DISTRIBUTION OF FAMILY NAMES

Every user profile lists next to the unique user-name also a field where the real name can be filled in. One may interpret the last string in the name field of the user profile as the surname and the first string as first name, in order to analyze the family and first names of Twitter users in the Netherlands. Through the "sample-stream", 220,369 distinct family names sampled via listening and detecting the language of messages were found.

Out of all provided names, 1,411 were left empty, 388,373 users provided only one word as their name, 420,212 two, 102,675 three, 24,7432 four, 5,117 five, 1,586, six, 1,050 seven, 662 eight, 442 nine, 279 ten. The reason for this high number of words is based on a high number of users writing their first names with spaces between every letter, adding symbols to the name or filling in bogus names. For example the name "S A M A N T H A :)", but also "One In A Million" or "We are the champions" were found as names.

The whole set of Dutch surnames as published by the Meertens Institute contains 123,990 names visualized in Figure 2.7 (green) as the probability density function. The distribution of names found in the Twitter sample filtered by language detection (#1), also shown in Figure 2.7 (red), shows a clear deviation from the original one.

Out of the existing names, 30,364 existed in the data the Twitter sample #1. The sampling error for the existing names, defined as the area covered by the complement of the intersection of both distributions was 53.76% for the language detected sample. In the GPS filtered data (#2 blue), 17,951 names were found to exist in the Meertens data, out of 39,296 distinct provided ones. The error between both distributions equalled to 6.6%.

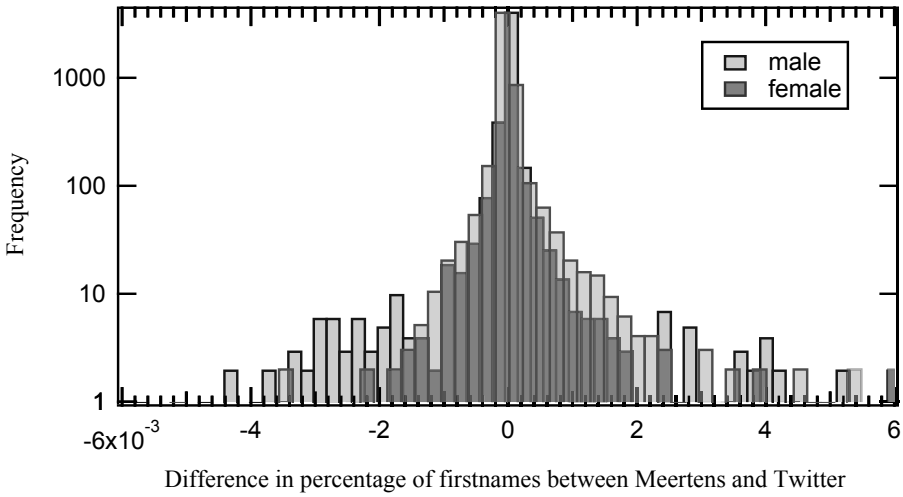


Figure 2.8: Distribution of sampling errors of first-names in the GPS sampled data set compared to the distribution given by the Meertens Institute.

DISTRIBUTION OF FIRST NAMES AND THE GENDER

By interpreting the first string in the name field of the user profile, as the first name, 406,653 distinct first names were found in the "sample-stream" filtered by language and 21,637 in the GPS filtered sample. The "ground truth" from the Meertens Institute contains 108,941 first names that appear at least 5 times in the Netherlands. For all these names, the Meertens Institute reported the number of male and females having this name. Comparing the frequencies of names denote that 0.5% of males and 0.6% of females in the Netherlands were found with a R^2 values of 0.12 and 0.17 in the "sample-stream" filtered by language. Surprisingly, for the GPS sampled data, 0.7% of males and 0.3% of females were found having an R^2 value of 0.54 and 0.58. The higher the R^2 value (up to a maximum of 1), the better the sample of Dutch inhabitants. Figure 2.8 depicts the sampling error. The names reported in the Hyves.nl user profiles summed up to 148,114 first names out of which 57,757 are also in the Meertens data set denoting a sampling of 2% of the inhabitants with a R^2 value of 0.11.

As the sampling of locations shows only little error, the first-names show a larger deviation from the ground-truth. A reason which might be the possible skewed distribution of the age of users of online social media.

AGE AND MARITAL STATUS

Twitter users are free to enter a description of them selves in their profile which denotes that a small fraction of users provide their age. Unfortunately only 0.007% of all users in the used data set provided their names which denotes only a marginally small amount, not sufficient to analyze the data.

Nguyen *et al.* [22] tried to estimate the age of Twitter users using machine learning techniques. They employed annotators to build a test set of 3,100 Twitter users from

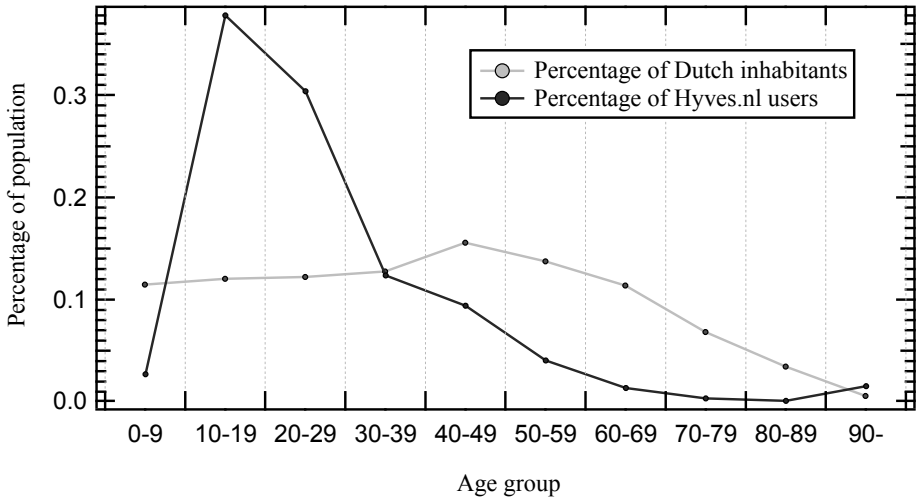


Figure 2.9: Comparison of the age of Hyves.nl users (dark gray) to the population of the Netherlands (light gray).

the Netherlands estimating the age of users by checking the description, provided in the Twitter user profile, Tweets, Facebook or Linkedin profiles if available. Their way of finding dutch users was by searching the Twitter API for common dutch words finding that 60% of the users had an age below 20, 26% between 20 and 40 and 14% had an age above 40 years. A finding that correlates with the age distribution obtained from the Hyves.nl data set. The age distribution of users in Hyves.nl shows an over-sampling of young persons when comparing to the age of the Dutch population as shown in Figure 2.9.

In the used data set 161,676 Twitter users connected their Hyves.nl account to their Twitter profile. Therefore the distribution of approximately 1% (1,588 Hyves user-accounts) of these users was possible to obtain because in Hyves also only a small number of users provide their age. This distribution reflected the one depicted in Figure 2.9(blue) quite well. Due to these findings, it is reasonable to assume that in terms of the age distribution, most sampling techniques of Twitter data will fail. A point especially crucial when trying to estimate the outcome of elections because a large fraction of sampled users might be under age in order to vote.

Similar findings can be reported for the marital status which, when searching profile information for the words married or getrouwd (Dutch for married), revealed that only 0.08% of all users in the Twitter data set mentioned this word in their profile information. As it would be completely wrong to simply assume that everyone else is either single, divorced or widowed even though the distribution of ages might explain this low value the approach of using Hyves data for further assistance could be valid. Figure 2.10 depicts the percentage of married users per age group taken from CBS data (shades of blue) compared to information obtained from the Hyves.nl data set (shades of red). The curves

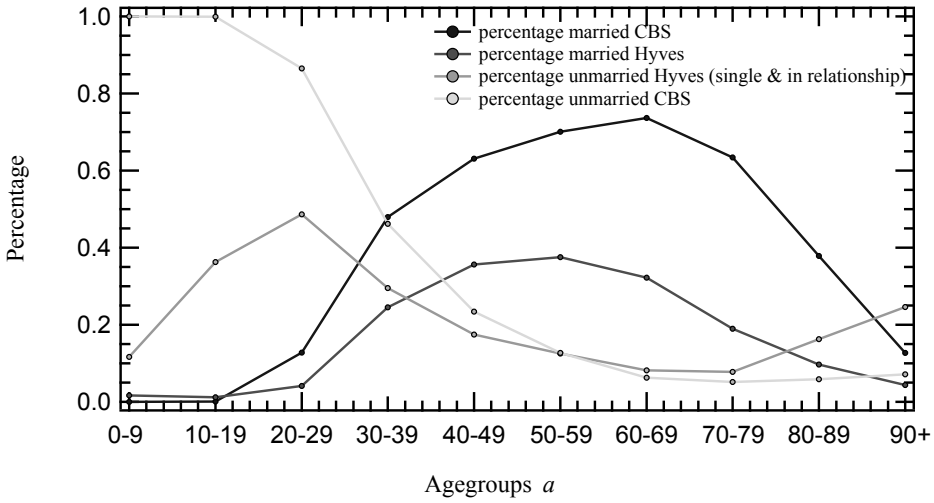


Figure 2.10: The age of Hyves.nl users to their self reported marital status compared to data of the central bureau for statistics in the Netherlands (CBS).

denoting the percentage of married individuals exhibit a similar shape but it seems that in reality there are more people married than claimed in the OSN.

2.1.2. INTERESTS OF USERS

Out of the 19 topic groups in Hyves.nl, some further strengthen location estimation. Topics like hangouts, schools, colleges, clubs, companies, food and sports contain implicit location information. Assuming that people like to visit bars, restaurants, sport clubs in the same city they live and work enables us to infer the city from these groups.

By using Bayesian analysis [23], the probability a user has joined a specific group given he lives in a specific city can be calculated. If the resulting distribution shows no significant peaks (larger than 1 standard deviation), this means that the users in this particular group are homogeneously distributed in the Netherlands. An over-representation of a particular group in a city however is a good indicator that this group can be used to infer a users city. In total 13,512 groups were identified that can predict the residence of a user. On average 64% of the members of the found groups live in the same city. This does not imply that the other 36% reside in different cities as some users simply do not provide their home town. When assuming that users who did not enter a city in their profile would live in the same city as most users of this group, the average predictability increases to 86%.

DIFFERENT TASTES IN AGE GROUPS

Groups do not only reveal location information but also insights about the age of a user. For example, musical interests have a strong correlation to the age of a particular user. Figure 2.11 exemplary depicts the age of users who like different singers or music bands.

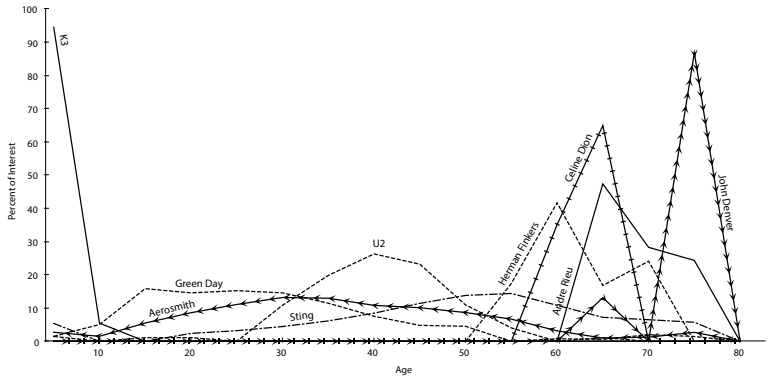


Figure 2.11: Probability users have a specific taste in music to the age of a user.

Conversely these correlations suggest that the specified age of most users in the Hyves.nl data set is accurate. Strong relations between interests and the age of a user were additionally found for movies, music types and game consoles.

In order to efficiently identify relations as depicted in Figure 2.11, association rule learning can be employed on a data set of publicly available information. Though this technique is able to unveil correlations without the need of individual analysis of interests towards age groups as explained in the following section.

ASSOCIATION RULES

Association rule learning is a popular method used in data mining in order to discover relations between attributes in data sets. Often utilized for market basket analysis, the input data set for association rule learning contains an item set of things a person has bought. A typical rule created out of a supermarket data set could therefore be the following: If noodles and cheese are bought then the customer will also buy bolognese sauce with a confidence of α percent where all products appear in β percent (support) of all purchases. The confidence α corresponds to the fraction of the support of all items in the rule to the support of the requisites. The naive way of calculating simple co-occurrences would result in a very large co-occurrence matrix because of ca. 1.1 million groups in our data set. Given the groups of all users as input, association rule learning will still calculate rules in a reasonable time, for a given minimum support and confidence.

An implementation called apriori [24] was used to calculate association rules with a given minimum support of 0.1% and a minimum confidence of 50%. The exact number of groups in our data set was 1,115,558. The support of 0.1% means that 1,116 user profiles should list a group in order to include the group in the rule. The calculated rules had a maximum length of 4 resulting that at most 3 groups lead to a consequence. Longer rules are clear subsets of shorter ones having a higher confidence but smaller support. An example for such a rule is the following. Users that are interested in the soccer club “Ajax Amsterdam" are also interested in the “Amsterdam Arena" with a support of 0.203% and a confidence of 58%. But if a user is interested in “Ajax Amsterdam" and “Adidas" he is more likely to be interested in the “Amsterdam Arena" with a confidence of 83% but

the rule has a support of only 0.113%.

As it is possible to set the privacy settings for groups to only show groups out of selected topics, association rules learning helps to infer others. By knowing only a few groups of a user it is possible to directly apply a rule with a high confidence to infer other groups of the user. Lets take the example of the “Ajax Amsterdam” fan again, were it is already known that he likes the “Amsterdam Arena” with a confidence of 58% and the stadium, the soccer club is playing in, with a confidence of 72%. Additionally one can estimate that he will also like different brands like “Nike” (confidence of 53%) or “Adidas” (confidence of 54%) or general terms like “soccer" (confidence of 56%). As mentioned earlier if combinations of these groups are found the confidence increases.

The same holds for the earlier mentioned age prediction as shown in figure 2.11 based on different groups. For example the probability to be of age 11 if the movie “Finding Nemo" was liked is 70%. Knowing that the user additionally likes the movie “Happy Feet", increases the probability to 87% as the rule gets more specific.

Interestingly the given example of soccer fans already depicts that group predictions work across different topics (sports to brands to hangouts).

It is visible that most rules are between groups of the same topic (same color). For every topic there seem to be a few hubs standing for the largest groups in this particular topic that can be predicted by multiple other smaller groups.

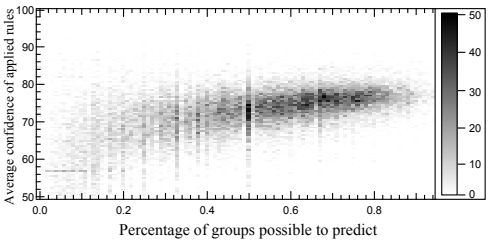


Figure 2.12: Joint 2D histogram of the percentage of groups that can be revealed using association rules versus the average confidence of the applied rules. The color indicates the number of users the rules apply to.

As association rule learning seems to be a good solution to obtain global information about group predictions, although it is not a user-centric method. This means it is not possible to observe effects of the underlying topology of the friendship network. This predictability is defined by two values. One is the number of groups that can be inferred using all rules whereas the second is given by the average confidence of rules applied to all groups of a user. The latter gives insights into the “predictability" of this user. Figure 2.12 depicts the predictability versus the fraction of predicted groups.

The Pearson correlation of the age of a user towards its predictability is slightly negative with -0.15, which in turn is based on the fact that the number of users in our data set decreases for older users. As previously shown, groups may have a certain dependency on the age which means that the groups older people follow do not reach the required minimum size of 1,116 users to be included as a result of association rule learning.

To which extent data from OSNs might be used to estimate the mood or opinion of a population in terms of different concepts or even to predict events, remains unclear and should be focus of future research. especially if one considers that most OSN systems are used to broadcast media whereas messages might reflect opinions but are also used to “influence” individuals. Analyzing data reflects therefore an “hen-egg problem” as an analyst, nowadays, can not infer the underlying intention of a user broadcasting media.

2.1.3. BEHAVIORAL ATTRIBUTES OF USERS

Behavioral attributes of a user are typically values describing an individual by traces, left on a system, like the time someone logged on, the duration a person staying online or the quantity of activity monitored through a server. Technically an individual using a certain online service, or an OSN, can be described by these kind of measures. For example in Twitter, if an individual composes messages attached with a GPS position, a trace of all locational information is enough to uniquely identify the person. Driven by a publication of Locard [25] from 1930 who showed that 12 points are sufficient to uniquely identify a fingerprint, Montjoye *et al* described in [26] that 5 geographic points are enough to uniquely identify users in a data set of 1.5 million users of a mobile phone operator.

The behavior of a user can be captured through the activities that appear if one knows the time a user has used a certain service. Figure 2.13 visualizes the activity of 100 randomly sampled users from the Netherlands, based on the time of day the person wrote a message.

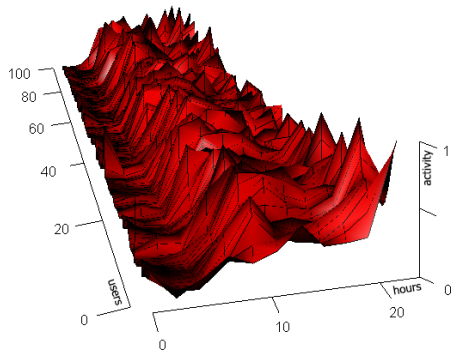


Figure 2.13: Visualization of hourly activity of 100 randomly sampled users from the twitter data set. Abscissa denotes 24 bins which correspond to hours, the ordinate normalized activity in this hour, applicate (z) axis denotes 100 randomly chosen users.

Most people are using the service clearly throughout the day, as there is only little activity recorded during the night hours.

In September 2013, a Spanish parliamentary report[4] claimed that Spain is less productive because it is mainly part of the central European timezone (GMT+1) whereas it should be part of the western European timezone like the canary islands, which belong to Spain, the United Kingdom or Ireland which are on similar longitudes.

Indeed, when analyzing the times at which users from the eastern part of the central European timezone, within the continent of Europe (Poland, Slovakia, Hungary, Serbia, Croatia, Bosnia and Herzegovina, Montenegro, Albania, Kosovo, Macedonia), are writing messages (Figure 2.14 (green)), one observes that during night hours, between 1am and 7am relatively small traffic is produced. For western Europeans, Spanish and French inhabitants (Figure 2.14 (red)), this time of small activity is between 2am and 8am, which indicates that, although people are living in the same timezone, different behavior exist.

When calculating the difference of the area between the two distributions within the night (22pm to 5am) and the morning hours (5am till noon 12pm), a difference of

[4] "Informe de la subcomisión creada en el seno de la comisión de igualdad para el estudio de la racionalización de horarios" (Report of the sub-commission created as part of the equality commission for the study of the rationalization of timetables) http://ep00.epimg.net/descargables/2013/09/26/ed87c0772aeb2b9406fa383995b93026.pdf

-0.03627 can be measured[5], which denotes an offset of 52 minutes. Although it is unclear how the time was estimated in the previously mentioned report, the result of comparing the twitter activity pattern is amazingly similar to the reported 53 minutes. This simple measurement of user behavior therefore might confirm the findings that Spanish people sleep less (during the night) than inhabitants of countries in the eastern part of the central European timezone.

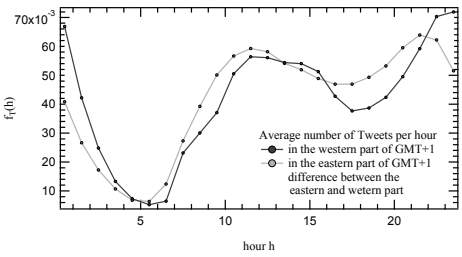


Figure 2.14: Activity of Twitter users based on 22 million messages. 3.6 million in the east (201,269 users) and 18.4 million in the west (697,161 users) of the central European timezone.

In a more in depth analysis of the activity patterns of users one may ask for different groups of users to avoid the Yule–Simpson effect [27] which describes that an observed trend in different groups will change if the results are combined. Therefore two different approaches were chosen to analyze clusters of user behavior. On one hand the k-means algorithm was directly applied to the histograms defining user behavior and on the other hand the two data sets of eastern and western Europeans were converted into similarity matrices in which every entry of a matrix denotes the similarity of two users based on different similarity metrics and spectral clustering was applied to these similarity matrices to identify clusters of users having similar behavior. The used similarity metrics were: the cosine similarity, the chi-square similarity, the earth movers distance, the geometric distance and the Pearson correlation which were all normalized to have a high value if the compared signals are similar and small if the input is dissimilar.

Figure 2.15 depicts exemplary the results of the k-means algorithm for the two data sets. As the number of clusters is an input argument to k-means, the elbow method, comparing the sum of squared errors for different numbers of clusters, was used to estimate that 6 clusters are an appropriate value. When comparing the clusters in Figure 2.15, one may notice that they exhibit similar distributions which may denote that certain user groups exist across boundaries. For both cases, clusters containing users who are active mainly in the morning (around 10am), around lunch (between 11am and 2pm) as well as in the evening (between 10pm and 1am) exist.

Closer inspection of the clusters describing the behavior of users mainly active in the morning showed nearly no difference between users in the western and eastern part of the central European timezone because the peak visualized in the two diagrams at the top in Figure 2.15 starts to ascent and peaks around the same time for users irrespective of their origin. Similar finding can be reported for other clusters. Within the the analysis of separate clusters of user behavior it was therefore not possible to measure the earlier mentioned “missing 53 minutes” because the signals within the analyzed clusters do not differ significantly. This finding, contradicting the earlier one shown in Figure 2.14 which showed that there exist a difference, can be explained by differently behaving groups of

[5]The difference is defined as the area between the two probability distributions, where a value of 1 would denote no overlap of the distributions at all.

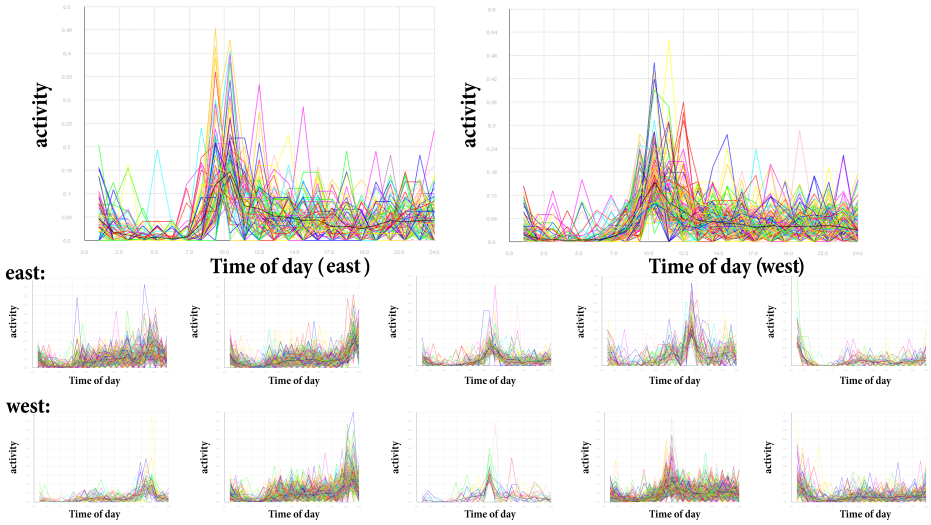


Figure 2.15: Clusters of user behavior based on k-means. The diagrams visualize clusters of user behavior found for users in the eastern and western part of the central European time zone. Different colors denote different users. The black lines depict average values within the clusters.

users. To which extent the group of working citizens is reflected cannot be validated as it might be possible that Twitter users are in an age group in which they are more flexible to chose their working hours.

2.2. RELATIONS OF AN INDIVIDUAL

As shown in the previous section, the attributes of users state important data for egocentric network analysis. The second important part is described through the analysis of relations between individuals. It is shown that attributes of individuals like their opinion, certain habits (smoking, etc.), obesity and happiness [28] are "contagious" and propagate along friendship relations. Such a statement sounds obvious, because most people are educated by their parents, teachers and behave according to social norms. But how far does one influence others or by how much is one influenced by others, and is there eventually an effect onto one self by unknown persons, like friends of friends etc.? These kinds of tasks can be approached by modeling and measuring the interrelations of an ego and all friends, acquaintances or family members of the ego. In the language of graph theory, a relation is called a link and a user is typically modeled as a node. All alters (nodes) having a direct relation to the ego are also called the direct neighbors of the ego.

In egocentric networks, the ego has by definition the highest degree and in most egocentric networks the ego is the only node with this degree. In the example given in Figure 2.1, which states again the egocentric network of the author of this thesis within the OSN Facebook, the ego has a degree 68. For all egocentric networks the diameter, defined as

the longest out of all shortest paths from every node to all other nodes, is at most 2. Also, all egocentric networks are by definition connected.

Expressing a network as adjacency matrix A is a common and simple way to store graphs. In such a matrix, an entry is 0 if the two nodes are not connected, or 1, if the nodes are connected. Table 2.4 shows an exemplary adjacency matrix of an undirected network.

	Ego	Alter I	Alter II	Alter III	Alter IV
Ego	0	1	1	1	1
Alter I	1	0	0	1	0
Alter II	1	0	0	0	1
Alter III	1	1	0	0	0
Alter IV	1	0	1	0	0

Table 2.4: Adjacency matrix

In most situations the strength a relationship towards or between alters is important, real numbers may be used in the adjacency matrix as well, expressing how close the friendship is, how long the relation already exists, how similar the two persons are to each other etc. Table 2.5 shows a weighted adjacency matrix, where the weight denotes the cumulative amount of communication taking place during one week in hours.

	Ego	Alter I	Alter II	Alter III	Alter IV
Ego	0	4	2	3	1
Alter I	4	0	0	6	0
Alter II	2	0	0	0	8
Alter III	3	6	0	0	0
Alter IV	1	0	8	0	0

Table 2.5: Matrix expressing the strength of relations.

In online social networks, the links are usually based on friendship relations. If the information about social interactions is not available, links can also be created artificially. Such artificial edges are then typically based on some similarity metric. If one considers for example the database of an online shopping service, then the links may connect pairs of users that bought the same or similar items. If social relations were known to the service, such recommendations would improve dramatically because friends influence each other and have similar tastes, a theory that is called homophily[6]. In marketing homophily is the basis for viral marketing which makes use of word-of-mouth spreading, namely the fact that customers will convince their friends to buy a certain item without a marking party interfering with the process. When observing how strong the relation of a ego towards alters is, companies may determine whether a customer can be deemed “influential" and should consequently receive better treatment than others [29]. Information on relationships, personal habits and interests can be taken into account when

[6] Homophily describes that similar individuals tend to bond with each other. It is sometimes also referred to as the “birds of a feather, flock together” effect.

assessing risks and rates when applying for health insurance [30], and face recognition performed on photos stored in online social media allows the re-identification of persons in other contexts, such as identifying passersby in camera recordings to deliver targeted billboard advertisements [31].

McPherson *et al.* [32] and Blenn *et al.* [33] showed that friends in OSNs like Facebook and Hyves, Digg, Twitter and other OSNs do share multiple attributes like their age, taste in music etc. and live close to each other.

As such technologies are developed and applied, concerns about the privacy of one's personal data are increasingly gaining track. Indeed, privacy filter usage has become a mainstream practice: in case of the largest national social network site in the Netherlands in 2012, Hyves.nl, 63% of the users had enabled privacy settings in their profile making their details invisible to the general public.

Personal information can actually be reconstructed from a social network's friendship graphs. The underlying justification our approach is driven by is the sociopsychological hypothesis, which was empirically verified for Digg.com [34] and Facebook.com [35]. Users form social ties with those around them who are similar in socio-economic status, interests and opinions [36]. In consequence, knowing a user's friends can therefore to a large degree tell the individual tastes and choices of a social network user even when his profile page is hidden.

The degree to which this technique can be successfully applied varies with the overall embedding of a particular ego in the social graph as well as other attributes, such as the ego's personal characteristics, the overall diversity of alters or the degree to which the friends are making use of privacy settings themselves.

Two major approaches, active and passive, are possible to access private information. Active approaches try to obtain data by directly attacking a particular user using fake profile information [37], surveys or third party applications that access the users profile in the OSN. Passive approaches are based on statistical analyses of users and the friendship network. These passive approaches may be based on the profile information a user specifies, tracking the friendship network through third-party applications, or the combination of different data sources.

Gross and Acquisti [38] analyzed patterns of information revelation in OSNs and privacy implications in the "early" stage of Facebook. An amazingly high number of 89% of users in their data set provided their real name. Other attributes like phone number, birthday, home town, address etc. were also given by the majority of the users. Different techniques to infer private information like re-identification of users by analyzing the postal code and their birthday are presented. Face recognition to identify users on different sites or even identity theft of the users social security number was shown to be feasible.

The role of third party sites in tracking users of OSNs and obtaining private information is investigated by Krishnamurthy and Wills [39, 40]. In most cases, a user has no possibility to control all applications that track profile data. Users are not aware which data is accessed by them and what the different services do with this data.

Based on the knowledge about friendships in OSN and the fact that those relations are mostly built between individuals having similar interests it is still possible to infer private attributes of a user from his friends even if the user has a profile which is not

visible to everyone. McPherson et al. [32] discussed “homophily" as a concept that limits individuals to connect only to others having similar attributes. The strongest divisions are based on race and ethnicity followed by age, religion, education, occupation and gender. Hence, ties between non-similar users are either not constructed or dissolve at a higher rate. This leads to social niches in the social space.

Aristotle already noticed three kinds of friendships [41]. Younger people seem to seek friendships for pleasure “for they live under the guidance of emotion, and pursue above all what is pleasant to themselves". Older people tend to form “useful" friendships such as companionship, help, guidance, among others. The third kind, according to Aristotle, is the perfect friendship which is the relation between individuals “who are good and alike in virtue". In general, he defined in all cases that people like to bond with people who are alike in specific characteristics or interests.

He et al. [42] constructed a Bayesian network assuming that direct neighbors have a higher overlap than users multiple hops away. It is shown that privacy can be indirectly inferred via social relations and mathematically over multiple hops. He et al. use an influence strength which is defined as the conditional probability ($P(A|B)$) that user A has a attribute given a friend (B) has the same attribute.

By using friendship information and group attendance information, Zheleva and Getoor [43] showed for different OSNs that it is possible to infer private attributes using group and friendship information.

Mislove et al. [35] claim that “you are who you know" because automatic community detection for multiple attributes of the users led them infer private attributes with an accuracy of 80% inside those communities. This approach needs the knowledge of the topology of the social network in order to detect communities. Due to the dynamic nature of OSNs, standard crawling techniques take rather long to obtain the whole network, it is thus unfeasible for attackers to first crawl the network in order to detect communities.

If a user has a private profile page it is still possible to uncover friendship relations because in Hyves, similar to Facebook and other OSNs, relationships are bidirectional. Therefore, an ego’s name is listed on profile pages of friends having a publicly viewable profile page. Based on the average number of friends in Hyves, every user should have 42 friends with a publicly viewable profile page. As stated in Bonneau et al. [44], “eight friends are enough” to reveal the whole network of users of a OSN.

2.3. “BIRDS OF A FEATHER"

As described by McPherson et al. [32], friends tend to have similar interests because they know each other, live close to each other, meet physically at places where they follow their hobbies or at places where they work together. Friendships in Online Social Networks do not necessarily follow this scheme as one may also create friendship relations towards users without knowing them in person. The hypothesis that personal preferences limit the possible number of users, still holds as online friendships are based on common interests as shown in [32, 34–36].

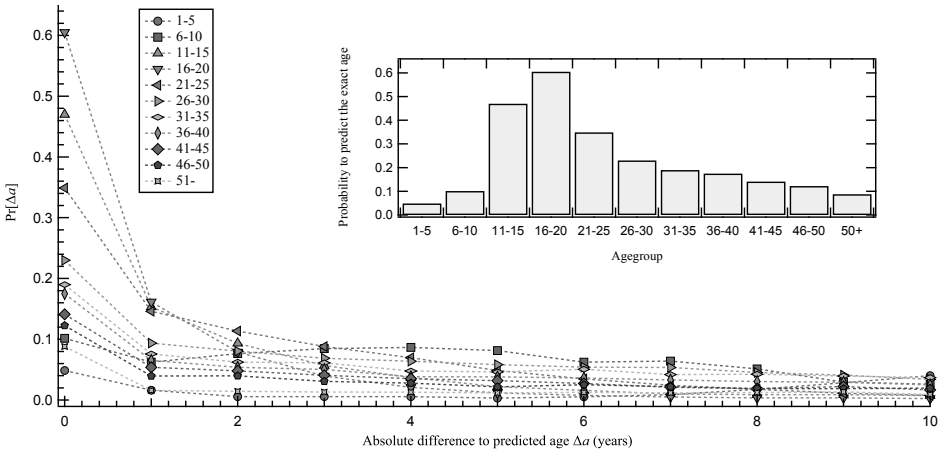


Figure 2.16: Prediction of the age of a user the using mode of his friends age for different age groups. Inset: Probability to correctly guess the age of a user.

2.3.1. AGE OF FRIENDS

One assumption is that a user is as old as most of his friends. Hence, for every user in the Hyves.nl data set providing an age, the most frequent age of his friends was used as an estimator and compared to the actual age. The results are shown in Figure 2.16 by the difference between actual age of a user to the mode of the friend's age.

As indicated by multiple traces (different markers, and colors) in Figure 2.16, the probability that most friends have the same age as a user is depending on the age group the user is in. The highest accuracy of this method (prediction rate) is found for the group of 16 to 20 year old users where 61% of friends have exactly the same age as the user. When allowing up to ± 1 year of difference the probability to predict the correct age of a user, increases to 77%. This prediction probability decreases for older age groups.

A reason for this high age overlap might be based on the fact that friendships in the group of 10 to 20 year old users are created in schools, where students are in the same class. Later in life, colleagues and friends are not exactly the same age anymore. Another explanation might be the decreasing average number of friends: In the group of 16-20 year old's, the average number of friends is 81 whereas the group of users at the age of 46-50 years have on average only 14 friends within the OSN Hyves.nl.

2.3.2. LOCATION DETERMINATION

Another intrinsic value of a user is the hometown. Again, the simplest assumption is that most friends of a user live in the same city the ego lives in. Actually, based on our data set, 67% of an ego's friends who provided their hometown live in the same city as the user, and 91% of all alters within Hyves.nl live within 50 km of the users location.

As the usage of Twitter.com differs from OSNs like Hyves.nl, because it is used as a broadcast medium, friends are not living that close to each other. Calculating the distance to all followers of 1,000 randomly selected Twitter users from the Netherlands by mapping the self-reported location using the geonames database [45] to GPS coordi-

nates led to the result that only 0.64% of followers within Twitter live in the same area. However, when counting the occurrence of followers locations, the resulting distribution peaks at the ego's position. Jurgens [46] showed that the precision in estimating a user's position can be increased by using the locational information of only mutual friends. This estimation can even be increased according to Jurgens, by using the location of users that were frequently mentioned in tweets.

Every user joined on average 26.6 groups. The concept of homophily suggests that friends have similar tastes which should result in a high overlap of group memberships between a user and his friends. If all friends are taken into consideration only a small overlap can be found. Amazingly most users have at least one friend who joined nearly the same groups as the user. The difference in groups a user is a member of compared to his friends can be seen as a similarity measurement between users. Based on this metric, only a fraction of all friends are close friends, whereas a high number of acquaintances appear in a user's friendship network. The fact that only a few friends in the friendship graph are close friends is also described by Granovetter [47] and analyzed by Mcpherson *et. al* [48]. Thus a way of identifying close friends could be by analyzing the information if the users are tagged on the same image. This would imply that the users physically know each other and the probability they are close friends increases.

By comparing the predictability of a user with the predictability of friends, a positive correlation of 0.29 was found, stating that a fraction of the friends of a user with a high predictability are also predictable. By correlating the number of friends that have a publicly viewable profile to the predictability, no significant relation was found, which means that a small number of friends having an open profile are already enough to guess the groups a user attends. If such techniques can be employed for users multiple hops far away through the representation of the observed (known) friendship network as Bayesian network where links are annotated by the "predictability" remains for future work. However, given the rather low correlation of 29% for friends' predictability shows that attributes of users who are multiple hops "far away" is only possible to a limited extend.

Further on, commonly used smart phone services upload the users' phone book to their servers, in order to connect directly to all persons in the list, who are also using the service. As this method states a quite convenient way of connecting to other users, a possible attacker having access to the database, is able to reconstruct private attributes of persons who are not using the service at all. Especially as everyone needs to trust the ability of a friends service provider, and the friend to keep the friends data secure. Obviously such an assumption can only be fulfilled if every person who's friends are using OSNs, trusts these friends and acquaintances and the services.

2.4. NEIGHBORHOOD OF AN EGO WITHOUT THE EGO

When analyzing egocentric networks it is sometimes useful to remove the ego, as he is connected to any alter. The appearing structures describe how alters are connected with each other and may therefore provide a view onto communities the ego is part of. Creating such a network out of the example shown in Figure 2.1 leads to the unconnected graph depicted in Figure 2.17. Multiple communities (indicated by shades of gray) can be observed which resolve to groups like school friends, friends and collaborators from

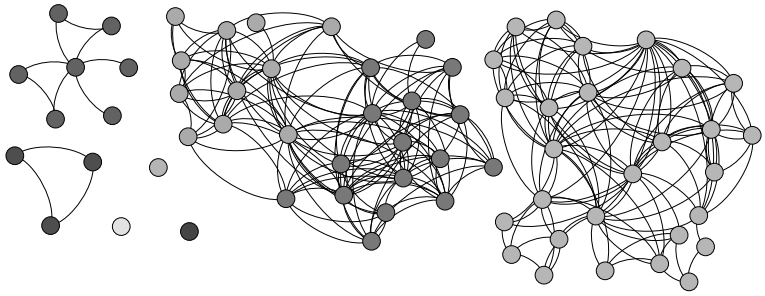


Figure 2.17: Facebook friendship network of the author. Shading represents different communities.

university, colleagues and friends from work, family members and acquaintances of the author. This observation depicts that the ego is part of at least 5 communities and has 3 acquaintances within the social graph of Facebook. A description for communities can be accessed through regions the users lived in, the school, university and companies he worked for, as well as the strength of relationship like family, friends and acquaintances.

Burt [49] was one of the first to analyze missing links between alters and called them, "structural holes". The basic assumption of structural holes states that the lack of ties among alters may benefit an ego. If alters are not connected, Burt assumes that the ego is more autonomous and has more control especially in terms of information spread. Structural holes will provide novel information to the ego, because the ego-node is connected to multiple communities being itself connected because of different reasons. The ego is therefore a broker, exposed to different ideas from different communities. Burt used data taken from discussions among managers in a large electronics firm to prove the correctness of the hypothesis. Burt proposed certain metrics to capture how expressed the structure around an ego is.

Redundancy of alters: If alters are connected to each other and the link between the ego and the alters have the same weight, the information the ego may obtain through connected alters won't differ that much. Therefore the alters are redundant. The redundancy of one alter is defined as the number of links towards other alters, divided by the total degree of the ego. The total redundancy of the network for all nodes N with an ego e is defined by the degree sequence D as:

$$R = \sum_{i=1}^{N} \frac{d_i - 1}{d_e}, i \neq e \tag{2.1}$$

The non-redundant portion of the ego's network, also called the *effective size* is defined as:

$$E = d_e - R \tag{2.2}$$

The effective size is maximal if the alters are not connected to each other. An extension of the redundancy for weighted graphs is straight forward. The *efficiency* of an ego is further on defined as the normalized effective size.

$$I = \frac{E}{d_e} = 1 - \frac{R}{d_e} \tag{2.3}$$

The efficiency value is one, if alters are not connected, and at minimum zero, indicating that alters are well connected [49].

In terms of content propagation, a higher efficiency of an individual would denote that a user might have a higher potential in spreading a message. to validate the hypothesis, data from Digg.com, described in Appendix A.1.3 was taken. In Digg.com, users submitted bookmarks to the web-service on which other users voted on. If a submission (i.e. story) got enough popularity by having a certain number of positive votes, the submission was displayed at the frontpage of Digg.com where it received high attention. By calculating the efficiencies for all ego-centric networks created out of the friendship network of Digg.com, the basic assumption that egos with a high efficiency are more successful seems valid. The average efficiency of all egos in Digg.com is 0.451 with an effective size of 20.27 for 444,003 users that submitted at least one story to Digg.com. The average efficiency of all nodes that succeeded in having a story promoted is 0.795 with an effective size of 102.21 for 29,318 individuals.

2.5. CHAPTER SUMMARY

This chapter described the difficulties to sample inhabitants of a country from data obtained from an OSN. It is by no means possible to obtain a purely random sample in terms of all attributes of users. Whereas certain properties of inhabitants, like the name, location and gender might state a proper sample, created by using messages annotated with GPS information, other attributes like the age of users will probably be biased. If data from OSNs is used to estimate the public opinion or predict the outcome of elections such bias should be considered.

In Chapter 2.3 the similarities of friends in terms of their properties are analyzed and it is found that most users have a few very similar (close) friends, next to acquaintances and colleagues. Due to the fact that a high fraction of users do not protect their profile from being publicly readable it is shown that the reconstruction of profile information based on publicly viewable profiles of friends is possible with an accuracy of up to 86%. This finding renders current mechanisms to protect a profile, by hiding information, nearly useless but it enables the improvement of recommendation systems.

In the last section of the chapter, the topology of ego-centric networks is described and analyzed. When removing the ego from an ego-centric network it is possible to estimate the number of overlapping communities a user is part of which cannot be found when detecting communities in complete networks of OSNs. In terms of content propagation the position of a user within the network of his friends and the connectivity of friends among each other is of importance because on one hand an efficiency value can be calculated expressing the ability to successfully spread content and on the other hand, discovered communities might provide insights into different interests of an individual. However, further research is needed to estimate the influence of community structure to content propagation.

3

Sociocentric Network Analysis

At the beginning of the 20th century, sociologists like Georg Simmel and Leopold von Wiese started looking at societies from a different point of view. Not individuals, as previously assumed, were the main descriptor for societies but the interactions between individuals in all their manifold. In 1908, Georg Simmel described the approach of formal sociology, which is based on the analysis of social relations. His statement, still used nowadays, describes that the form of relationships is the basis for societies as interactions of individuals are defining a society.

Sociocentric network analysis quantifies relationships and analyzes network structures within defined groups, like a classroom of children, the inhabitants of a city, the population of a country or all users of an OSN. In Online Social Networks the definition of a group or population is a difficult task as already stated in Chapter 2.1.1. On the other hand, analyzing a dataset of all users of an OSN is often not feasible because of problems related to the size and dynamics of social services. For example, the OSN Facebook.com reports 1.28 billion monthly active users at the end of 2013 [50], Hyves reported 10 million user accounts in 2010 and Twitter reports 241 million users [51]. The problem arising from large social networks are manifold. On the one hand, the data collection process might take quite long, up to multiple months which indicates that data gathered in the beginning of the process is outdated once the collection is finished. On the other hand storing and processing huge amounts of data states a quite expensive problem as machinery and manpower is needed to deal with the data.

The research question within Chapter 3.1 is therefore related to the amount of data necessary to estimate topological metrics correctly and which method to sample data should be used. Due to the fact that OSNs have specific topologies based on communities or groups of users an approach that traverses these communities might seem favorable. As shown earlier in Chapter 2.4 however, individuals are likely to attend overlapping communities which will be described in Chapter 3.3. Other sociocentric network metrics and their relation to content propagation will be described in Chapter 3.4. The

question of interest within this last section in this chapter relates to the applicability of certain sociocentric graph metrics in terms of content being transported across the network.

3.1. OBTAINING NETWORK DATA

Data collection in a network is steered by a data sampling strategy, which can be roughly categorized into random sampling techniques (for example by randomly picking from a set of previously known node ids [39]), stratified sampling or traversal-based sampling. In practice, most research favors the latter as graph-traversal algorithms generate connected topologies from the very beginning, even if only a small subset of the graph has been obtained.

Among the class of graph-traversal algorithms, the classic breadth-first-search (BFS) and depth-first-search (DFS) are most widely adopted, as they are easy to understand and implement and comprehensively covered across standard textbooks (e.g. Cormen [52]). From these fundamental algorithms, a number of derivations have been introduced for specialized applications and types of utilizations, for example snowball sampling explained in Goodman [53] which for each node only visits n randomly-chosen, unknown neighbors, or forest fire described in Leskovec *et al.* [54] which probabilistically skips neighbors during its breadth-first-search. While these traversal algorithms visit nodes according to the specific order they were discovered in, other algorithms steer their search based on metrics computed at run time. Derivatives of random walk algorithms keep a transition matrix that is continuously updated based on the node degrees encountered in the search. One example of such algorithms is “non-backtracking random walk with re-weighting” described by Lee *et al.* [55] which eliminates the possibility of traversing back into the already known graph to increase search efficiency. Random traversal algorithms however have the disadvantage that coherent community structures of a graph only become visible comparatively late in a crawl, which makes these methods unsuited for example for user prediction and privacy research exploiting the commonalities of users within small-size clusters as shown in chapter 2.3 of this thesis.Other algorithms, such as mutual friend crawling described in chapter 3.1.2 and in Blenn *et al.* [56], therefore aim towards the modular structure of networks and utilize link structure statistics to steer its search to sequentially visit and remain as long as possible inside clusters in order to retrieve closed communities of users.

While graph-traversal algorithms such as BFS and DFS have been used for decades, an analysis about the representativeness of their output has only very recently begun with the emergence and analysis of large scale networks shown in Kurant *et al.* [57]. The fact that graph sampling can lead to a skewed result was already observed in sociological studies of small scale interaction graphs, as the discovered “your friends have more friends than you" corollary described by Feld [58]. An in-depth study of this high-degree bias of BFS was conducted by Kurant *et al.* [57], who were able to theoretically show the bias made when estimating the degree distribution. Their theory indicates that the average node degree only asymptotically approaches the actual degree after more than 40% of the network is sampled, and simulations of artificial 10,000-node networks match these predictions. This observation was further tested on samples of the Facebook graph in Gjoka *et al.* [59], which crawls of 81,000 users each obtained by scrapes of the so-

cial network. This however brings with it the complication that only open profiles and friendship relations are contained in the data set. As in some networks such as Hyves nearly half the profiles are marked as non-public depicted in Blenn *et al.* [33], a focus on only visible relations could potentially introduce some other behavioral bias.

Algorithm 1 BFS (DFS, RFS) Graph Traversal

```
1: Function{Traversal}{G, s}
2: visited ← ∅
3: Q ← List(s)
4: while |Q| > 0 do
5:    u ← Q.removeFirst {DFS: Q.removeLast}
6:    {RFS: Q.removeRandom}
7:    visited ← visited ∪ u
8:    for v ∈ neighbors(u) do
9:       if v ∉ Q and v ∉ visited then
10:         Q.addLast(v)
11:      end if
12:   end for
13: end while
14: EndFunction
```

A common baseline of previous bias evaluations is the performance of a random walk on the graph. Due to the scale-free structure of OSN graphs, leads to a large estimation error of node degrees shown in Kurant *et al.* [57] based on its strong linear preference towards high degree nodes [60]. Therefore such random baseline has to be included as a "ground truth", modified however with a small but important twist: the random-first-search (RFS) used in this section, which randomly chooses a node from the list of discovered, but still unprocessed neighbors. This performs a random walk on only the discovered but not visited part of the graph without re-visiting nodes, which for a graph crawl is undesired overhead.

The main conceptual difference between BFS and DFS (see algorithm 1) can be summarized by how each algorithm chooses the next vertex to visit among those it has already discovered: As BFS processes each node in the order it was discovered, BFS will extend its search in a circular fashion from its starting point, first exploring all nodes of distance 1, before continuing to nodes at distance 2, and so forth. Due to the typically scale-free degree distributions of social networks, BFS will likely start at a low degree node in the periphery of the graph, from where it will quickly proceed to the well connected nodes in the center. Exploring nodes through a FIFO queue, BFS will remain in this area during the first part of the crawling process and thereby over-sample high-degree nodes. DFS on the other hand will continue its exploration process starting from the last discovered node and add newly discovered nodes to the front of its list. As a result, DFS will also reach the well-connected center fast but continue its search until it reaches leaf nodes in the periphery, thereby sampling the low-connected leaf nodes at a higher rate than BFS and resulting in an underestimation of nodes' degrees.

Although all previous initial work on crawling biases has focused only on node de-

gree, the behavior of these crawling procedures and their preference for high-degree nodes in the center or low-degree nodes in the periphery also heavily influence other commonly used topological graph metrics.

3.1.1. METRIC CONVERGENCE

The convergence of a number of topological graph metrics commonly used in social network analysis as a function of the amount of graph crawled can be investigated in order to quantify the introduced bias, based on different crawling techniques. The chosen key metrics are the following: (1) assortativity which measures to which extent pairs of nodes of similar degree connect to each other, (2) the average node degree, (3) the correlation between a node's degree and the average degree of its neighborhood, (4) the network diameter describing the longest shortest path from any node in the network to any other node, (5) graph density and (6) a fitting of the power-law exponent on the node degrees.

The analysis of metric convergence is conducted on the network topology of friend and follower relations within the social media aggregator Digg.com, for which a complete graph topology was collected and presented in [61]. For further explanations about the Digg.com dataset please consider the Appendix A.1.3.

To analyze the convergence of above named network metrics, the Digg.com graph was crawled using the previously discussed breadth-first-search (BFS), depth-first-search (DFS) and random-first-search (RFS) traversal algorithms from 100 randomly selected starting points, tracking the development of each graph metric value. As the stability of metrics and their variance between individual runs changes over the course of the crawl, the interval at which the network metrics are assessed dynamically varied. At the beginning of the crawl, which is showing the highest fluctuations, all graph metrics are computed every time 5,000 nodes have been visited. When the sample contains between 5% and 15% of the total network size, the metrics are computed every 15,000 nodes, and for the remainder of the traversal metrics are evaluated every 50,000 nodes. This results in an approximately equal sampling of the three regions.

From a practitioner's perspective of network crawling, it is important to note the difference between discovered and explored nodes. The former have simply been seen by a graph traversal algorithm while the latter group has been completely processed during an iteration of the algorithm. While this seems an insignificant distinction, it implies a significant difference in practical social network crawling. When obtaining data from a social network site, a crawler typically needs to initiate individual requests for each aspect of a profile, as this keeps the processing and data transmission demands for the site's operator low and matches the vast majority of consuming application programs. For a crawler, this however implies that repeated individual requests are necessary to obtain a complete view about a person's profile and friends: when requesting a user's profile page on Digg.com (but also Hyves, Twitter or Facebook), this initial response only contains the information about the single user in question. When requesting a list of friends, social networks typically only return a list of unique identifiers without providing any further context information about the friends themselves. In other words, while it is known that the user has x friends there is nothing that can be said about those x persons as their profile pages have to be individually retrieved. When crawling a social network, it is therefore possible to quickly compose a large list of users on a given site by retrieving the

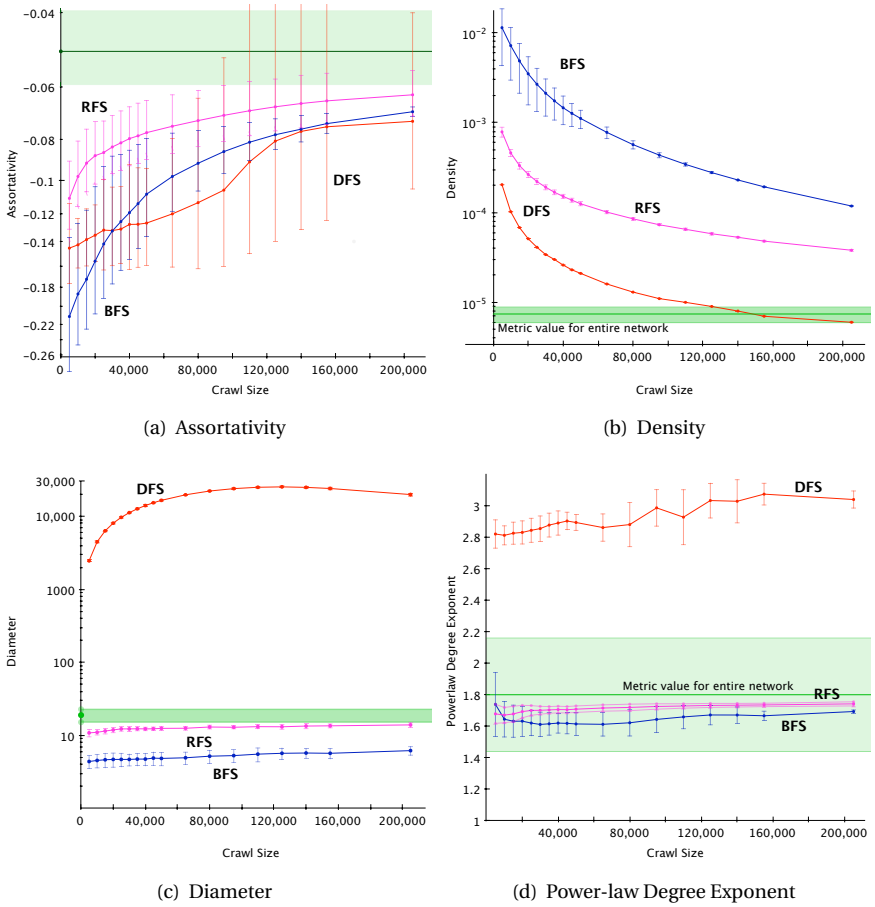
neighborhoods of only a few users (this set is referred to as the set of *discovered nodes*). A correct computation of most metrics will in practice however require that the profiles of the discovered users will also have been retrieved (referred to as *known nodes*). The size of the crawled graph is therefore always measured in terms of *known nodes*.

In this section, plots are shown depicting the development of metric values, crawled through a BFS (blue line), DFS (red line) and RFS (purple line). The line shows the arithmetic average of 100 randomly chosen starting points for a particular metric value and crawling method. The error bars show the standard deviation across these 100 runs. All plots will show the crawl size in a linear scale on the x-axis, while the metric on the y-axis will be shown either in linear or logarithmic scale to maximize readability. The solid green line displays the actual metric value for the entire Digg.com social topology if entirely crawled, the green colored area around this line displays a +/- 20% deviation from this value.

Figure 3.1(a)-(f) show the values and changes of assortativity, average node degree, degree correlation, density, diameter and power-law degree exponent respectively, during the first 200,000 crawled network nodes. These initial 20% were split out separately, as most network crawls in practice are below this size and a large amount of variability takes places within these first phase.

As can be seen in Figure 3.1, DFS, RFS and BFS all converge to the final metric value, however the speed of convergence is surprisingly small and largely depends upon which metric is being used. The majority of graph traversal algorithms has not entered the +/- 20% corridor after the first 20% of the total network size have been crawled, there is more than a factor of 2 difference in 12 of the 18 metric/algorithm combinations after 100,000 known nodes (~10% of the graph's size). This only improves slightly after twice the iterations, after 200,000 crawled nodes still 11 out of 18 combinations are off by a factor of 2. Despite the generally modest performance results across the three algorithms, the approximations made by random-first-search are in nearly all cases (with the exception of graph density) significantly more accurate and approaching the correct value more rapidly than those of the other two algorithms.

Between the, in practice most commonly if not exclusively used BFS and DFS approaches, the results do not indicate a clear winner. For an estimation of graph density, DFS approaches the true value faster, while for an approximation of the power-law degree exponent BFS proves to be a superior choice. This metric-specific under- and overestimation of BFS and DFS can be intuitively explained for most of the presented metrics based on their algorithmic design. As BFS polls nodes in the order they were discovered, BFS quickly reaches the network core from its initial starting point and – having now encountered and added a long list of neighbors from these well-connected nodes – remains there for extensive periods of time. The result can for example be seen in the average node degree of the *known* nodes, which during the initial 10% of the crawl remains nearly an order of magnitude larger than its true value. This behavior has a similar impact on the graph density, which is overestimated at nearly two orders of magnitude, an estimate made on the initial data obtained from the well-connected core of the network. That BFS tends to directly reach towards the center can also be inferred from its estimate of the diameter, the longest shortest path that has been found so far. BFS's diameter lifts off but then remains constant at about half its final value, again another indication that



(a) Assortativity (b) Density

(c) Diameter (d) Power-law Degree Exponent

Figure 3.1: Metric convergence during initial crawling of 200,000 network nodes.

the traversal has reached from the outskirts half way across in the network into its center. The exact opposite can be associated with DFS, which builds a stack of discovered nodes exploring always the last discovered node first. This behavior drives this algorithm across the network core into the periphery of the graph, which can be clearly seen at the average node degree. At a value at or slightly above 2, DFS must have explored a non-trivial amount of leaf nodes in its search, if one considers that leafs have a degree of 1, the node with the next lowest possible degree (besides bridges with degree = 2) is a node connecting two leaves to the remainder of the graph already implies a degree of 3. In consequence, DFS will drastically underestimate node degree but conducting most of its search across the leaf nodes in the periphery of the graph, which can be seen in the extremely long chains of nodes being created while simultaneously also driving the estimate of the graph's density to a minimum.

This behavior can additionally be visualized through the average number of neighbors (ANoN), which expresses the average degree of those neighbors directly connected to the currently processed node. This momentary average across the *discovered* but not yet *known* nodes therefore gives a rough approximation where in the graph a traversal algorithm currently resides. High numbers of ANoN indicate well-connected neighborhoods typically found within the core of a scale-free topology, while low numbers indicate a graph's periphery with a large portion of leaf nodes (where in the extreme case ANoN will be close to 1). Figure 3.2 shows the average number of neighbors as a function of the DFS, BFS and RFS crawl size, binned into the current reporting interval (5,000, 15,000 or 50,000 nodes). With an initial 40 fold increase over the true long-term average number of neighbors during the initial phases of the crawl, BFS clearly first explores nodes in the well connected center of the graph while only later turning into the periphery, while DFS with a ANoN below half of its true value and being between 2 and 6 traverses the vicinity of the network.

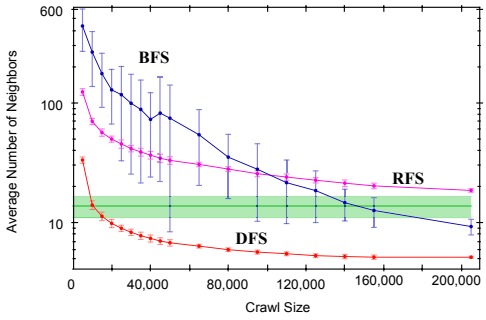


Figure 3.2: The "Average Number of Neighbors" - a metric to asset crawl locality.

While the general patterns of metric convergence can be very well explained from these behavioral characteristics of graph traversal algorithms, the extremely long convergence times and estimation performance are nevertheless astonishing, so much that the study was replicated using several approaches to rule out possible methodological errors. Figure 3.4 shows the convergence of the six target metrics for the entire crawl, and while all algorithms rebound across all metrics towards the actual final metric value, a proper convergence into the +/- 20% error margin can for some combinations take until the very end of the exploration. For many situations, the convergence is however not that grim: both the average node degree and the power-law degree exponent can be properly estimated if using the right crawling strategy after 5-10% of the entire network, a solid estimate of assortativity becomes possible after approximately 30% of the entire graph. Besides the different crawling patterns, this instability is for some metrics also inherent to the metric itself, assortativity for example has been shown to be dependent on the size of a graph [62], which can be also seen in Figure 3.4.

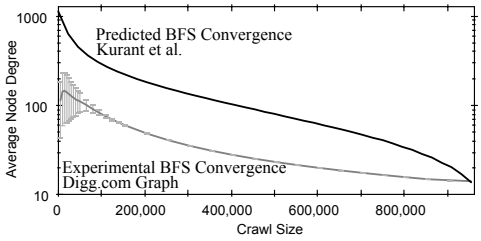
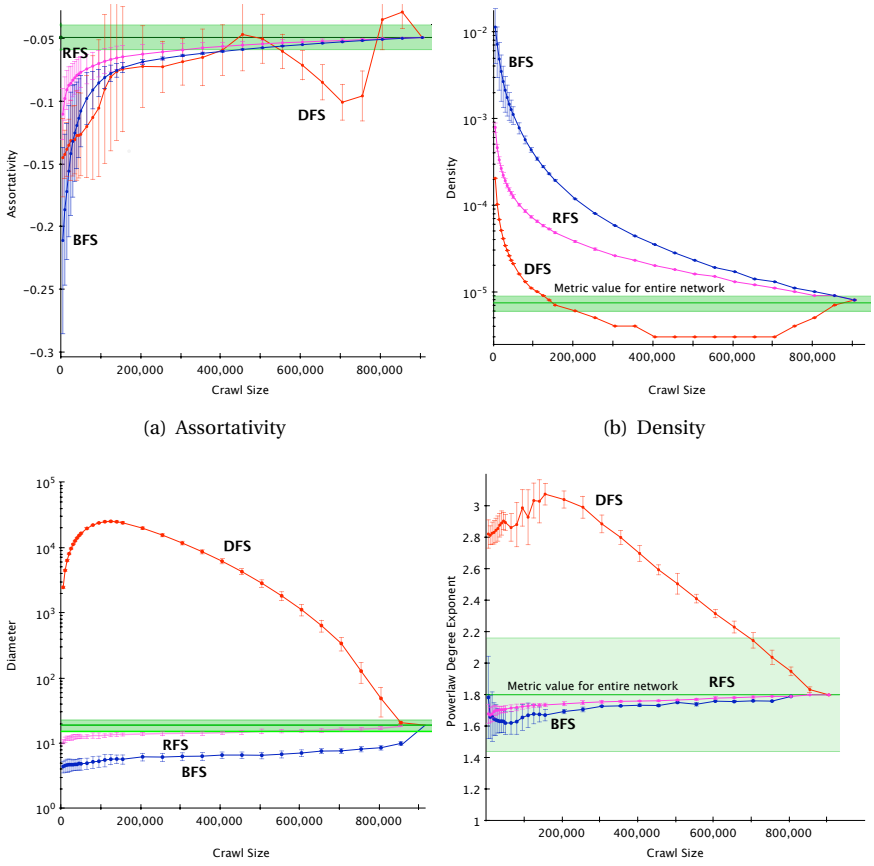


Figure 3.3: Comparison of BFS convergence.

As an additional validation step, the BFS convergence behavior and metric estimates for the average node degree as theoretically derived in [57] can be validated with those



(a) Assortativity (b) Density

(c) Diameter (d) Power-law Degree Exponent

Figure 3.4: Metric convergence during entire network crawl.

observed when crawling the Digg.com social graph. Figure 3.3 shows this comparison. While not providing an exact match, both approximate the true value asymptotically in a similar fashion.

3.1.2. MUTUAL FRIEND CRAWLING

As mentioned in the beginning of this chapter, sociocentric network analysis focuses on groups of individuals and the relations between them. The general problem is stated by the question of finding a group of users in a social network. If attributes of the users are available, it is a rather simple task as one may analyze the sub-graph containing only users that match a desired parameter. But, as not every individual provides the desired attribute and techniques explained in chapter 2.3 may be too time consuming, a topological approach can be considered.

Milgram's [1] famous experiment, often called the "six degrees of separation" phenomenon, showed that every person can be reached through 6 others. Although the experiment set-up and the conclusions drawn by Milgram are questioned [63], the general observation that most users within a population (of a country or an OSN) are connected and form a so called "large component" seems to be valid. Large components typically contain 70%-90% of all user accounts in OSNs (measured in the data sets described in the appendix A.1 of Digg.com, Twitter.com and Hyves.com). The remaining 10%-30% can be found in numerous smaller components, disconnected from each other. By analyzing the large component of OSNs from a topological point of view, identifying groups is therefore related to the problem of finding communities.

The automated process of downloading users typically requires a robot (a software program) to look at the profile page of a user and store the names of all friends. Such an operation may take 0.1 to 2 seconds as it includes multiple HTTP requests to a server in order to iterate through the whole list of friends. An optimistic[1] calculation shows that with one crawling computer, obtaining LinkedIn's database of 120 million users (as of November 2011) would take approximately half a year. The same calculation for Facebook's dataset of 1.2 billion users leads to a crawling time of circa 5 years. By using massively parallel crawling techniques those times can be decreased. Clearly, by the time the last records have been obtained, most of the retrieved information will be outdated.

A lot of work on social network analysis is conducted on communities of users as a level of abstraction. A natural question is therefore whether it is possible to direct a crawling procedure in such a way that it is obtaining a network community-wise. This would enable researchers to analyze sub-graphs of the whole network, even if users hide their information, while still obtaining data. In contrast, using the standard crawling methods like BFS, DFS or RFS, as mentioned in the previous section 3.1.1, one literally needs to wait until the whole network is crawled before starting to analyze the data because there might be a few users critical to a particular single community still missing from the dataset, and their existence and criticality cannot be determined until all data is collected.

In this section, a simple approach to crawl a network community-wise and detect communities at the same time is described. The algorithm called *Mutual Friend Crawling* achieves similar results, compared to existing community detection methods.

Communities are defined in terms of the fraction of nodes of a network that share more connections with each other than with the rest of the network.

A well known metric to capture the community structure of a network is modularity. Modularity m as defined in Clauset *et al.* [64] is "the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random." The definition of modularity is given in equation 3.1.

$$m = \frac{1}{2L}\sum_{i}^{N}\sum_{j}^{N}(a_{ij} - \frac{d_i d_j}{2L})1_{\{i \text{ and } j \text{ belong to the same community}\}} \tag{3.1}$$

For a given graph G with N nodes, L links and a given partition, the modularity denotes how well the community structure is expressed. The element a_{ij} denotes the element

[1] In this context, optimistic means that no mechanisms against crawling or screen scraping are enforced.

corresponding to the ith row and jth column of the adjacency matrix of G and d_i is the degree of node i. $1_{\{i \text{ and } j \text{ belong to the same community}\}}$ is the indicator function returning 1 if i and j are in the same community otherwise 0.

A modularity value of 0 defines that the number of links belonging to the same community is equal to the number a random graph would have. The higher the modularity, the more pronounced the community structure, except for the trivial case of modularity = 1, in which all links of G are in the same community. Conversely, this means that negative values are a definition of something like an "anti-community" structure. An overview over modular graphs and how to achieve high modularity is given in Trajanovski *et al.* [65].

Mutual Friend Crawling (MFC) is crawling community structure in real world OSNs having linear complexity and results are to be compared to techniques like BFS and DFS as described in Cormen *et al.* [52]. In both techniques (BFS and DFS), the graph is crawled node per node adding all discovered nodes to a list of nodes to visit. The difference between BFS and DFS is based on the procedure how the next node to visit is selected. In BFS the first node of this list is selected to be visited next and removed from the list whereas in DFS the last node in the list is selected and marked as visited. Both techniques are directing the crawling procedure towards the inner core of the network due to the friendship paradaxon, first observed by Feld [58]. It states originally that friends of an arbitrarily chosen user have more friends than the user itself, which will force a crawl towards nodes having a high centrality in the network. A related effect, described by Kurant *et al.* [57], describes that BFS and DFS is introducing an bias towards high degree nodes for an incomplete traversal of the network already described in section 3.1.1.

BFS and DFS are the most used techniques to traverse a graph but in order to reveal the community structure of a graph BFS and DFS are not the best choice. One possible technique of crawling a network and detecting communities at the same time is to facilitate random walks. Random walks are known to stay inside communities as described in Pons and Latapy [66] and Lai and Lu [67]. The main idea behind random walk community detection is that a community has more links between nodes of the community than between communities. Based on this definition, a random walk would traverse nodes of the same community more often than the ones of different communities. However, a random walk allows steps backwards to already visited nodes which is increasing the time taken to crawl the network.

Different community detection algorithms like fast and greedy community detection by Clauset *et al.* [64], Spinglass by Reichold and Bornholdt [68], edge betweenness clustering by Girvan and Newman [69] or label propagation by Raghavan *et al.* [71] cannot be used to detect communities during crawling as those algorithms are meant to be applied onto the full topology of a network.

A related approach to detect community structure while crawling is presented by Nguyen *et al.* [72]. Their algorithm Quick Community Adaptation (QCA) assumes that the community structure is already known for a complete network and manages to calculate community structure in dynamic networks. QCA tries to maximize modularity by assigning a "force" which attracts a node towards a community. However, this method also needs the whole network including assignments of nodes into communities. In their approach the used algorithm to estimate the initial community memberships is

presented by Blondel *et al.* [73] called the Louvain(-la-Neuve) method. This algorithm calculates a modularity maximizing partition of a given graph by using the change in modularity when discovering a new node and adding it to an existing community. If the difference is not positive the node stays in its initially assigned community.

To compare the result of different clustering algorithms the Jaccard similarity coefficient next to the already defined modularity can be facilitated. The Jaccard similarity coefficient defines the similarity of sample sets by measuring the quotient of the intersection and the union of both sets defined by Fortunato and Castellano [74] given in equation 3.2.

$$I_J(s_1, s_2) = \frac{n_{11}}{n_{01} + n_{11} + n_{10}} \tag{3.2}$$

In equation 3.2, n_{11} denotes the number of node pairs found in the same community whereas n_{01} and n_{10} are the number of pairs of nodes assigned to the same community by algorithm s_1 but not s_2 and vice versa.

In contrast to BFS and DFS, MFC assumes the knowledge about the degree of neighboring nodes. This assumption is reasonable in OSNs as the number of friends is easily to obtain, whereas the process of receiving the actual links towards them needs more effort in practical terms. For example in the OSN Twitter, each message contains the number of followers and friends the originating author had when writing the Tweet. Also, OSNs are most commonly crawled using screen scraping. Here, the OSN is accessed in the same way a user does by using HTTP requests to analyze web pages for relevant data. The number of friends is usually listed at the profile page of a user. Several clicks (requests) on the list of friends are needed to obtain all node ids (friends) having a relationship with this user. If one is interested in more details of a user, for example the real name or group affiliations, the profile information will need to be obtained in any case and at the same time a friend count is mostly available without any additional overhead. In this way the needed crawling effort does not increase but the order in which data is gathered is changed.

MFC is based on the "reference score" S_R defined in equation 3.3. This score denotes the fraction of the number of already discovered links (references) r_f pointing to node f so far in the crawling process and the total degree d_f, i.e. the total number of friends, of node f.

$$S_R = \frac{r_f}{d_f} \tag{3.3}$$

During the crawl, the next node to process is chosen from the list of the already discovered nodes having the largest S_R. The full algorithm is specified in pseudo code in algorithm 2.

MFC is based on a BFS algorithm having two major differences. The first one is a map used to store the number of found references as indicated in line 2, 4 and 15. The second difference is based on the way the next node to visit is chosen: instead of choosing the next one from the list of discovered nodes as BFS does, MFC calculates the reference score (lines 6-9) and chooses the next node based on the maximum of the reference score (line 10 & 11).

Algorithm 2 MUTUAL FRIEND CRAWLING

```
create a queue Q
create a map R
add starting node to Q
store starting node and 0 as number of found references in R
while Q is not empty do
    for all elements in R do
        reference_score ← value in R / degree of the node
        max_score ← max(max_score, reference_score)
    end for
    next_node ← dequeue element having max_score from Q
    delete next_node from R
    if next_node has not been visited yet then
        for all neighbors of next_node: do
            add neighbor to Q
            increment number of found references to neighbor by 1 and store it in R
        end for
        remember that node (next_node) was visited
    end if
end while
```

The algorithm will therefore visit nodes first, having a large S_R. If a network has a community structure based on the definition of having more links in the community than links connecting communities, MFC will crawl communities one after another.

In order to apply MFC on weighted graphs, a simple definition of the strength of a node as the sum of the weights of adjacent edges is sufficient. In this case the reference_score S_R is defined as the fraction of the sum of weights of already discovered links to the strength of the node as defined in 3.4.

$$S_R = \frac{\sum (\text{weights of found references to f})}{\text{strength of node f}} \tag{3.4}$$

3.1.3. COMMUNITY CRAWLING

Figure 3.5 illustrates MFC intuitively using a small example. Simple visual inspection shows that there are six clusters. If one wants to explore each cluster one after the other, the algorithm should first explore all nodes having one color in Figure 3.5 before continuing with the next group of nodes. The nodes labels denote one possible order in which the graph could be traversed in order to visit communities one after another, thus leading to the intended and perfect order of exploration.

In order to test the proposed algorithm on multiple graphs, one may measure how strong the community structure in a given graph is expressed, by using the ratio of links inside communities to the total number of links. This value, defined as P_{in} is reflecting the probability an arbitrary chosen link is an intra-community link.

MFC was tested on a total of 100,000 artificial networks with different P_{in} values using a graph generator, described in van Kester [75], each with 10,000 nodes, 100,000

links and 100 equally sized communities. Two general types of graphs commonly found in network science were evaluated: The degree distribution of the first type follows a uniform distribution and the degree distribution of the second type is approximating a power-law function. All graphs have been crawled from all possible starting nodes. During the crawl, keeping track of the order of visited nodes allows to analyze if a complete community has been crawled before going to the next one.

Figure 3.6 shows the crawling trajectories of those crawls. The figure is expressing how many nodes have to be crawled in order to visit all nodes of a community. In order to crawl the network community-wise, the optimal traversal would need to visit all nodes of one community first before visiting the next node belonging to the next community. As all communities are equally sized, the optimal traversal of the graph would lead to a diagonal line in Figure 3.6. A bended line is expressing that nodes from different communities were visited before all nodes of the previous visited community have been finished. The order in which multiple communities are crawled is not reflected in Figure 3.6. The colors express trajectories of different crawling methods where green lines denote DFS, blue ones BFS and red MFC.

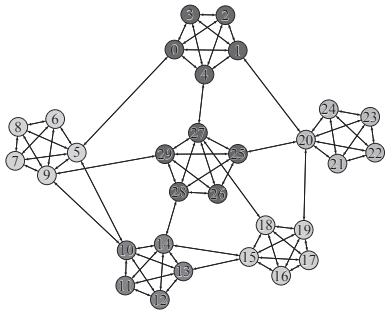


Figure 3.5: A simple example graph. Nodes are labeled by the order of traversal during the crawling process. Different colors denote different communities.

For high P_{in} values, Figure 3.6 illustrates that MFC performs as expected and leads the walk on the graph to all nodes contained in one community before crawling the next. For BFS and DFS a larger fraction of the network has to be crawled to finish one community. Interestingly, BFS perform "closer" to the optimum than DFS. This is because BFS explores the local neighborhood whereas DFS explores the nodes furthest away from the starting node. Thus, the "chance" of BFS to visit all nodes of one community earlier than in DFS is higher.

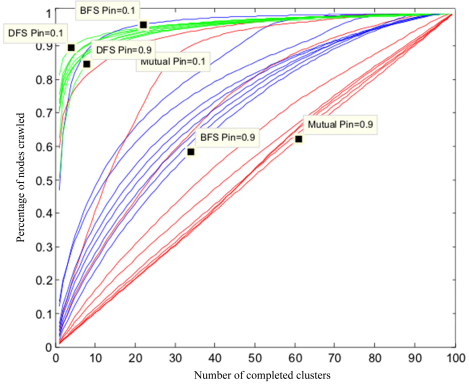
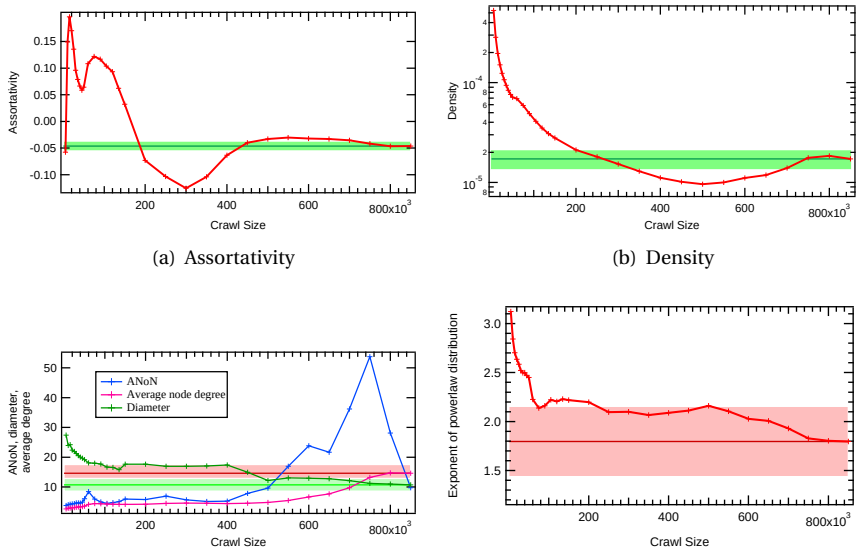


Figure 3.6: Crawling communities. Depicts the percentage of nodes that have to be visited in order to crawl a full community. Green lines represent depth first search, blue lines breadth first search and red lines mutual friend crawling, for different P_{in} values.

MFC performs reasonably better than BFS and DFS in terms of crawling along the community structure. For P_{in} values larger 0.3, the order in which the nodes are traversed fulfills the require-



(a) Assortativity (b) Density

(c) Diameter (d) Power-law Degree Exponent

Figure 3.7: Metric convergence during the entire network crawl for 200 randomly selected seed-nodes.

ments. However, P_{in} values smaller than 0.5 somehow define negative communities in terms of the definition and therefore a BFS approach by chance performs better.

When comparing the metric convergence of MFC to the results presented earlier for BFS, DFS and RFS, MFC gives interesting results, shown in Figure 3.7. For example the average number of neighbors (ANoN-value) stays rather small until nearly half of the network is crawled, which indicates that small communities are traversed first and high degree hubs are only visited once nearly all nodes in the network are visited. This behavior also explains that the average node degree stays at smaller values, compared to BFS or DFS. In addition, the curve of assortativity can be explained by the fact that in communities, users are typically connected to others having a similar degree.

3.2. COMMUNITY DETECTION

As demonstrated empirically, MFC crawls communities of a graph one after another. In order to detect communities while crawling the graph, traces of the reference score of visited nodes can be analyzed. Figure 3.8 shows the trace of reference scores performed on the example graph (Figure 3.5) starting from node 0. As MFC will always select the next node to visit having the highest reference score (line 8 in algorithm 2). Hence the reference scores inside communities should always increase or stay roughly the same while traversing the graph. When detecting a node that is interconnecting communities, a large number of links of this node are referring towards a previously unknown community. Therefore, its reference score will be smaller than the ones of nodes within in

the currently traversed community. As such a node will be selected last, a drop in the trajectory of reference scores of visited nodes can be observed.

Figure 3.8 shows five major drops of the score. Those drops in the reference score of the chosen nodes reflect that the crawler entered different communities. The difference between the next reference score to the previous one is defined as $\Delta refscore$.

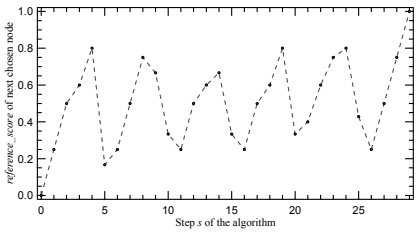


Figure 3.8: Plot of reference_scores versus the number of visited nodes.

While traversing the graph, all nodes are added to the same community as long as $\Delta refscore$ is positive or higher than a certain threshold. If the score decreases a new community will be visited. To prevent the creation of single node communities, the drop in the reference_score should be at least half the difference between the maximum S_{max} and the minimum S_{min} of the reference_score in the previous community.

Via the method of tracking S_R, *Mutual Friend Crawling* found six communities on the example graph having the community assignments as indicated different shading in Figure 3.5. One problem however still remains: a possibly incorrect classification of certain nodes during the first visit of a new community. In case neighbors of the starting node as well as neighbors of the first node of a community have the same degree as the visited node, a node of a different community could be assigned incorrectly.

Such misclassification is exemplified for the case of three equally-sized, fully-connected communities which are pairwise connected through one node as shown in Figure 3.9. In this case, when starting in one clique, all nodes of this clique are added to the first community. When reaching one of the 3 nodes connecting communities (10, 11 or 21 in Figure 3.9) the score drops, a new community is generated and all following nodes are added to this community. When reaching for example node 10, the reference score for the 2 peers (11 and 21) is the same. One of them is chosen to be next node to visit. Now, (e.g., by visiting node 11) 2 links towards the third node are discovered, doubling its score. Having a higher score than all other neighbors of 11, 21 will be added to the same community as 11. Afterwards one of 11's or 21's neighbors are visited where the reference score drops

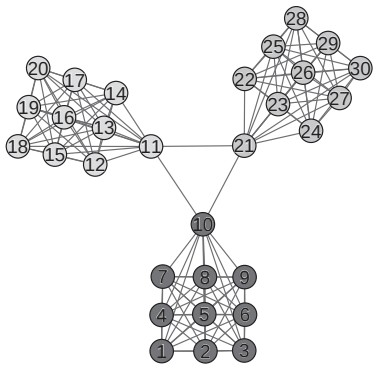
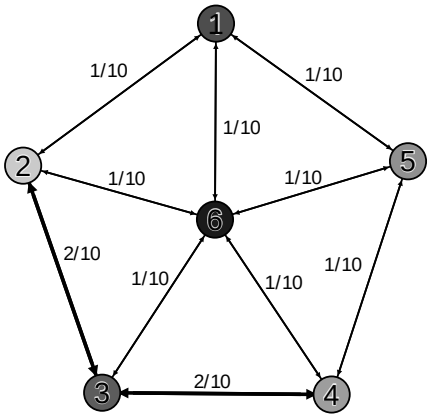
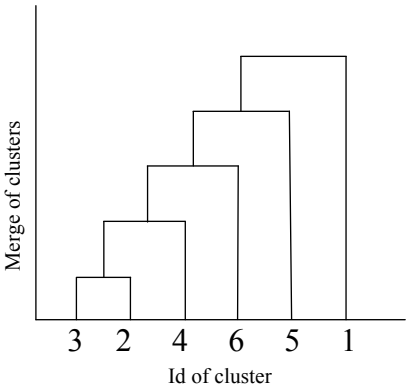


Figure 3.9: Example graph where a misclassification may occur when visiting node 11.



(a) Graph of communities estimated out of the given example graph.



(b) Dendrogram representing the hierarchy of communities in the example graph.

again leaving 11 and 21 in one community. All other nodes are then correctly classified. The solution is to check for this kind of misclassification by iterating through all nodes in a already discovered community checking if a node has more connections with another community then inside the "own" community. If so this node is merged to the connected community. As this procedure is raising the density in the community the node is merged to, the modularity value will only increase as stated in Trajanovski *et al.* [65].

To express a hierarchy of communities the dendrogram can easily be built using the graph of found communities. In this graph, every community is represented by a node. Links between nodes are weighted by the number of links connecting the 2 communities one level lower in the hierarchy or, if the original graph was weighted, the sum of all weights of links between two communities. The graph shown in Figure 3.10(a) depicts the first hierarchical level of the example network. Through iterative pairwise merging of communities connected by a link, multiple levels of hierarchy are created. The merging step is based on the fraction of the number of links in the original graph connecting two communities to the number of intra-community links of the community that is to be merged with a second one, because just using the total number of links between two communities states an unfair measure as it does not take the differences in size and density of communities into account. When assigning this value to the directed links in the graph of communities, the directions and magnitude of link-weights define which communities should e merged and in which order. This hierarchy is depicted in Figure 3.10(b).

To prove the correctness of *Mutual Friend Crawling* in terms of community detection, a comparison of the results of community assignments to existing approaches is needed. However, as the variety of community detection algorithms is too large to compare against, just those algorithms were chosen which (with slight modifications) can be used to identify communities while crawling.

Dataset	Method	Number of communities	P_{in}	Modularity
Karate club	Original partition	2	0.86	0.36
	Louvain method	4	0.74	0.42
	Fast and greedy method	3	0.74	0.38
	Random walk method	5	0.63	0.35
	Mutual friend crawling	2	0.86	0.36
Football	Original partition	12	0.64	0.554
	Louvain method	10	0.708	0.604
	Fast and greedy method	5	0.746	0.544
	Random walk method	9	0.726	0.603
	Mutual friend crawling	9	0.736	0.57
Digg	Louvain method	26646	0.94	0.478
	Fast and greedy method	37591	0.92	0.393
	Mutual friend crawling	78308	0.83	0.142

Table 3.1: Comparison of Mutual Friend Crawling to well known community detection procedures on different datasets.

The chosen methods are:

1. Newman and Clauset's fast and greedy modularity maximizing method [64]
2. Pons & Latapy's random walk method [66]
3. the Louvain(-la-Neuve) method by Blondel *et al.* [73]

As all mentioned approaches are not directly providing a map of community ids to node IDs, the partition resulting from the merges of nodes leading to a maximum of modularity was chosen. This partition is then compared to the output of MFC. For the given example graph in Figure 3.5, all community detection algorithms found the same result as indicated by the colors in Figure 3.5.

A comparison of the mentioned methods on some selected data sets is given in Table 3.1. The chosen data sets were: "Zachary's karate club" [76], Girvan and Newman's "American College football games" [69] and the network of all Digg users as described in Tang et. al. [77].

As given in Table 3.1, the partitions found by MFC are comparable to existing and well known procedures when compared in terms of the P_{in} value and modularity, except for the last dataset, a large-scale directed network of all users of Digg.com, where the number of detected communities is higher than the result of the Louvain(-la-Neuve) or fast and greedy method.

However, this could be based on the resolution limit of modularity, as described in Fortunato and Barthélemy [78]. A partition having a high modularity could lead to a relatively small number of large communities which is not reflecting the real community structure. The communities found by *Mutual Friend Crawling* are smaller than the ones found by the other methods still having the same properties like a power law shaped community size distribution. Also the number of users in a group given by MFC is reasonable. The largest community found by Mutual Friend Crawling has a size of 9,443 users whereas the largest one found by the Louvain method contains 186,271 users. Without further investigation one may argue for both numbers to be better than the other one. Therefore this question is left out to be solved by further research.

method	original	Louvain	Fast and greedy	Random walk	Mutual Friend Crawling
original	1	0.719	0.354	0.615	0.468
Louvain		1	0.424	0.721	0.483
Fast and greedy			1	0.422	0.324
Random walk				1	0.487
Mutual Friend Crawling					1

Table 3.2: Comparison of the partitions discovered by MFC to other community detection algorithms on the college football dataset [69] using the Jaccard similarity index.

As the P_{in} value and the modularity are global values which cannot be used to compare two partitions directly the Jaccard similarity index may be used. As mentioned earlier, if pairs of nodes are assigned to the same community this similarity will have a high value.

Table 3.2 shows the similarity of node assignment into communities between the different community detection algorithms. While this metric is not very sensitive to the number of communities, it shows that MFC is equally good as well known methods. A more complete analysis of the Jaccard similarity index is given in van Kester [75].

However, using MFC will only partition the topology of a network into communities whereas users in OSNs are typically members of multiple “overlaying” or “overlapping” communities which cannot be captured by this kind of community detection.

3.3. OVERLAPPING COMMUNITIES

Communities overlap with each other when nodes belong to multiple communities. Such overlap exists widely in real-world complex networks, particularly in social and biological networks [81–83]. In social networks, human beings have multiple roles denoting that people are members of multiple communities at the same time, such as companies, universities, hobby clubs etc. as exemplary shown in Figure 2.17.

Other examples are movie actor networks, where nodes are actors and actors are connected if they acted together in one or multiple movies, one could regard the set of actors in one movie as a community. According to such a view onto a movie actor network, the communities of all movies are by definition cliques which overlap with each other if certain actors participated in the making of multiple movies. Similar networks are coauthorship networks (nodes represent scientists, nodes are connected if they coauthored one or more articles where a community denotes all authors of a publication), journal editor networks or sports player networks (players who played in the same games are connected).

These types of networks are a special kind of social networks, called affiliation networks which naturally contain fully connected sub-networks, called cliques or complete sub-graphs. The clique structure of social networks increases largely the percentage of triangles among the three hop walks, resulting in high clustering coefficient. Besides sociocentric statistics in affiliation networks such as clustering coefficient, characteristic path length and nodal degree, the following metrics are related to overlapping structures: the number of communities, the number of individuals in each community, the communities each individual belongs to, the number of individuals every pair of com-

munities has in common and the number of communities each group is adjacent to (two communities are adjacent if they have individuals in common).

Palla *et al.* [81] defined four metrics to describe how communities in networks overlap with each other: the membership number of an individual, the overlapping depth of two communities, the community degree and the community size. Palla *et al.* [81] showed that communities in real-world networks overlap with each other significantly. They reported that the membership number of an individual, the overlapping depth of two communities and the community size follows power-law distributions, except that the community degree features a peculiar distribution that consists of two distinct parts: an exponential distribution in the beginning and a power law tail. Poller *et al.* [84] proposed a model in which both the community size and the community degree follows a power-law distribution, by applying preferential attachment to community growth. A model proposed by Toivonen *et al.* [85] succeeds in reproducing common characteristics of social networks: community structure, high clustering coefficient and positive assortativity.

In order to describe overlapping communities in OSNs a complete set of metrics to fully characterize the structure is needed. As such a structure is by no means trivial, hyper-graphs can be facilitated to describe social networks.

3.3.1. REPRESENTATION OF SOCIAL NETWORKS WITH OVERLAPPING COMMUNITIES

Suppose the network under consideration has N individuals and M groups, where an individual may belong to multiple groups. The membership number m_j of an individual j is defined by the number of groups of which j is a member. The degree d_j of an individual j equals the number of individuals who have the same membership in one or more groups. The interest-sharing number $\alpha_{i,j}$ of individuals i and j is defined by the number of groups to which they both belong, which indicates how many common interests they share.

The group size s_k of group k is the number of individuals that belong to group k. The group degree u_k of group k equals the number of groups sharing individual(s) with group k. The overlapping depth $\beta_{k,l}$ of two groups k and l equals the number of individuals that they share. An affiliation network is linear if $\beta_{k,l} \leq 1$ for all $k, l \in [1, M]$, where M is the number of groups. The affiliation network is called a m-uniform affiliation network if the membership number $m_j = m$ for $j \in [1, N]$.

The graphs in Figure 3.10 exemplify the definitions of d_j, m_j, $\alpha_{i,j}$, s_k, u_k, and $\beta_{k,l}$. The graph in Figure 3.10 (a) has five

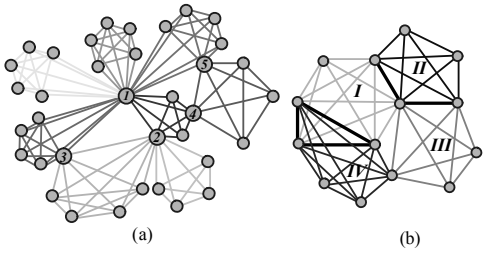


Figure 3.10: Two example graphs illustrating community structure. Nodes denote individuals, communities consist of links of the same shade. a) Labeled nodes belong to multiple communities. b.) Communities are labeled from $I-IV$, black links belong to multiple communities.

labeled nodes which are members of at least two communities. Obviously, $d_1 = 24$, $d_2 = 12$, $d_3 = 10$, $d_4 = 8$ and $d_5 = 9$. Nodes $1-5$ belong to 5, 3, 2, 2 and 2 communities respectively, thus $m_1 = 5$, $m_2 = 3$ and $m_3 = m_4 = m_5 = 2$. The individuals 1 and 2 belong to only one common community, hence $\alpha_{i,j} = 1$. As shown in Figure 3.10 (b), the groups $I - IV$ have 6, 5, 5 and 6 nodes, hence, $s_I = s_{IV} = 6$ and $s_{II} = s_{III} = 5$. Evidently, the overlapping widths: $\beta_{I,II} = 2$, $\beta_{I,III} = 1$, $\beta_{I,IV} = 3$, $\beta_{II,III} = 2$, $\beta_{II,IV} = 0$ and $\beta_{III,IV} = 1$. The group degrees are: $u_I = u_{III} = 3$ and $u_{II} = u_{IV} = 2$.

An affiliation network is usually described by a graph in which nodes represent individuals and two nodes are connected by a link if they both belong to one or several communities. If a set C_I of individuals belong to group I, the set C_I of individuals comprise a fully connected clique. If a set C_{II} ($C_{II} \subseteq C_I$) of individuals belongs completely to group II, one cannot represent group II by this graph description, because the set C_{II} of individuals is already fully covered by community I.

Scott [58] discussed the generation of an affiliation network with simple graphs. Newman *et al.* [86] suggested a bipartite graph model with all information preserved through representing groups by one type of nodes and individuals with the another, where links connect nodes of different type, as shown in Figure 3.11. Lattanzi and Sivakumar [87] proposed a bipartite-graph based generative model for affiliation networks as well.

HYPERGRAPH REPRESENTATION

A hypergraph, $H(M,N)$ with M nodes and N hyperedges, is a generalization of a simple graph whereas a simple graph is an unweighted, undirected graph without self-loops or multiple links between pairs of nodes. The term "hyperedge" is used instead of "hyperlinks" in order not to confuse it with hyperlinks between Internet web-pages. Its nodes are of the same type as those of a simple graph, shown in Figure 3.12(a).

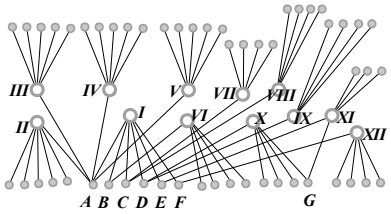


Figure 3.11: The bipartite graph representation of the affiliation network of the NAS group.

Hyperedges of hypergraphs can connect multiple nodes, like the hyperedge A in Figure 3.12 (a) connecting nodes $I, II, \cdots, V$. A hypergraph is linear if each pair of hyperedges intersects in at most one node. Hypergraphs where all hyperedges connect the same number m of nodes are defined as m-uniform hypergraphs with the special case that 2-uniform hypergraphs are simple graphs. If an affiliation network is linear, the representing hypergraph is linear; if an affiliation network is m-uniform, the representing hypergraph is also m-uniform.

An affiliation network with M groups and N individuals can be described by a hypergraph $H(M,N)$: M nodes representing M groups; N hyperedges represent N individuals; and an hyperedge is incident to a node if the corresponding individual is a member of the corresponding group.

The line graph of a hypergraph $H(M,N)$ defined as the graph $l(H)$, in which the node set is the set of the hyperedges of $H(M,N)$ and two nodes are connected by a link with weight t, when hyperedges share t nodes. The degree d_j of an individual j, equals the

Table 3.3: Names and members of all communities of the exemplary social network of NAS.

Index	Name of community	Members (individuals)
I	NAS-TU Delft	A, B, C, D, E, F
II	A research group at MIT	$A, A_1, \cdots, A_5$
III	A research group at Cornell Univ.	$A, A_6, \cdots, A_{10}$
IV	IEEE/ACM ToN editorial board	$A, A_{11}, \cdots, A_{15}$
V	A research group at KSU	$A, A_{16}, \cdots, A_{20}$
VI	A research group at Ericsson	$B, B_1, \cdots, B_4$
VII	A research group at KPN	$C, C_1, \cdots, C_4$
$VIII$	Piano club	$C, C_5, \cdots, C_8$
IX	A research group at TNO	$D, D_1, \cdots, D_4$
X	A rock band	D, D_5, D_6, D_7, G
XI	A soccer team	E, E_1, E_2, E_3, G
XII	Bioinformatics group at TU Delft	$F, F_1, \cdots, F_4$

number of individuals that connect to j in the line graph $l(H)$. The line graph $l(H)$ is an unweighted graph when the corresponding hypergraph is linear; otherwise is weighted, and the weight of link $i \sim j$ equals the interest-sharing number $\alpha_{i,j}$.

AN EXAMPLE OF AN AFFILIATION NETWORK

Table 3.3 describes an affiliation network based on the memberships of individuals within the NAS research group (Network Architectures and Services Group at Delft University of Technology). Individuals A, B, C, D, E, F descibe members of NAS and others are members of groups which overlap with the NAS group. Figure 3.11 depicts the bipartite graph representation of the NAS affiliation network where gray circles represent the groups and the gray disks represent individuals. Nodes are linked when the corresponding individual belongs to a community.

The hypergraph representation of the example network $H(12,53)$ is shown in Figure 3.12 (a). Nodes of the hypergraph denote groups and individuals are denoted by hyperedges. There are 12 groups as described in Table 3.3, corresponding to 12 nodes in Figure 3.12 (a), in total there as 53 individuals among whom, 6 NAS members with membership ids $m_A = 5$, $m_C = m_D = 3$, $m_B = m_E = m_F = 2$. If an individual belongs to multiple groups, the corresponding nodes are connected by the hyperedge specifying that individual.

Figure 3.12 (b) depicts the line graph $l(H)$ of the hypergraph $H(12,53)$ in Figure 3.12 (a). In the line graph $l(H)$, individuals are denoted by nodes and the groups are denoted by links of the same color therefore indicating incident nodes. The line graph $l(H)$ is unweighted since the NAS affiliation network is linear.

3.3.2. TOPOLOGICAL PROPERTIES

The line graph $l(H)$ has N nodes and L links. The topology of $l(H)$ can be described by its adjacency matrix A, a $N \times N$ matrix, where the element a_{ij} equals the link-weight of link $i \sim j$ if there is a link between node i and node j, else $a_{ij} = 0$. Since $l(H)$ is

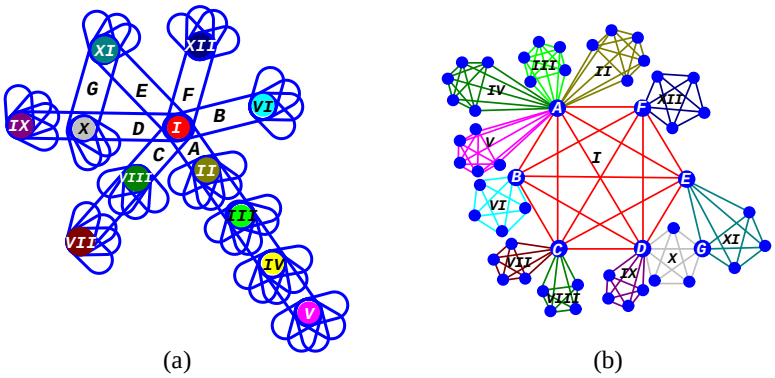


Figure 3.12: (a) The hypergraph representation of the network described in Table 3.3. Hyperedges are blue ellipse-like closed curves, nodes are disks with different colors marked. A node and a hyperedge are incident if the node is surrounded by the hyperedge. (b) The line graph of the hypergraph in (a), nodes denote individuals.

undirected, the adjacency matrix A is symmetric.

The following equalities are valid for all affiliation networks,

$$N = \sum_{k=1}^{M} s_k - \sum_{k=1,l=1}^{M} \beta_{k,l} \tag{3.5}$$

$$L = \frac{1}{2}\sum_{j=1}^{N} d_j = \sum_{k=1}^{M} \frac{s_k\,(s_k-1)}{2} - \sum_{k=1,l=1}^{M} \frac{\beta_{k,l}\,(\beta_{k,l}-1)}{2} \tag{3.6}$$

$$\sum_{j=1}^{N} (m_j - 1) = \sum_{k=1,l=1}^{M} \beta_{k,l} \tag{3.7}$$

If $\beta_{k,l} \leq 1$ for all $k, l \in [1, M]$, where M is the number of groups, which implies that the affiliation networks are linear, one arrives at,

$$d_j = \sum_{\substack{\text{All the groups to} \\ \text{which individual } j \text{ belongs}}} (s-1) \tag{3.8}$$

where s is the group size; and

$$u_k = \sum_{\substack{\text{All the individuals} \\ \text{that group } k \text{ contains}}} (m-1) \tag{3.9}$$

where m is the membership number of an individual. When the affiliation network is linear $\alpha_{i,j} \leq 1$.

The adjacency matrix $A^{l(H)}_{N\times N}$ of the line graph $l(H)$ of a hypergraph $H(M,N)$ which represents an affiliation network with M groups and N individuals, can be expressed by the unsigned incidence matrices $R_{M\times N}$ of $H(M,N)$

$$A^{l(H)}_{N\times N} = \left(R^T R\right)_{N\times N} - \mathrm{diag}(R^T R) \tag{3.10}$$

where the entry r_{ij} of R is 1 if node i and hyperedge j are incident, otherwise $r_{ij} = 0$. Basically, the adjacency matrix $A^{l(H)}$ equals the matrix $R^T R$ setting all diagonal entries to zero. The interest-sharing number $\alpha_{i,j}$ of individual i and j equals the entry $a_{ij}^{l(H)}$ of $A^{l(H)}$

$$\alpha_{i,j} = a_{ij}^{l(H)} \tag{3.11}$$

The membership number m_j of an individual j equals,

$$m_j = \sum_{i=1}^{M} r_{ij} = (R^T R)_{jj} \tag{3.12}$$

The group size s_k of group k is

$$s_k = \sum_{l=1}^{N} r_{kl} = (RR^T)_{kk} \tag{3.13}$$

Let $W_{M\times M} = \left(RR^T\right)_{M\times M} - \text{diag}(RR^T)$, then the overlapping depth $\beta_{k,l}$ of two groups k and l equals,

$$\beta_{k,l} = w_{kl} \tag{3.14}$$

where w_{kl} is an entry of $W_{M\times M}$.

The individual degree d_j equals the number of nonzero entries in the jth row/column of $A_{N\times N}^{l(H)}$, with the special case $d_j = \sum_{i=1}^{N} a_{ij}^{l(H)}$ when the affiliation network is linear. Similarly, the group degree u_k equals the number of nonzero entries in the kth row/column of $W_{M\times M}$.

3.3.3. SPECTRAL PROPERTIES

A m-uniform affiliation network can be represented by m-uniform hypergraphs $H_m(M,N)$, of which the unsigned incidence matrix R has exactly m *one-entries* and $M-m$ *zero-entries* in each column. Thus, all the diagonal entries of $R^T R$ are m. The adjacency matrix of the line graph of $H_m(M,N)$ can be written as,

$$A_{N\times N}^{l(H_m)} = R^T R - mI \tag{3.15}$$

where $R^T R$ is a Gram matrix [88, 89].

For all matrices $A_{N\times M}$ and $B_{M\times N}$ with $N \geq M$, it holds that $\lambda(AB) = \lambda(BA)$ and $\lambda(AB)$ has $N-M$ extra zero eigenvalues

$$\lambda^{N-M} \det(BA - \lambda I) = \det(AB - \lambda I)$$

and (3.15) yields,

$$\det\left(A_{N\times N}^{l(H_m)} - (\lambda - m)I\right) = \lambda^{N-M} \det\left(\left(RR^T\right)_{M\times M} - \lambda I\right)$$

The adjacency matrix $A_{N\times N}^{l(H_m)}$ has at least $N-M$ eigenvalues $-m$. Also,

$$x^T\left(R^T R\right)x = (Rx)^T Rx = \|Rx\|_2^2 \geq 0$$

and

$$x^T\left(RR^T\right)x=\left(R^Tx\right)^TR^Tx=\left\|R^Tx\right\|_2^2\geq 0$$

where $x_{L\times 1}$ is an arbitrary vector. Hence, both $\left(R^TR\right)_{N\times N}$ and $\left(RR^T\right)_{M\times M}$ are positive semidefinite, hence all eigenvalues of $\left(R^TR\right)_{N\times N}$ are non-negative. Due to (3.15), the adjacency eigenvalues of $A^{l(H_m)}_{N\times N}$ are not smaller than $-m$.

A non-uniform affiliation network with maximum membership number $m_{\max}$ can be represented by a non-uniform hypergraph $H(M,N)$. The unsigned incidence matrix R of $H(M,N)$ has at most $m_{\max}$ *one-entries* in each column. Therefore, the largest diagonal entry of R^TR is $m_{\max}$. The adjacency matrix of the line graph of non-uniform hypergraph $H(M,N)$ is,

$$A^{l(H)}_{N\times N}=R^TR+C-m_{\max}I \tag{3.16}$$

where $C=\text{diag}\left(\begin{array}{cccc} c_{11} & c_{22} & \cdots & c_{LL}\end{array}\right)$ and $c_{jj}=m_{\max}-(R^TR)_{jj}\geq 0$ for $j\in[1,N]$.

Since

$$\begin{aligned} x^T\left(R^TR+C\right)x&=x^T\left(R^TR\right)x+x^T\left(\sqrt{C}^T\sqrt{C}\right)x\\ &=\|Rx\|_2^2+\left\|\sqrt{C}x\right\|_2^2\geq 0\end{aligned}$$

where $x_{L\times 1}$ is an arbitrary vector and $\sqrt{C}=\text{diag}\left(\begin{array}{cccc}\sqrt{c_{11}} & \sqrt{c_{22}} & \cdots & \sqrt{c_{LL}}\end{array}\right)$, R^TR+C is also positive semidefinite, thus, the adjacency eigenvalues of $A^{l(H_m)}_{N\times N}$ are not smaller than $-m_{\max}$.

EXAMPLES OF AFFILIATION NETWORKS

The arXiv data of subjects of "General Relativity and Quantum Cosmology" (GR-QC) and "High Energy Physics - Theory" (HEP-TH) in the period from January 1993 to April 2003, collected by Leskovec *et al.* [90] can be used to construct an affiliation network and the corresponding hypergraph with the papers as nodes and the authors as hyperedges. A hyperedge is incident to a node if the corresponding author authors or coauthors a corresponding paper. In this manner the hypergraph of the arxiv GR-QC coauthorship network with 5855 authors and 13454 papers, and the hypergraph of the arXiv HEP-TH coauthorship network with 9877 authors and 21568 papers was constructed. The data of s, β, m, d and α can be fitted by a power function $f(x)=x^{-\gamma}$. Values of γ are shown in Table 3.4. In the coauthorship networks of both subjects, papers with only one author and with more than ten authors are very rare. Most of papers have two or three authors. The group degree u follows a power-law tail. The group overlapping depth β follows a power-law distribution as well. Most of the pairs of groups have no overlap. The membership number m of an individual denotes the number of papers he or she authors and coauthors which also follows a power-law distribution. The interest-sharing number α, denoting the number of papers in which two individuals participate together, is also well aproximated by a power-law distribution. The ArXiv coauthorship networks of both subjects possess high clustering coefficient, positive assortativity and short average path length as shown in Table 3.5.

Data of the IMDB movie actors collaboration network with 127,823 movies and 392,340 actors from the Internet Movie Database (based on www.imdb.com) was used as described in A.1.6. In the hypergraph of IMDB movie actors collaboration, the movies are

nodes and the actors are represented as hyperedges. A hyperedge is incident to a node if the corresponding actor appears in the corresponding movie. The parameters of fitted data of s, u, β, m, d and α to a power function $f(x) = x^{-\gamma}$, are shown Table 3.4. Data of s is fitted with two power-law functions in different regions. The group degree u appears also to follow two power-law distribution in two regions. All the values of γ are shown in Table 3.4. The IMDB movie actors collaboration network exhibits high clustering, assortative mixing and short average path length as shown in Table 3.5.

SourceForge is a web-based project repository assisting programmers to develop and distribute open source software projects. SourceForge facilitates developers by providing a centralized storage and tools to manage the projects. Each project has multiple developers. The hypergraph of the SourceForge software collaboration network was created by taking software projects as nodes and the developers as hyperedges. A hyperedge is incident to a node if the corresponding developer participates in the corresponding software project. The SourceForge software collaboration network has 259,252 software projects and 161,653 developers.

MODELING OF SOCIAL NETWORKS WITH OVERLAPPING COMMUNITIES

One needs to notice that the number of groups M is larger than the number of individuals N in the ArXiv and Sourceforge network, and vice versa in the IMDB network. Making a movie seem to need more labor than writing a paper or developing an open-source software. In a growing hypergraph model, one may take $\frac{M}{N} = 1$, assuming that each coming individual start a new group. Note that the group size of real-world affiliation network follow a power-law distribution. Employing preferential attachment of individuals to the existing groups to achieve power-law distributed group size may model empirical observations. The tricky issue is to determine the membership number of each arriving individual, namely to decide how many nodes a new hyperedge should connect to.

The hypergraph model is described by the following procedure:

1. Start with a seed hypergraph $H_0(M_0, N_0)$ with M_0 groups and N_0 hyperedges.
2. Suppose that the desired number of individuals (hyperedges) of the network to be generated is $N + N_0$. Determine the membership numbers for N new hyperedges: $\Gamma = \begin{bmatrix} \bar{m}_1 & \bar{m}_2 & \cdots & \bar{m}_N \end{bmatrix}$. The membership number vector Γ is a input parameter to the hypergraph model.
3. At growing step j, $j = 1, 2, \cdots, N$, add a new hyperedge j and a new group to the hypergraph. Make the new hyperedge j and the new group incident, and the membership number of j becomes 1.
   (a) Connect the new hyperedge j to the existing group k with probability $p_k = s_k / \sum_{i=1}^{j-1} s_i$, where s_k is the group size of group k and $\sum_{i=1}^{j-1} s_i$ is the sum of group sizes of all the existing groups.
   (b) Repeat 3a) $\bar{m}_j - 1$ times so that the membership number of the hyperedge j increases to the expected membership number $\bar{m}_j$.
4. Repeat 3) until the number of hyperedges increases to $N + N_0$.

Algorithm 3 Growing hypergraph model

IN: A seed hypergraph $H_0(M_0, N_0)$ with M_0 nodes and N_0 hyperedges, The membership numbers for new hyperedges $\Gamma = [\bar{m}_1\ \bar{m}_2 \cdots \bar{m}_N]$
OUT: A hypergraph $H(N+M_0, N+N_0)$

$H \leftarrow H_0(M_0, N_0)$
for each $j \in \{1,2,3,\cdots,N\}$ **do**
  add a new hyperedge j to H
  $m_j \leftarrow 0$
  add a new node to H and let it be incident to the hyperedge j
  $m_j \leftarrow m_j + 1$
  while $m_j < \bar{m}_j$ **do**
    $k \leftarrow$ a random natural number between 1 and $j-1$
    $r \leftarrow$ a random real number between 0 and 1
    if $r < s_k / \sum_{i=1}^{j-1} s_i$ **then**
      let the hyperedge j be incident to the node k
      $m_j \leftarrow m_j + 1$
    end if
  end while
end for

The model is also presented with pseudo-codes in Algorithm 3. Compute the metrics d_j, m_j, $\alpha_{i,j}$, s_j,u_j and $\beta_{i,j}$ using the methods given in Section 3.3.2 including the formulas (3.10) to (3.14).

PROPERTIES OF THE GROWING HYPERGRAPH MODEL

A hypergraph $H(20,20)$ is used with the membership number $m_j = 1$, $j = 1,2,\cdots 20$, as the starting seed. A total of 5,000 new hyperedges (individuals) and 5,000 new nodes (groups) were added to the starting seed through 5,000 growing steps. Hence, all the generated hypergraphs had 5,020 nodes and 5,020 hyperedges.

In the growing process, the constant membership number $m_j = 2$, $j = 1,2,\cdots,5000$ is applied, obtaining the uniform hypergraph H_2. In the same way, H_3, H_5, H_7, H_{10} and H_{15}were constructed. Then hypergraph $H_{U[1,100]}$ was created with a uniformly distributed membership number in the interval $[1,100]$. These hypergraphs are constructed in order to study the properties of H_{pow} which is obtained by applying the sequence of membership numbers with the pdf $\Pr[\Gamma = m] = m^{-2.02}$.

The group size and group degree of a random group are denoted by S and U, the group overlapping depth of a random pair of groups by B, the individual degree of a random individual by D, and the interest-sharing number of a random pair of hyperedges by Φ.

Due to the principle of preferential attachment [91], one may expect that the group size of all the generated hypergraphs will follow power law distributions, which was empirically confirmed. The exponents of the power laws are shown in Table 3.4.

Nacher *et al.* [92] and Manka *et al.* [93] showed that the nodal degree of line graphs of simple graphs with power law degree distribution follow a power law distributions

Table 3.4: The exponents γ of power-law fittings $f(x) = x^{-\gamma}$ of s, u, β, m, d and α of the arXiv GR-QC and HEP-TH coauthorship networks, the IMDB actor collaboration network, the SourceForge software collaboration network, and the growing hypergraph model with different sequences of membership numbers.

Network	$\gamma(s)$	$\gamma(u)$	$\gamma(\beta)$	$\gamma(m)$	$\gamma(d)$	$\gamma(\alpha)$
ArXiv GRQC	5.50	2.14	3.93	1.95	1.84	3.56
ArXiv HEP-TH	6.24	1.63	3.56	1.72	1.68	2.86
IMDB actors	2.04/5.35	0.407/3.40	4.80	1.81	1.91	3.62
SourceForge	3.91	2.45	3.76	3.48	2.61	4.60
H_2	2.12	2.39	3.38	n.a.	2.35	n.a.
H_3	2.55	2.46	3.07	n.a.	2.16	n.a.
H_5	2.38	2.09	3.19	n.a.	2.12	n.a.
H_7	3.06	2.81	3.11	n.a.	2.59	n.a.
H_{10}	3.22	2.22	3.53	n.a.	2.38	n.a.
H_{15}	2.90	1.95	3.34	n.a.	2.66	n.a.
$H_{U[1,100]}$	3.66	2.85	3.82	n.a.	3.01	n.a.
H_{pow}	3.91	2.45	3.76	3.48	2.61	4.60

as well. The individual degree distribution of H_2 is just the degree distribution of line graphs of scale-free graphs.

The clustering coefficients C, the assortativity coefficients ρ_D and the average path lengths l of all the generated hypergraphs H_2, H_3, H_5, H_7, H_{10}, H_{15}, $H_{U[1,100]}$ and H_{pow} are reported in Table 3.5. All the generated hypergraphs exhibit high clustering coefficient, positive assortativity and short average path lengths as reported and shown for real-world affiliation networks show.

Many real-world networks, especially social networks, exhibit an overlapping community structure. Affiliation networks are an important type of social networks. The proposed hypergraph representation reproduces the clique structure of affiliation networks. The topological and spectral properties of affiliation networks are shown analytically, and formulas were presented which facilitate the computation for characterizing the real-world affiliation networks of ArXiv coauthorship, IMDB actors collaboration and SourceForge collaboration. Numerical analyses show that the proposed hypergraph model with power-law distributed membership numbers reproduces the power-law distributions of group size, group degree, overlapping depth, individual degree and interest-sharing number of real-world affiliation networks, and reproduces the properties of high clustering, assortative mixing and short average path length of real-world affiliation networks.

3.4. USEFULNESS OF FRIENDSHIP RELATIONS

As several hundred million Internet users regularly frequent OSN sites as a place to gather and exchange ideas, researchers have begun to investigate how this comprehensive record can be used to understand how and why users join a community, how these networks grow by friendship relations, how information is propagated among friends, and who are the most important and influential users in such social groups. A good understanding of these principles would enable many application scenarios, such as the prediction

Table 3.5: The clustering coefficients C, the assortativity coefficients ρ_D and the average path lengths l of the arXiv GR-QC and HEP-TH coauthorship networks, the IMDB actor collaboration network, the SourceForge software collaboration network, and the growing hypergraph model with different sequences of membership numbers.

Network	C	ρ_D	l
ArXiv GRQC	0.637	0.584	6.50
ArXiv HEP-TH	0.289	0.382	4.89
IMDB actors	0.762	0.682	4.29
SourceForge	0.636	0.401	7.06
H_2	0.616	0.508	6.13
H_3	0.581	0.576	6.71
H_5	0.491	0.498	7.85
H_7	0.613	0.644	7.62
H_{10}	0.686	0.519	6.89
H_{15}	0.722	0.478	6.56
$H_{U[1,100]}$	0.566	0.422	7.22
H_{pow}	0.636	0.401	7.06

of elections, competitions and trends [94], effective viral marketing [95], targeted advertising [96] or the discovery of experts and opinion leaders [97].

These investigations and applications in social networks however make the fundamental assumption that the friendship relations between users are a critical ingredient for the proper functioning of social networks [98], i.e., they assume that information, opinions and influences are sourced by single individuals and then propagated and passed on along the social links between members of the community. The extent, density, layout and quality of the social links and the network of links as a whole will therefore determine how information can be spread effectively.

However, the importance of individual friendship relations and the friendship network as a whole is less than previously perceived. In these social news aggregators, users submit news items (referred to as “stories”), communicate with peers through direct messages and comments, and collaboratively select and rate submitted stories to get to a real-time compilation of what is currently perceived as “hot” and popular on the Internet. Yet, despite the many possible means to communicate, interact and spread information, an analysis of ten million stories and the commenting and voting patterns of two million users over a period of four years revealed that the impact of the friendship relations on the overall functioning and outcome of the social network is actually surprisingly low. Users indeed form friendship relations according to common interests and physical proximity but these friendship links are only activated with 2% probability for information propagation. Furthermore, in about 50% of all stories that became “hot”, there was no prior contribution by the friend network to the extent that would have led to emerging popularity of the story; instead, a critical mass was reached through participation of random spectators.

COMMON ASSUMPTIONS ABOUT THE IMPORTANCE OF FRIENDSHIP RELATIONS

Ever since the publication of Katz and Lazarfeld's argument for the origin and spread of influence through communities [99], researchers have investigated the mechanisms by which ideas and opinions are passed along social relationships. Since then, the role of individuals, as well as the characteristics and importance of their relationships have been investigated in a variety of different research fields.

A common way to describe the structure and relations between individuals in a community are by "weak" and "strong" ties. Originally proposed by Granovetter [47] in a sociological context based on the intensity, frequency and amount of personal contact between individuals, this characterization of interpersonal relationships has spread and been adopted by many other subject domains, such as marketing, political science and economics. According to the theory, weak and strong ties behave differently in communication and information dissemination: while a lot of interaction is taking place between "strong ties", i.e. persons with frequent and long-lasting contacts, these ties within tightly knit clusters carry a lot of redundant information; thus new and novel information can best enter from outside these clusters across "weak ties".

These aspects of novel information transmission and redundancy in weak and strong ties are further analyzed by Burt [49] within the context of organizational networks, who finds that information transfer in a company is best achieved when individuals possess a high number of overall, but relatively low number of redundant contacts. People switching between different positions within an organization keep their previous ties, and companies with a well-connected social network are exhibiting a larger agility to react to problems. Burt refers to areas within organizations having too few or too weak "weak ties" as structural holes. Similar findings are also reported by Krackhardt [100] who investigated the importance of informal interpersonal networks in organizations in times of crises. In a game, two hypothetical companies were created in which the units in one company contained friends working together in one division and in the other company friends had been in different units. During crises, simulated through a drop in available resources, the organization having a well-connected network of units performed significantly better than the other one.

Hansen [101] added to this so-called search transfer problem the notion that besides the existence of weak and strong ties, the absolute strength of a connection is also of noteworthy importance. In a study of information sharing between subunits of large multi-national companies, well-connected units again scored better than others, but among equally well-connected subunits the ones with more intense ties performed even better due to increased collaboration. When sharing complex knowledge, weak connections did provide exposure and information about possible solution approaches, successful adoption however was aided by an increased intensity of the social tie.

When facing problems or difficult questions, we turn to friends or acquaintances around us to get clues or a solution, as claimed by Homans [102] and Coleman *et al.* [103, as cited in [104]]. Consequently, social interactions and ultimately the social network provide a fertile ground for the promotion of new ideas, information and innovation. A prime example to assess such knowledge dissemination is Coleman et al.'s study "Medical Innovation" [103], investigating whether and how a group of physicians are adopting a new drug after recommendations from their social network. Later reanalyzed in [104],

Burt however does not find convincing evidence that social ties were indeed the driving force behind the adoption of the new medication, as the number of ties to physicians who had adopted the drug had no influence on whether a physician was prescribing it in turn, which should occur in a contagion process.

Tsai [105] investigated the knowledge dissemination measured in terms of innovation ability within an organizational network of 60 business units and argues that the ability to obtain beneficial information depends on the location within a social network. Individuals or groups placed and connected in the center of the network get more exposure to information simply through their topological location in the system, which in his example manifested itself in higher innovation performance.

Although the potent abilities of networks to transmitting information and relaying messages have been known for quite some time since Milgram [1] conducted his famous experiment, leading to the previously mentioned well-known phrase “six degrees of separation", interactions between the individual and the surrounding social network have received in the recent past new and diverse attention. For example researchers in human dynamics, public health or epidemiology found on social networks. Christakis [7] for example demonstrated in a longitudinal study of some 12,000 participants that a person’s risk of becoming obese, one’s ability to stop smoking or maintain happiness is to a significant extent influenced by those surrounding the person. Both benefits and risks are being propagated by social ties, and the influences are contagious between friends and friends of friends. These recent outcomes have lead to new insights and impulses in public health, for example that certain diseases are better approachable at the individual and the network level or that epidemics may be more efficiently prevented under resource constraints when first protecting the high-degree entities by vaccination in the network, who act as fast spreaders of ideas and disease. The same underlying principle of the high importance of high-degree nodes can in turn however also create a significant problem, as these hubs pose a significant vulnerability in scale-free network topologies, for example when targeted by malicious attacks [106].

In recent years, two distinct trends have emerged in network analysis: First, with the availability of new data sets and fast processing options, the focus has shifted from empirical observation of small groups of a few dozen to a few hundred participants in larger studies. With the advent and wide-spread popularity of online social network platforms, this field of study has gained additional momentum as these newly available communities now provide an easily accessible, machine-readable data source for a broad-scale analysis of established research topics.

Second, the bulk of recent work has investigated the structural properties of complex networks, with a lesser focus on understanding the friendship and information propagation processes taking place inside such large scale social networks.

Along these lines, Mislove *et al.* [107] studied the topological properties of four OSNs at large-scale: Flickr, YouTube, Live-Journal, and Orkut. By crawling publicly accessible information on these networking sites, they evaluated different network metrics, e.g. link symmetry, node degree, assortativity, clustering coefficient of the four networks. The OSNs investigated were characterized by a high fraction of symmetric links, and composed of a large number of highly connected clusters, thereby indicating tight and close (transitive) relationships between users. The degree distributions in the OSNs follow a

power law and the power law coefficients for both in-degree and out-degree are similar, showing the mixed importance of nodes in the network - there are few well connected and important hubs to which the majority of users reach to.

In [108], Leskovec *et al.* presented an extensive analysis about communication behaviors and characteristics of the Microsoft Messenger instant-messaging (IM) users. The authors examined the communication pattern of 30 billion conversations among 240 million people, and found that people with similar characteristics (such as age, language, and geographical location) tend to communicate more among each other. The constructed communication graph was analyzed for topological properties of the graph in terms of node degree, clustering coefficient, and the average shortest path length; it was shown that the communication graph is well connected, robust against node removal, and exhibits the small-world property.

Backstrom *et al.* [109] studied the network growth and evolution by taking membership snapshots in the LiveJournal network. They also presented models for the growth of user groups over time. Benevenuto *et al.* [110] examined users activities on Orkut [111], MySpace [112], Hi5 [113], and LinkedIn [114]. A clickstream model was presented in [110] to characterize how users interact with their friends in OSNs, and how frequently users transit from one activity (e.g. search for peoples profiles, browse friends profiles, send messages to friends) to another. There are also many researchers who aim to discover content popularity and propagation in OSNs. For instance, in Cha *et al.* [115], photo propagation patterns in the OSN Flickr are studied. The results discovered in [115] reveal different photo propagation patterns, and suggest that photo popularity may increase steadily over years. An in-depth study about content popularity evolution and content duplication was performed with YouTube and Daum (a Korean OSN) in [116]. In this publication, Cha *et al.* studied the popularity distribution of videos uploaded to the two websites. It was shown that video popularity of the two applications is mostly determined at the early stage after a video content has been submitted to the OSNs. A similar comparison was conducted for YouTube and Digg by Szabo and Huberman [117], who presented a model of predicting the long term popularity of user generated content items.

Besides high-level observations, there is only little known yet about the exact content propagation mechanisms taking place inside an online social network and the possible roles and impact different types of users might assume during information dissemination. If unearthed, such insights would have many application domains, ranging from the discovery of experts and opinion leaders to efficient innovation adoption and marketing. For this reason, the search for the “influentials” has been a significant endeavor in the viral marketing literature where it is argued that “few important trends reach the mainstream without passing the influentials in the early stages, ... they give the thumbs-up that propel a trend". Their recommendation and word-of-mouth dissemination lets information spread exponentially [118, p. 124].

3.4.1. INFORMATION SPREAD THROUGH THE NETWORK OF FRIENDS

This section will dissect the process of friendship relations defining the success of information dissemination and investigate for the case of the Digg OSN, whether the propagation of news is indeed the result of the activation of user ties.

SELF-ORGANIZATION OF THE FRIENDSHIP NETWORK

According to sociological theory, friendship relations in OSN grow directed by common interests and tastes [36]. Within the Digg social network, all news stories are classified within eight major topic areas, further subdivided by 50 special interests. When matching the users' concrete digging behavior with the topic area a story was classified in, one observes that the subscribers exhibit quite strong and distinct preferences and tastes for individual topic areas: Even when following the content published in several genres, most of their attention is focused on a few areas.

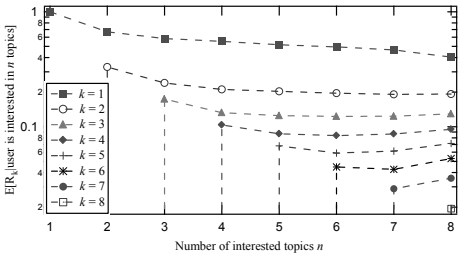


Figure 3.13: The share of diggs a user devotes on average on the 1^{st},2^{nd},...k^{th} most frequented topic areas (y-axis, in logarithmic scale) as a function of the total number of categories a user has been active in (x-axis).

As shown in Figure 3.13, if a particular user reads, diggs and is therefore interested in two distinct topic areas, say for example "Science" and "Technology", almost 70% of all consumed stories fall within the most preferred genre. For three subscribed topic areas, say for example "Lifestyle", "Business" and "Entertainment", the ratios drop to 65%, 25% and 15%, thus the most preferred topic still attracts on average nearly two thirds of all clicks. Even for users interested in eight categories the top two will on average account for 60% of read stories.

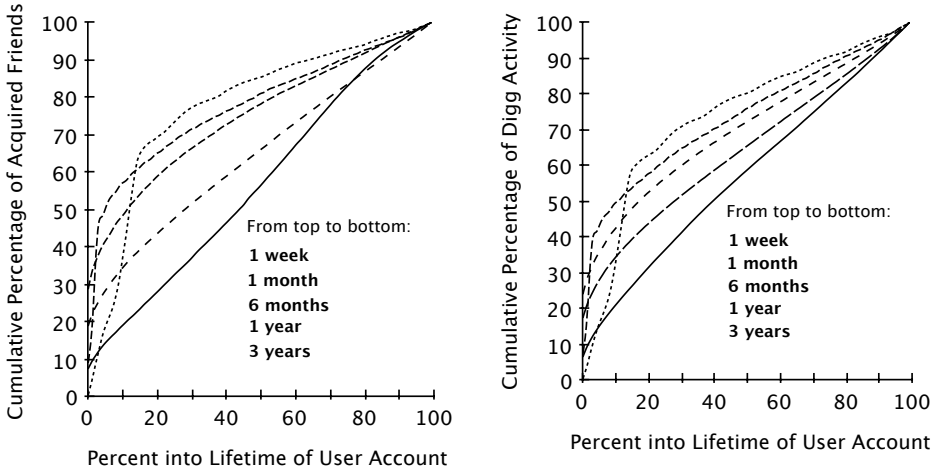
Since the relative preferences between categories are quite pronounced and stable, these ranks of user interest provide a direct measure of how similar the tastes and preferences of users in their information acquisition are.

When comparing the interests between two users and their ranking of topics, the "similarity hop", can be used to reflect the distance of a user's favorite topic with respect to the k-th favorite topic of his friends [77]. The k-th topic after ranking is denoted as $T_{(k)}$. For a friend pair i and j, two sets $\{t_{i(1)}, t_{i(2)}, ..., t_{i(8)}\}$ and $\{t_{j(1)}, t_{j(2)}, ..., t_{j(8)}\}$, define the ranked topics, in which $t_{i(k)}$ and $t_{j(k)}$ are the names of the k-th favorite topic of user i and user j, respectively. Since $t_{i(1)}$ is the most favorite topic of user i, one may compare $t_{i(1)}$ with $t_{j(k)} \in \{1 \leq k \leq 8\}$ of user j. The similarity hop, defined in (3.17), measures how similar two friends are regarding their preferred tastes.

$$h_{ij} = (k-1)\, 1_{\{t_{i(1)} = t_{j(k)}\}} \tag{3.17}$$

The indicator function, $1_{\{x\}}$ is defined as 1 if the condition of x is satisfied, else it is zero. The similarity hop h_{ij} ranges between $0 \leq h_{ij} \leq 7$. A value of zero denotes that two users have identical interests. A small similarity value, say $h_{ij} = 1$, indicates high overlapping interests between two friends. While a large number (e.g., $h_{ij} = 7$) suggests that users do not have common interests.

A network-wide analysis of the similarities between friends shows that users directly connected to each other have a very high alignment of their preferences and tastes: 36% of rank lists are identical, 20% require one transformation, and within three transfor-



(a) Friendship acquisition

(b) Digg activity

Figure 3.14: Acquisition of (a) friends and (b) diggs through the lifetime of user accounts. The y-axis shows the cumulative percentage of digging/friending activity to date as a function of the cumulative lifetime of a user account on the x-axis, defined as the total timespan between registration and the last activity of a particular account.

mation steps 80% of all friendship relations are aligned. People acquire and maintain friendships based on whether these future friends have previously demonstrated a similar taste and composition in their behavior on Digg.com.

Interestingly, the rate at which a user initially acquires friends and diggs on stories seems to be related to the overall lifetime of the user's account, determined as the timespan since the first registration until the last action performed by this account. Visitors who sign up and immediately form a lot of friendship relations within their first day but slow down on their second, typically abandon their profile after one week or less. The slower and more continuous friends are added to a profile, the longer a person continuous to participate on the Digg website, as shown in Figure 3.14. The most sustainable rate of digging activity and friendship acquisition is exhibited by those who remain active for 3 years and thus can be considered heavy users of the platform.

INCENTIVES FOR COMMON DIGGS

While there exists a perfect overlap between the interests and tastes of individual friends, there is a surprisingly low amount of common activity among friends and on average only 2% of all friend pairs actually do react and digg on the same story.

The hypothesis that common interests result in the formation of friendships in order to gain information from neighboring peers [119] would also predict that the more similar the tastes between friends are, the closer the alignment of clicking patterns would be. In practice, this however seems not to be entirely the case; although there is a generally decreasing trend between interest overlap and common clicks, the differences are not statistically significant.

ACTIVATING THE FRIENDS OF FRIENDS

Friends and friendship pairs however do not exist in isolation, but are embedded within a larger network of the friends of friends. This very dense structure in OSNs as also shown by Mislove *et al.* [107] may work as a powerful promoter, as theoretically a large number of nodes can be reached if information can be passed on from a friend to a friend and propagated over several steps: In theory, an information may reach an exponentially growing number of recipients as the number of hops it traverses increases. Given that there exists a critical threshold that needs to be met to promote a story to high popularity and a limited number of friends are actually active on the site on a particular day, the *network* of friends, in other words the friends of friends, could make the difference between stories that spread or fall into oblivion.

The analysis shows that information can indeed travel over multiple hops from the original submitter in the Digg OSN (see Figure 3.16(a)) and on average reaches 3.7 hops from the source until the propagation dies down. The actual contribution of the multi-hop network, i.e. the amount of friends of friends that can actually be activated by this process, is however rather limited. As shown in Figure 3.15, nearly 70% of the ultimately participating *network* of friends consists of the submitter's direct contacts, while the incremental benefit of the additional hops decreases exponentially. This result is not astonishing given the generally low activation ratios of friends and possible redundancies in the spread as indicated by the dashed line in Figure 3.16(a), i.e. a person receiving several notifications from various friends in the previously activated friendship network.

This aspect is further visualized in Figure 3.16(b), which shows the share of the total redundant notifications observed at a particular distance from the original source. A notification can be classified as redundant if a particular user has been informed about a particular story before and the incoming trigger consequently provides no additional information, or if a notification arrives after the receiving user has already digged on a story earlier on.

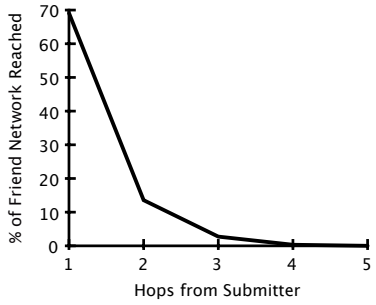
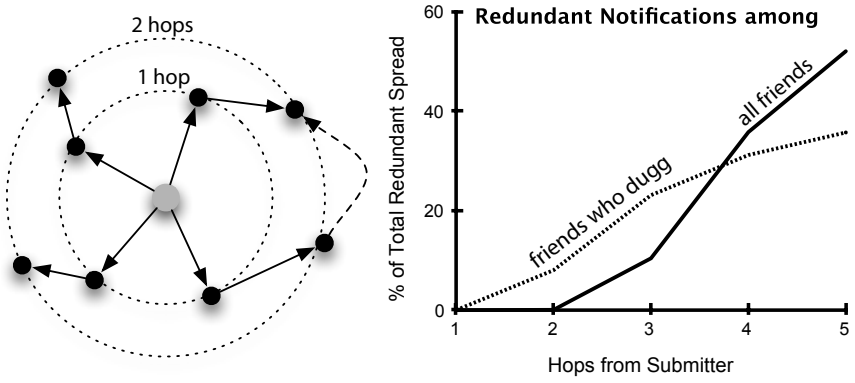


Figure 3.15: Activation of the friendship network in information spread. The Figure shows the percentage of the total activated friendship network as the information spreads out hop-wise from the original submitter.

As can be expected, due to the tree topology of the first hop friendship network, no duplicate notifications are initially generated, while the number of redundancies increases rapidly as the spread progresses, both as a result of back-links into the already explored network and due to exhaustion of the pool of possible candidates. The slope of both curves shown in Figure 3.16(b), the redundant notifications within the activated friendship network indicated by the dashed line and the theoretical maximum of redundant notifications if all friends would react to an incoming trigger indicated by the solid line, is however bounded: the former one declines with a dwindling network activa-



(a) (b)

Figure 3.16: Redundant activation in the friendship network. As the friendship graph contains common acquaintances, a spreading process along the friendship links will result in duplicate, redundant notifications as indicated by the dashed arrow in subfigure (a). Subfigure (b) quantifies the percentage of the total redundant notifications passed along friendship links (y-axis), depending on the stage in the spreading process (x-axis).

tion after three hops, the latter one slows down as the network and all friendship links are getting saturated.

REACHING CRITICAL MOMENTUM

All news stories submitted to the Digg social network are initially collected in the "upcoming" list, which with more than 20 000 submissions per day has a very high turnover rate (more than 800/h) and a total capacity of 24 hours after which stories will disappear. In order to become promoted to the frontpages, a story therefore has to attract sufficient interest, i.e. a large enough number of diggs, within this timeframe of 24 hours. As shown in Figure 3.17, the majority of stories that passes this threshold does so after the initial 16 hours. We experimentally determined that about 7 diggs per hour are necessary to qualify for the promotion, thereby stories should gather on average around 110 diggs.

A story can rally this support initially from random spectators or friends of the submitter, who were notified about the newly placed story. To successfully spread via friendship links, a critical mass of friends needs to vote on the item. For this to happen however (assuming that all or a high percentage of them will react to the incoming notification), a sufficient number of friends first need to be active and active on the Digg.com website within this promotion window to become aware of the story and be able to contribute to its promotion. This probability can be inferred from previous records, as the data set contains all instances when a particular user submitted, digged and commented on stories or created friendship links since account registration. Thus, an analysis of the combined actions of a particular user provides a lower bound[2] of that person's probability to be active during a typical 24 hour time window. Combining such estimates with the

[2] As the user could have been active without having been logged in or visited and seen the website without performing any action visible in the logs.

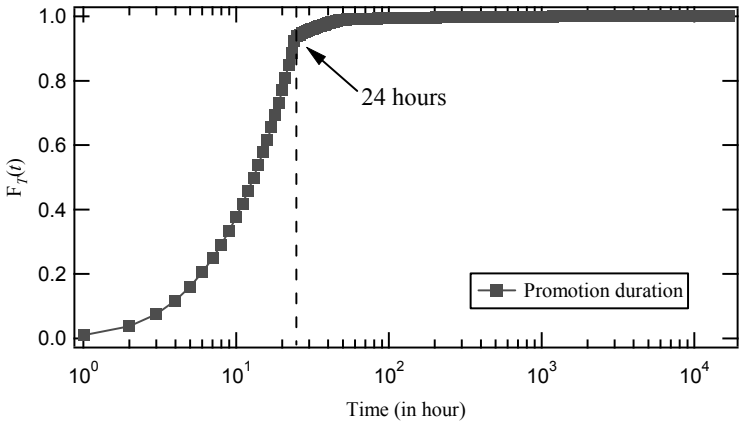


Figure 3.17: Promotion probability of submitted stories over time. Stories have to gain initially enough momentum within 24 hours to be selected from the pool of fast-moving submitted news items. The figure shows the cumulative probability for a story to become promoted within a particular time period after original submission.

structure of the submitter's friendship network provides an approximation of the probability that a particular number of friends are active during the promotion window.

Figure 3.18 shows the average likelihood for a given number of friends to be active on the website on the same day, and therefore in theory be available to provide the required support. While the probability that the required 110 friends are indeed present corresponds with the actual promotion success ratio of 0.01, this fine line between failure and success strongly depends on the performance of the underlying stochastic process, whether at a certain time a sufficient number of friends are online and willing to support the story. In the remaining 99% of the cases, additional support needs to be rallied from users outside the submitter's friend network.

3.4.2. ARE USERS FOLLOWING THE HERD?

As the impulse of a user to follow a friend's previous action is relatively low, it might simply be that more than one trigger event is needed to activate a user. There exists an established body of literature on behavioral mimicry [120], indicating that people are subconsciously copying the behavior of those around them; for example, it has been reported that the likelihood for a person to buy a computer is influenced by how many computers are owned within that person's neighborhood [121].

As the data set contains both the social relationships and actions of users, one may use this combined information to quantify to what extent such network externalities are indeed influencing the behavior of individual users, i.e. does a person's likelihood to recommend some content depend on the number of friends that have previously reacted positively to a particular item? For this behavioral mimicry to unfold, a chain of conditions however needs to be met: (1) There needs to be some mechanism that lets a person

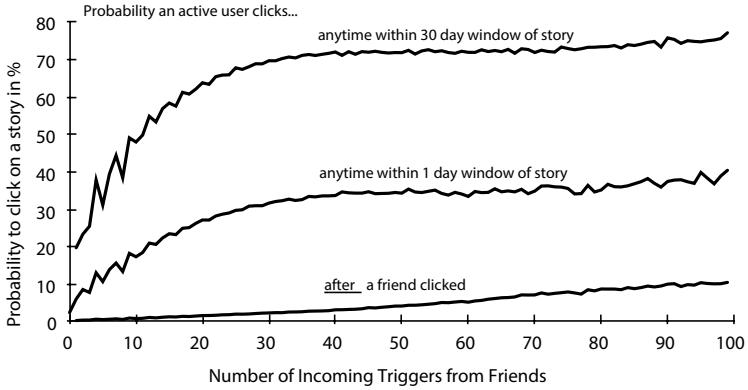


Figure 3.19: Probability of user activation after triggers. The likelihood for a user to digg on a story (y-axis) in principle increases with the number of diggs performed by the friendship network (x-axis). The effect however is drastically limited when only considering a window of 1 day, the time from submission until the promotion cut-off date when a user's contribution will have the most effect, instead of the total 30-day lifetime of a promoted story. Additionally, when also enforcing the requirement that notifications are strictly arriving before the receiving user has digged, the probability to digg even after a high number of friendship network votes drops to less than 10%.

learn and observe the behavior of those around them. (2) The person needs to be active and able to receive and perceive the surrounding triggers. (3) A trigger needs to be timed in such a way that it can serve as an influencer to a person's behavior.[3] (4) If possible, a causal relationship between trigger and action should be established.

In the case of Digg, the activities of users are publicly visible to everyone, and by establishing a friendship link users can keep track of their friends' activities through notifications. As friends might not be able to receive and view such notifications due to abandoned accounts, extended absences or non-aligned activity periods (an issue further discussed in section 3.4.3), only those users will be considered during the analysis that were active at least once on the Digg website during a particular story's life time, and thus could in theory have received triggers resulting from their friends' activities.

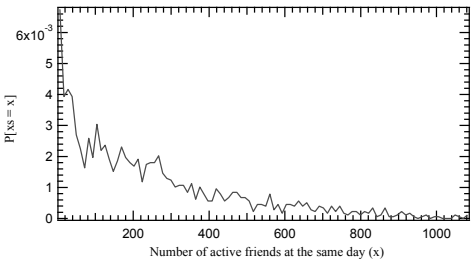


Figure 3.18: The figure displays the probability (y-axis) for a given number of friends (and friends of friends) to be online within same day (x-axis).

For these generally active users, Figure 3.19 shows the probability that a person will click on the same story as one of their friends, depending on the total number of triggers

[3] For the case of the adoption of computers as presented in [121] for example, the person should have bought a computer after those around it have done so. If for example the computer has been ordered months ago but not delivered yet, intermediate purchases from friends and neighbors could not have served as a trigger to that person's decision.

Table 3.6: Ratio of friends and non-friends among the total number of diggers for popular stories. The table lists the share of diggs coming from the submitter's friendship network out of all diggs, both before and after reaching the promotion threshold.

	Before popular		**After popular**	
Average ratio	Friends	Non-friends	Friends	Non-friends
a) friend-promoted	0.72	0.28	0.25	0.75
b) non-friend promoted	0.23	0.77	0.14	0.86

received through their following relationships. As can be seen in the figure, this likelihood significantly increases with the number of incoming notifications, but saturates beyond 40 triggers. The overall activation level however highly depends upon the time frame of observation: If a user may react anytime within a 30 day time window, given enough triggers users on average click nearly on 75% of those stories as their friends have done before. This 30 day time window is however the maximum lifetime of promoted stories, which with 1% of all submitted stories ($\approx$ 160 daily) only encompass a small fraction of the overall news content on the site. New submissions have to reach the promotion threshold within 24 hours and for this time window the maximum saturated activation probability across all stories drops to about 35%.

The presented activation ratios (and many of those studied in the previous literature) have so far only looked at a user's individual behavior and the existence of a social relationship, in other words any temporal information is not yet utilized, that can help answer whether the flow of information was aligned in such a way that the incoming trigger could indeed have initiated the resulting behavior by the follower. By making this distinction and only consider diggs that have been made on a story *after* a friend has digged on it, the probability of a reaction drops below 3% for a low to moderate number and below 10% for even 100 incoming notifications. Given that an active user receives on average 13.7 triggers, this further explains the low conductivity of friendship links, and the resulting linear relationship between number of triggers and digging probability can be reduced to a stochastic counting process [122]. It can be expected that the percentage of *causal* triggers will even be considerably lower; to establish this number however direct feedback from participants would be necessary explaining the motivation for every digg.

PROMOTION WITHOUT FRIENDSHIPS

The fact that the likelihood that a story can become popular solely through the activity of the submitter's friendship network is rather slim (given the slow activation ratio of friends, the limited contribution of the network of friends and low probability of a sufficiently large critical mass of friends that are active on the same day), in most cases the contribution of non-friends is necessary to promote a story up to the threshold level.

When analyzing the ratio of clicks from friends in the submitter's network to the total number of diggs before reaching the promotion threshold, the body of stories can be divided into two distinct groups - one with a high average contribution of friends and one with a low average contribution.

Table 3.6 shows the ratio of friends and non-friends active on a story both before and

after the promotion for all stories that became popular within the Digg network, divided into two groups using the arithmetic mean of friendship contribution ratios of popular stories with a friendship contribution (50%) as a dividing threshold. Stories with more than 50% friendship network contribution were tagged as (a) “friend promoted", with less than 50% as (b) “non-friend promoted".

Although being a rather simplistic decision point, it provides a rather pronounced differentiation of all stories into two groups. In about 54% of all cases, a story was marketed predominantly by friends, although a contribution of non-friends (28%) was necessary until the story reached critical mass. Figure 3.20(a) shows this aggregated pattern for a prototypical story from this class; in the beginning of the stories' lifetime, the submitter's friends dominate the process until about one hour before the promotion is reached, a number of unrelated users push the story over the threshold. In the remaining cases (46%), stories were spread and consumed predominantly by users outside the submitter's friendship network. Figure 3.20(b) shows a prototypical example for this pattern. Once the promotion threshold is crossed, both types of stories are read more by non-friends, as the quantity is usually significantly larger and the possible contribution of the submitter's friendship network may have already been exhausted. At this time stories also experience an immediate and drastic boost in the number of incoming diggs due to the prominent placement at the top of the Digg home page. This effect however quickly dampens down again as other more recently promoted stories displace the item from its prime position and the story moves on to later front pages. Experimental measurements have determined that stories attract practically no notable number of diggs after front page 4 or 36-48 hours.

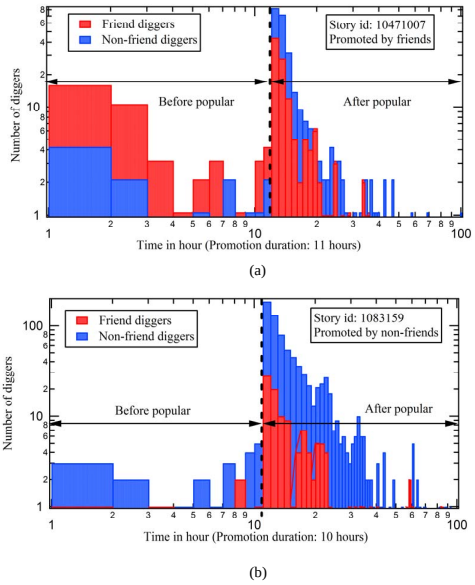


Figure 3.20: Comparison of friend/non-friend digging activities over a story lifetime.

THE CRITICALITY OF INDIVIDUALS

As previously shown, the successful spread of information cannot be explained directly from the social ties inside the investigated online social network, neither through the relationships among individual friends nor from the usage and outreach of users into their friendship network, in other words activating the larger body of friends of friends. In both cases, the average activation of users is generally too small to cross the threshold to criticality, thus resulting in the fact that only 1% of all stories and items submitted to the network ever reach popularity and in only 50% of the popular stories this promotion is due to the action of friends. This however naturally raises the question whether

all users are equal inside the network, or whether there are some individuals in the social community (a) who themselves have better (or earlier) access to important content and are therefore able to submit a high number of stories that will become popular, (b) can use their friendship network more efficiently, act as motivators and are able to overpropotionally recruit friends to click and spread the word, or (c) are able to early on spot content that will later resonate with the masses and become a hit. These questions will be the focus of this section.

There exist a number of ways to define the importance or criticality of individuals in networks. In complex network theory and social network analysis, importance is typically defined from a structural perspective, using topological metrics such as node degree or betweenness [123], which measure how well a particular node is connected to its surrounding peers and how many theoretical communication paths between nodes in the network will pass this entity en route.

Based on this definition, most studies of online social networks find a small number of topologically critical nodes [107, 108, 110], resulting from the typical power-law degree distribution of these complex networks; there exist a few well-connected nodes with whom a large number of users are friends. In this analysis, these findings can be confirmed and will thus serve as a definition to study critical individuals.

Contrary to other online social networks however, one may not only observe a skewed distribution in the degree and connectivity of nodes, but also in the symmetry of relationships among users. While most OSN show high levels of link symmetry[4], for example 74% of links in LiveJournal and 79% of links in YouTube are found to be bi-directional [107], the relationships in Digg are less reciprocative (38% on average) and also vary with the degree of the node: the more connections an individual B already has, the less likely it is to match an incoming new friendship request from A. In Digg, A thus becomes a "fan" of B, thereby receiving notifications about the activities of B, but this link and propagation of information remains unidirectional as B will not be informed about the actions of A.

This finding is consistent with sociological theory and ethnographic studies of social networks which identified that friendship requests in OSN are often driven by users' interest to become passively informed by means of these social ties [119, 124]. The fact that the average symmetry is significantly lower and also dependent on the degrees of remote nodes, underlines (a) that users are engaging in friendships in the Digg OSN with the intention of information delivery and (b) the existence of individuals which act (or views themselves) as sources and broadcasters of knowledge, which according to [118] would embody the critical influentials in the network.

SUBMITTING SUCCESSFUL STORIES

When looking at the entire body of stories submitted to the social news aggregator in 4 years, similar patterns of varying importance become visible.

[4] If user A names B a friend, B also refers to A as friend.

Table 3.7: Fraction of symmetric links in the Digg network. The likelihood to reciprocate incoming social ties and turn followers into bi-directional friends decreases with the total number of followers a particular user has.

Degree of node	Number of users	Symmetric link ratio
$0 < D_{in} < 10$	282536	0.53
$10 \leq D_{in} < 100$	49416	0.42
$100 \leq D_{in} < 1000$	13993	0.39
$D_{in} = 1000$	111	0.31

While a large number of people is watching the content published on Digg[5], only a limited number of registered users are actively submitting content to the social network.

The activity patterns of these users is furthermore biased, as shown in the Lorentz plot in Figure 3.21: the 80% least active users of the network are together submitting only about 20% of the entire content as shown by the dashed red line. This indicated a very uneven and biased system[6], nearly the same skew – commonly referred to as the 80-20 rule – has been found repeatedly in economics and sociology.

This skew becomes more drastic when only considering those stories that gained enough support and were promoted to popular items. As the figure shows, these successful stories can be attributed to a select minority of only 2% of the community, which is able to find and submit 98% of all stories that will go viral. This effect is however not the result of the pure quantity that users participate in the story submission process, in other words there exists no statistically significant relationship between the number of stories a person has submitted and the ratio of stories that will become popular (r^2=-0.01).

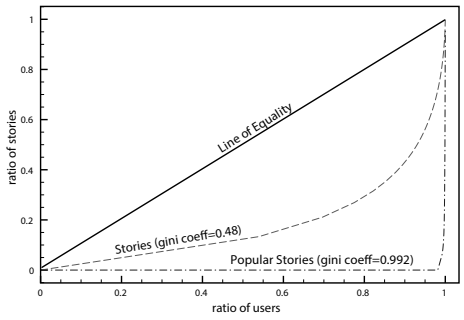


Figure 3.21: Equality of story submission. The figure shows a Lorentz curve of the total and popular story submission compared to the Digg user population.

While the presence of such a highly skewed distribution pointing out a few users might indicate the existence of a few "chosen ones", a closer inspection reveals that these highly successful submitters are not those users critical for the effective spread of

[5] A combined analysis of user comments, Diggs, and the number of visitors that followed a link associated with a particular story indicated that per registered digging user, the content is additionally seen by 12.9 passive spectators. The topics and generated clicks between spectators and digging users also reveal a near perfect overlap between digging and identified reading and usage patterns (r^2 = 0.96), thus the registered digging users may be viewed as a true proxy for the behavior of the entire Digg population. These two user groups and their clicking behavior account for more than 95% of the page hits referrals, Digg.com is generating in the Internet according to [125].

[6] An unbiased, equally balanced population is described by the "line of equality", where the top k% of users would contribute exactly k% of the content.

information. First of all, the average ratio of popular to submitted stories of the top 2% successful members of the community is only 0.23, therefore, even though they are the submitter of eventually highly popular content, they do not always generate top hits but a high proportion of their submitted content will not reach far. Second, the group of users who rank among the top successful members of the community is highly volatile. When comparing the top submitters between adjacent months or quarters, the set of successful users changes substantially between each studied time interval. As we do not find a significant number of stable members who are able to continuously repeat their previous successes, it has therefore to be concluded that there exists no conceptual difference or strategic advantage with those who do score successful stories. It appears that they were simply in the right place at the right time.

However, one may confidently say that it is not predominantly the well-connected nodes that are the originator of wide-spreading content, as there is no significant relationship between a user's success ratio and its level of connectivity with those around it ($p>0.5$).

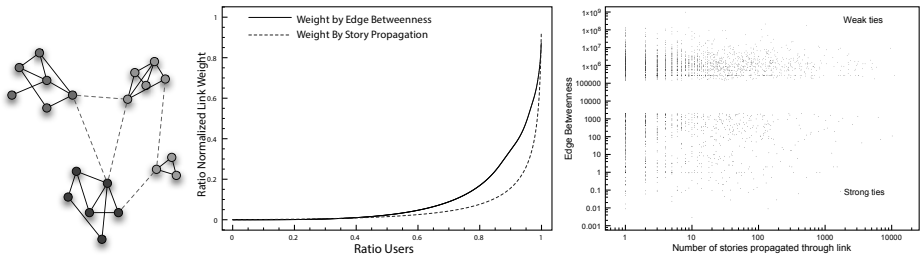
ACTIVATION OF THE SOCIAL NETWORK

While there do not exist any particular nodes that are over-proportionally injecting popular items into the network, there is the possibility that these nodes are highly successful in activating their surrounding friendship network, and therefore would be a key component in helping either their own or a friend's story reach widespread popularity.

It turns out however that the activation ratio of a node's direct friends is surprisingly low. On average, a particular node is only able to generate 0.0069 diggs per friendship link. This is mainly due to a combination of the already low conductivity of friendship links with low activity cycles of users. This low level of recruitment is furthermore quite stable with the structural properties of the network nodes. While the literature predicts that nodes in a social network achieve an exponentially increasing influence compared to their own importance [118, p. 124], a solid linear relationship (r^2=0.76) was found between the size of a nodes' friendship network and the amount of users a person can recruit to click on a story, and a low slope of the linear regression (a=0.102). In consequence, there is no over-proportional impact of higher-degree nodes: 1 activated user with 100 friends is on average about as effective as 10 activated users with 10 friends.

While no *quantitative* difference in the friendship network surrounding the important nodes was found, there may be a *qualitative* difference in terms of structural characteristics and the information propagation along links. As complex networks evolve, certain growth processes such as preferential attachment [126] create sets of highly connected clusters, which are interconnected by fewer links. According to social network theory [47, 127], these links among clusters, commonly referred to as "weak ties", act as a critical backbone for information propagation, as information within a cluster is communicated and replicated between nodes thereby creating high amounts of redundancy, while the weak ties transport other, previously unknown information between groups of nodes (see solid vs. dashed lines in Figure 3.22(a)).

To evaluate this hypothesis, the network was structurally classified into weak and strong ties according to their edge betweenness and compared with their theoretical importance to the actual amount of content that was propagated between each two nodes.



(a) (b) (c)

Figure 3.22: Information propagation along weak and strong ties.

Figure 3.22(b) shows a Lorentz plot of the link weight distributions for the topological betweenness and the actual information conductivity, demonstrating that the distributions are in general comparable and of the same class. As there is no hard threshold for what characterizes a weak or strong tie, we classified the top and bottom 20% of the distribution as weak and strong ties respectively and compared them to the number of stories propagated along a certain link. As shown in Figure 3.22(c), there does not exist any relationship ($r^2 = 0.00006$), thus information is not propagated more effectively along weak ties. Other topological definitions of how central a user is within a network, such as coreness or eigenvector centrality, also do not show any significant relationship to the propagation of information along edges ($r^2 = -0.0112$ and $r^2 = -0.0116$, respectively).

EARLY PREDICTORS

Finally, the question remains: if the assumed critical individuals – while not able to submit more popular content or activate more users – are able to early-on identify content that will later on become popular (see for example Keller and Berry [118]). By analyzing the voting patterns of all registered users on all stories from the months of April-May 2009, to determine how successful users were in finding and clicking on content that within the next hours or days would reach the popular stage, it was found that, users identified and reacted on average only to 11.9% (9% when eliminating those users who clicked on less than 5 stories in total over the period of two months) of content before it got promoted. With the absence of any high performers, it is impossible to identify specific individuals who are able to consistently and repeatedly find emergent trends.

This observation did not change either for the case of the high degree individuals or the users with a high success ratio of submitting content that will go viral; there exists no statistically significant difference in their ability to find content in the social network before it actually reaches widespread popularity.

3.4.3. BEYOND STATIC FRIENDSHIP RELATIONS

From the previous discussions it becomes evident that neither the importance of individual users nor the dynamics of the individual friendship relations or the network of

friends can at a statistically significant level consistently explain why a certain story will become a success while another one will not. Furthermore, as in nearly 50% of all stories the promotion process took place without any dominant interference by the friendship network, we further investigated how the low participation values of the friendship network may be explained and which features are the dividing force between those stories pushed by friends and those promoted by the general public.

SPREAD WITHOUT FRIENDS - A MATTER OF TIMELY RELEVANCE

To get to the root of why one story is propagated by the help of friends while another one is pushed by random users from the community, a survey was conducted in which a group of non-experts was presented with the title, description, image and the type of story (news article, video, or image) of the 158 most successful stories that were promoted in the last year. As in retrospect one may classify these stories as friend or non-friend promoted, the survey items were balanced in terms of topic areas to mimic a similar distribution as on the Digg.com website. Given only the contextual information about the story, the participants were asked to rate the general appeal, their own personal interest and the general importance of a particular story. Using a similar representation as on the Digg website, one story was presented at a time to the participants to rule out any influences from adjacent items or possible other cognitive biases such as the primacy effect [128].

The survey results indicated that the differentiation in the promotion process of stories was a direct result how important and relevant the participants rated the topic of a particular story. Either a high rating of "general interest to the public", in other words it is likely that one would hear about the topic in the evening news, or a high level of timely relevance, i.e. will this story be as important next month as it is now, was able to serve as a reliable predictor that a particular story has reached popularity on its own without driving help of friends (both factors statistically significant at p=0.05).

EXPLAINING CRITICAL MASS THROUGH TEMPORAL ALIGNMENT

As a large number of factors previously hypothesized to be of critical importance to information spread in OSNs turned out to be rather insignificant and furthermore highly volatile between observation periods, one further investigation was about the influence of time on the story propagation process. Some of the unexpected low or highly fluctuating factors are to some extent dependent upon the temporal alignment of users, i.e. whether users in general (and friends in particular) are visiting the site within the same narrow time window or not.

Figure 3.23 visualizes this idea of temporal alignment on a snapshot of the frontpages from April 2009, which shows the position of all popular stories with at least 100 diggs over a 48 hour time interval on the first 50 frontpages. As can be seen from the figure, there exists a high flux in the amount of stories passing through; within on average 3 hours the entire content on the first frontpage has been replaced by newer items. From a combined analysis of voting patterns and such frontpage traces, we are able to determine the usual search strategy and depth of users inside the social network, i.e., when, how often and how deep they are looking through the entire site. This process revealed that stories accumulate 80% of the entire attention they will receive after promotion from

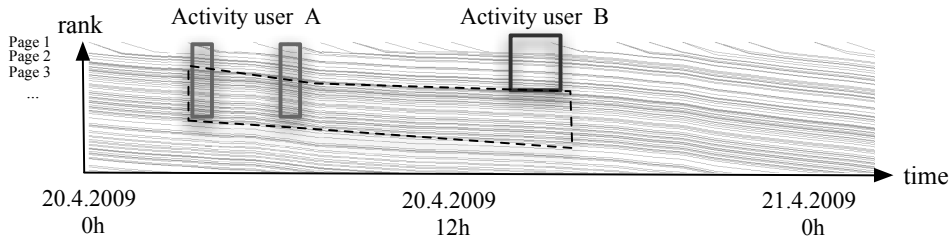


Figure 3.23: Story placement on front pages over time. The figure shows the development of the absolute position of stories on the front pages (y-axis) as stories age and are displaced by newly promoted material over time, based on a 48-hour snapshot in April 2009 (x-axis).

users on the first and second page only, while the ratio of users who are going over more than the first 4 front pages is practically zero.

Considering the case of two users active on 20 April 2009, this can explain the surprisingly low amount of common friendship activations, as nearly 70% of the stories visible to user *A* during the two morning visits are already outside of user *B*'s attention window as the user visits the social network just six hours later.

Unless *B* actively looks for and follows up on *A*'s activity, the abundance of content and high turnover rate of information combined with limited attention span will therefore largely bury the potential for commonality unless users proactively follow up through friendship relations. This finding demonstrates that whether a story reaches critical mass depends to a significant extent upon who and how many people are currently active on the site within a short time window. A combination of this temporal perspective with interest and friendship data can go a long way, as it was possible to improve the analysis accuracy of the activation ratio of certain friendship links and parts of the friendship network by a factor of 15. Note however that while a temporal view is currently able to reveal in retrospect why certain users clicked on a particular story and how and along which parts the information did propagate, it is not yet possible to predict how users will interact on a story in the future for a variety of reasons. Most importantly, an accurate prediction will require a good model of users' future activity periods at a fine enough resolution to minimize the prediction error of which stories users will see. Fur-

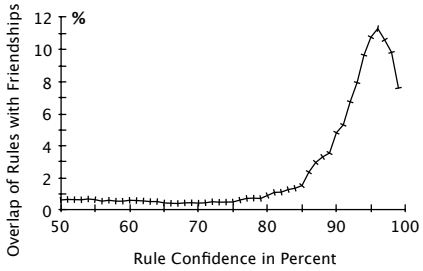


Figure 3.24: Overlap of friendship topology with behavioral rules. When comparing the overlap in behavior between pairs of users (x-axis) and their likelihood to have a friendship relation (y-axis), it is found that the vast majority of persons who commonly click together on stories are not related at a topological level. The probability for two users *A* and *B* showing nearly identical behavioral patterns (95% of *A*'s diggs are mirrored by *B*) to be friends is less than 12%.

thermore, it will necessary to further understand the concrete decision process that will lead to a user actively clicking on a story.

3.4.4. BEYOND BARE FRIENDSHIP TOPOLOGY

While critical mass can be significantly better explained when accounting for time differences and the shift and alignment of user activity periods, the individual friendship relations and the network of friends of friends still cannot fully describe the information propagation processes observed in the Digg social network. This section will present the case that a social network can only be partially captured through the topology of the direct *friendship network*, but that there may exist an unknown number of different *logical network layers* on top, whose topologies may reveal where and how interaction and collaboration actually take place.

ASSESSING THE IMPACT OF THE TOPOLOGICAL LAYER

In order to discover patterns of people and groups of people who commonly act together instead of only those who seem connected through a friendship relation, the corpus of diggs was analyzed for the existence of association rules, a machine learning technique which has previously provided merit in software debugging and marketing shown in Jeffrey *et al.* [129] as explained in chapter 2.1.2.

The minimum values for support and confidence in this analysis were 0.01% and 50%, meaning that any user considered for a particular rule must have participated in at least 0.01% of all stories (thereby eliminating abandoned and very low-volume user accounts) and establishing only a relationship if users share at least half of their diggs together. For the entire corpus, nearly 1.2 million common activity patterns could be discovered, which were mapped against the topology of the actual friendship network.

Figure 3.24 shows the percentage of friendship links between user pairs that were found to exhibit high levels of co-participation on the same stories as a function of the rule confidence in percent. As can be seen from the figure, the vast majority of similarly behaving user pairs in the Digg network have not formed a friendship between them. For any confidence value between 50-80%, meaning that in 5-8 out of 10 cases a digg by user A on a particular story will result in a digg from user B, there is less than a 1% probability that user A and B are directly connected. Even for extremely high performing rules, when in 95 out of 100 cases two users behave in an identical manner, less than 12% of those user pairs are friends. We can therefore conclude that although there exist some patterns in the common behavior of users, the bare topology of the friendship graph is unsuited to fully capture it.

THE "DIGG PATRIOTS": A HIDDEN LOGICAL LAYER

The existence of strong patterns and structure in the behavior of users on the Digg.com website suggest that users may engage in community building, forming a "logical network topology" characterized by specific semantics on top of the underlying social media platform. Concrete mechanisms to discover and identify the size and shape of these communities are still subject to research. As a proof-of-concept however, the case of the "Digg Patriots", an activist group of Digg users aiming to game the promotion algorithm through coordinated collective digging on stories, in some reported cases after payment by third parties [130], is possible.

When an email list archive of alleged members of the “Patriots" was exposed in 2010, it was possible to link the email communications of 102 members to a particular Digg profile and cross-reference their identities against the discovered highly-aligned activity patterns of users. Nearly half of the exposed “Patriots" also appear in the body of discovered association rules, Figure 3.25 shows the percentage of rules between 50-80% confidence from user interaction that were either linkable to either friendships or the “Digg Patriots". Remarkably, the collection of 102 coordinated “Patriots", known via the email archive, provides nearly a fifth of the discovered behavioral rules that can be extracted from the entire social network graph of 2 million users and 7.7 million friendship links between them.

It is therefore evident that effective social network analysis needs to go far beyond the analysis of the bare friendship topology and actually classify the semantics and characteristics of visible friendship and invisible other logical ties between social network users. How dramatic the logical ties between these particular users, undiscoverable from a graph theoretical perspective, might have been to the Digg social network can be exemplarily seen in Figure 3.26 which shows the diggs made over time on three promoted stories: Coordinating and orchestrating diggs in the early life time of a story, the “Patriots" (indicated in red) might have been the driver influencing the trajectory enough to push them over the promotion threshold, thereby leveraging the mass attention resulting from a front page through very little effort, yet invisible topologically.

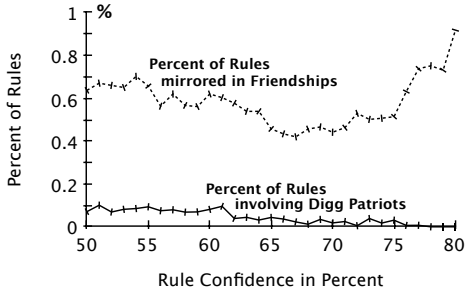


Figure 3.25: Explanatory power of the “Digg Patriots" layer.

The common assumption made in social network analysis that the deciding factor determining whether some information goes viral or not are the individual friendship relations among users was compared to empirical measurements. While evidence of some structure in how these friendship relations are formed was found (there is a high overlap of interests), the actual effectiveness and common clicking rate of friendship links is surprisingly low and does not confirm the high importance that is attributed to these social ties. As the wider *network* of friends stretching over multiple hops (friends tell their friends who tell their friends) provides a much smaller contribution in practice than it could in theory (it could reach an exponentially increasing number of entities), the impact and the propagation along friendships and the network of friends is in most cases not enough to reach critical mass. Furthermore one notices that although there exists a significant skew in the characteristics of

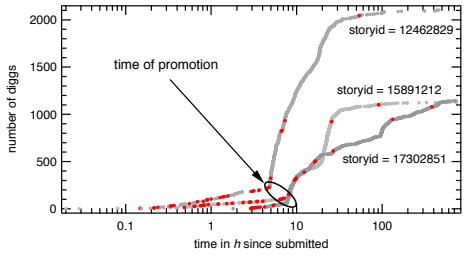


Figure 3.26: Concentrated activities of “Digg Patriots".

network nodes from a topological perspective, no evidence was found that these network nodes are indeed behaving differently and more effectively in terms of spreading information. They have no better access to information, are not more efficient in triggering their friends or do not predict trends better. The fact that there exists no group that can consistently and at a high level generate "hits" and individuals' success ratios fluctuate largely across observation periods leads to the conclusion that even successful members do not actually seem to have the recipe for success.

Various outcomes however, point to two factors that in the past have not received sufficient attention: time alignment and existence of non-topological relationships between users. When incorporating these factors, the conductivity of information propagation and the ability to explain it in retrospect improves manifold. Accurately predicting when users will be active and developing methods to detect and characterize these logical links between users should be the focus of future research.

3.5. CHAPTER SUMMARY

This chapter described the analysis of large scale friendship networks obtained from OSNs. In the first section (3.1), the question: How much data is needed to estimate topological metrics correctly, is answered by estimating how metrics evolve when using commonly used techniques: depth first search, breadth first search and random first search on a complete friendship graph of Digg.com. The results indicate that it is necessary to obtain more than 20-30% of the entire network to arrive at a solid estimate, for some metrics and traversal algorithm combinations nearly the entire network is necessary. In order to obtaining useful sub-graphs of an OSN, *Mutual Friend Crawling* is presented, an algorithm to crawl a large scale OSNs in such a way that community structure is detected and communities are crawled one after another. If only partial crawls of these social networks are used, this particular type of algorithm provides better topological samples than commonly used techniques like BFS or DFS.

However, when estimating community structure in OSNs the influence of overlapping communities did not receive significant attention in the past. That is why a hypergraph representation is introduced in Chapter 3.3, modeling the clique structure of affiliation networks.

When estimating or predicting the successfulness of users within OSNs, certain topological sociocentric measures, like betweeness centrality or eigenvector centrality exist. In Chapter 3.4 the usefulness of the friendship network was estimated, showing that these metrics do not reflect the ability of users to repeatedly succeed in distributing content. Additionally it is shown that small communities of users may dominate the propagation process by coordinating themselves though the usage of external communication. As these communities are not detectable through topological information, the important factor to identify such groups as well as successful spreaders is given through the timely alignment of friendships.

4

Evolution of Online Social Networks

Whenever a person registers a user account, befriends another one or follows messages of others, the network of an OSN is evolving. Interestingly, these purely structural changes are not purely random and follow certain physical and mathematical models.

Therefore this chapter gives insights into observable effects of evolving networks and the involved factors. Chapter 4.1 describes patterns of human behavior and claims that human activity might be log-normal distributed. As nowadays not only individuals but also robots are able to create links within a OSN, Chapter 4.2 provides insights into these "fake" relationships. Finally Chapter 4.3 gives a brief introduction into saturation effects in OSNs and a possible method to estimate if a network is saturated, by analyzing the degree distribution.

A model by Barabási and Albert [91] describes the structure of certain online social networks quite well. The model is based on the assumption of preferential attachment which denotes that a new node will connect to another already existing one, depending on the others' nodal degree. The probability to link to a node already having a high degree is higher than the probability to connect to a low degree node. Networks generated by the algorithm proposed by Barabási and Albert [91] exhibit the "scale-free" property which denotes that the degree distribution of graphs follow a power-law distribution. In terms of social networks, a user that joins an OSN is more likely to connect to another one who is already well connected (and therefore more visible) rather than connecting to relative unknown individuals, a behavior describing the process of preferential attachment quite well. Barabási and Albert stated that only the combination of growth and preferential attachment will lead to a power-law (scale-free) degree distribution having an power-law exponent γ of 3.

A power-law degree distribution can be found in the tail of multiple OSNs as shown in Figures 4.1, 4.2 a and b, as it appears as a straight line in a log-log plot. The degree distributions are measured in the following ways:

- **Twitter.com** Every user account on Twitter has a numeric ID in the range of 11 <

ID < 2147483648 (March 2014), whereas the ID increases for every new user. The ID-space is not completely used because of deleted accounts and Twitter's implementation. The profiles were sampled by choosing 100 random id's from blocks of 5000 id's. Therefore a random sample of all profiles was obtained. This sampling technique also explains the lack of extremely high degrees in the distribution, as the probability to sample these is marginally small.

- **Hyves.nl** The profile information on Hyves was obtained trough crawling the largest connected component of Hyves using RFS (explained in chapter 3.1.1 and A.1.5).
- **Deviantart.com** Profile information from Deviantart was obtained through different perspectives. On the one hand, if a user submits an image the image appears on a section of the website called "new deviations". Additionally the friends of a user can be crawled. An overlaying directed "network" of "watchers" was also crawled.

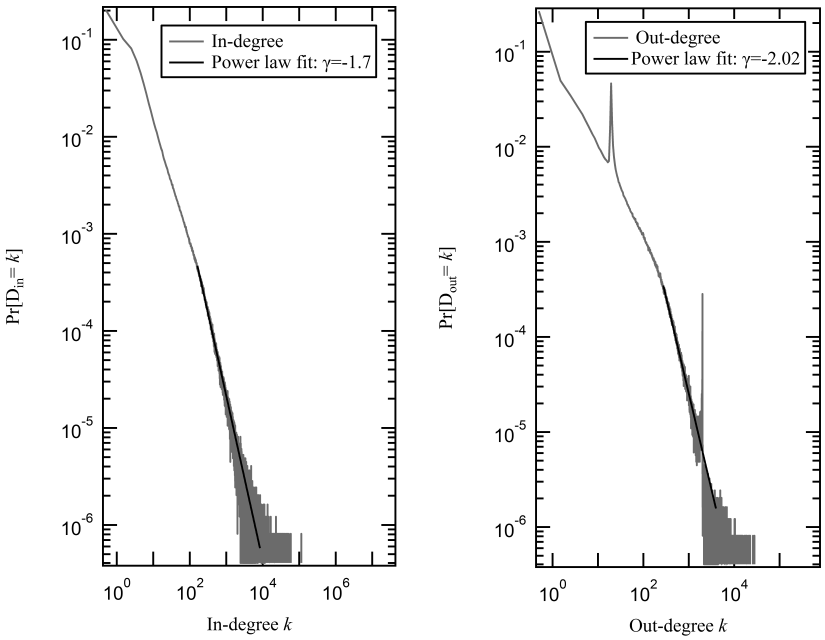
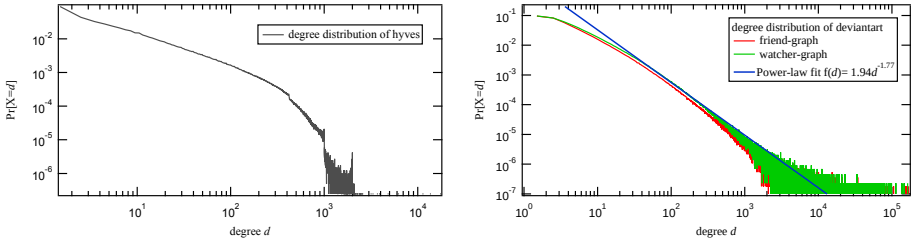


Figure 4.1: Degree distribution of in- and out-degrees in Twitter.

A power-law degree distribution suggests that most individuals have a very log degree of 1, denoting that these people have only one relationship. From the Figures 4.1 and 4.2 b, it seems that this finding is correct. But the distribution of the nodal degrees of Hyves.nl shows a clear deviation from a straight line in a log-log plot. Two reasons may account for the deviation. One lies in the used crawling technique which overestimates high degree nodes whereas on the other hand one may claim that in "real life" the num-



(a) Degree distribution of Hyves.nl.

(b) Degree distribution of Deviantart.com.

Figure 4.2: Degree distributions of Hyves.nl and Deviantart.com.

ber of individuals with no or only one relation towards friends, family or acquaintances is rather small.

The discrepancy between the expected degree distribution, using the model of Barabási and Albert [91] and some measurement data, can be analyzed by observing the quantitative and qualitative activities of individuals after having joined an OSN.

4.1. HUMAN INTERACTIVITY

Online Social Networks evolve if and when users are interacting with the services. Users may only create, edit or obtain information from a online service if they are using the Internet. Interaction between individuals through OSNs but also email or messaging services is therefore bound to the time span a person is using the. When thinking of content propagation as a form of viral spreading, the application of epidemic models becomes possible. In models like the susceptible infected susceptible (SIS) or the susceptible infected removed (SIR) model, the interactivity time, i.e. the duration between two spreading events of one user, is commonly assumed to follow an exponential distribution which allows the usage of Markov processes to model content propagation. However, multiple publications [108, 131–136] report that the duration between tasks like sending emails, accessing web pages, instant messaging and phone calls follow power-law distributions. A power-law random variable $X \geq \tau$ has the probability density function

$$f_X(t) = c\,t^{-\gamma} \qquad\qquad t \geq \tau \tag{4.1}$$

where $c = \frac{1-\gamma}{\tau^{1-\gamma}}$ and $\tau > 0$ is the lower bound for X. Because of these heavy tailed inter-activity distributions, word-of-mouth spreading, viral infections or dynamics of memes are expected to endure longer than previously expected based under the assumption of exponential inter-activity times. A virus or Internet meme would therefore survive longer and spread slower than expected, as described in [137, 138]. Heavy-tailed distributions may arise from a model based on a priority queue proposed by Barabási [131], where individuals have to finish tasks of which the majority of tasks can be completed in a short time, but some tasks take rather long. Barabasi's priority queue model fits quite well to the distribution of durations between events, leading to a power-law distribution with an exponent γ around 1. Unfortunately, these results complicate the applicability of

epidemic models because Markov Theory cannot be applied. Non-Markovian behavior is only recently addressed in the work of Cator *et al.* [139], Iribarren and Moro [137] and Van Mieghem and van de Bovenkamp [140].

However, when assuming that inter-event durations are power-law distributed, the following concerns arise:

1. In many cases, only a part of the data (the tail larger than τ) is modeled by a power-law. The lower bound τ in (4.1) usually does not correspond to the physical minimum of the random variable X, but τ is fitted from the data by ignoring the smaller values that do not obey the power-law. Often, these smaller values may have a large probability to occur so that their neglect is difficult to justify. In other words, only a part of the process (above τ) is then modeled by a power-law (4.1), while the other part (below τ) is not. In such cases, it is questionable whether the process or distribution of X indeed follows a power-law.

2. Apart from the lower bound τ, an upper bound K is frequently invoked, at which the power law is cut-off. Most processes or measurements have both a lower as well as an upper bound. However, it is often unclear whether the process in the deep tail still obeys a power law distribution or some other, much faster decreasing distribution. The upper bound K is usually empirically determined, rather than based on the physical maximum of X. As long as

$$\Pr[X > K] = c \int_K^\infty t^{-\gamma} dt = \left(\frac{K}{\tau}\right)^{1-\gamma}$$

   is small (with respect to the measurement precision), the upper bound K is justified, else other validation arguments are needed. The exponents γ of measured power-laws are mostly below 2, which many indicate that $\Pr[X > K]$ exceeds the measurement precision. In that case, the process or X may not be power-law distributed for large values exceeding K.

3. A crucial point shown in this chapter is that binning of data (either by the data-provider or the researcher, in a linear or exponential way) alters the shape of a log-normal distribution, so it may appear as a power-law.

It will be shown that one may claim that human interactivity (using technology) is better described by log-normal distributions than power-laws.

A log-normal random variable is defined as $X = e^Y$ where $Y = N(\mu, \sigma^2)$ is a Gaussian or normal random variable. Hence, $X \geq 0$. The probability density function (pdf) of a log-normal random variable X follows [23, p. 57] from the definition, for $t \geq 0$, as

$$f_X(t) = \frac{\exp\left[-\frac{(\log t - \mu)^2}{2\sigma^2}\right]}{\sigma t \sqrt{2\pi}} \tag{4.2}$$

where (μ, σ) are called the parameters of the log-normal pdf, while the mean and variance are [23, p. 57]

$$E[X] = e^{\mu} e^{\frac{\sigma^2}{2}}$$

and

$$\mathrm{Var}[X] = e^{2\mu} e^{\sigma^2} \left(e^{\sigma^2} - 1\right)$$

It will be shown, that a log-normal distribution might look like a power-law distribution, especially within the regime of small values for γ, while fitting the measurements equally well or better.

When working with empirical data one needs to be cautious as for example binning of data will alter the shape of a log-normal distribution, so it may appear as a power-law. When analyzing measurements from Digg.com [61, 77], and the Enron data set[1], a collection of emails sent by employees of the company Enron, one may argue that a log-normal distribution is a valid candidate for the distribution of human inter-event durations.

4.1.1. OBSERVATIONS AND MEASUREMENTS

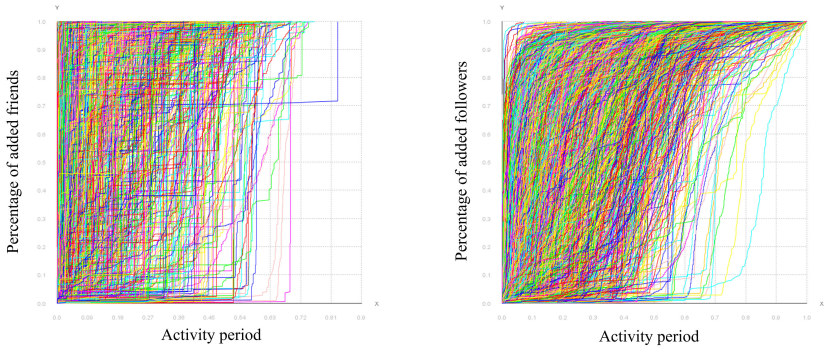
Whenever an online social network evolves, i.e. nodes and links are being created over time, all friends of a user must be added during a period that the user was logged-in to the system. Users need to add friends or send friend requests through the interface of the OSN themselves. Having access to a complete data set including all activities of users from Digg.com for a time of 4 years, described by Tang *et al.* [77] and Doerr *et al.*[61], allows to analyze the timeframe in which users of Digg.com add their friends. Figure 4.3(a) depicts traces of 1000 randomly picked users of Digg.com, visualizing at which time, friendships were created.

Staircase-like patterns shown in Figure 4.3(a) appear, because people are adding multiple friends in rather short time intervals. These bursts of activity also observed in email communication [131, 137, 141, 142] seem to be omni-present in human activities. For simplicity, the total duration users were active in Digg (abscissa) and the total number of friends (ordinate) is normalized.

As the network of Digg.com is directed (like in Twitter.com or other OSNs), one may only follow other users, while adding followers is not possible. This means the process of adding friends is solely based on the user himself, whereas obtaining followers depends on the activities of other users. This difference explains why the number of followers as shown in Figure 4.3(b) increases more smoothly over time compared to the trajectories in Figure 4.3(a).

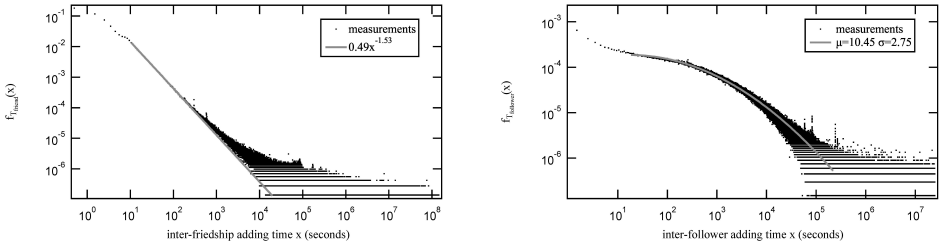
Figure 4.4(a) and Figure 4.4(b) depict how the network of Digg.com "grew", by visualizing the distribution of durations between adding friends, T_{friend} and receiving followers, $T_{follower}$, for the 7.4 million friendship relations in Digg. The histogram of T_{friend} can be fitted by a power-law with an exponent $\gamma = 1.5$, whereas $T_{follower}$ is fitted best by a log-normal (4.2) with parameters $\mu = 10.5$ and $\sigma = 2.8$. These parameters resemble the findings of Doerr *et al.* [143], who showed empirically that reaction times in a retweet network from Twitter and Digg are close to a log-normal distribution with similar values ($\mu = 10.1$ and $\sigma = 2.2$). The parameters are similar because Doerr *et al.* [143] analyzed the duration between consecutive logins to Digg.com, where being logged into the social service is a necessary condition for the analysis of friendship creation times as well, as a user can only create relations during the time he is online. The reason that the

[1] Enron Email Data set, Leslie Kaelbling and Melinda Gervasio, http://www.cs.cmu.edu/~enron/



(a) (b)

Figure 4.3: Trajectories of a sample of 1000 users of Digg.com, depicting the amount of friend- (a) and follower-relations (b), created during the time an account existed. Colors indicate different users.



(a) adding friends in seconds. (b) receiving followers in seconds.

Figure 4.4: Distributions of time differences between adding and receiving followers in seconds.

distribution of $T_{follower}$ does not fit the log-normal distribution along the whole range lies in the nature of Digg.com. Tang *et al.* [77] showed that only a few users were active over a long period, and just a fraction of them was actually submitting stories. This implies that users, whose submissions were promoted to the frontpage of Digg.com, have a higher visibility. Actually, as stated in Doerr *et al.* [61], only 2% of all registered users were actually successful in having their submissions "promoted" to the frontpage. Since the username of a submitter is written next to the story, these users received a lot of followers during the period their story has been on the main page. For this reason, $f_{T_{follower}}(x)$ is large for small x in Figure 4.4(b).

4.1.2. FITTING A LOG-NORMAL DISTRIBUTION

Fitting a log-normal random variable needs careful crafting and different techniques. The two main approaches are based on fitting the PDF and the EDF (empirical distribu-

tion function[2]) of the log transformed data. Figure 4.5 depicts the data fitted[3] to the CDF of a normal distribution $F_X(x) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$. The legend in Figure 4.5 illustrates that the parameters of the CDF are in the same range as the estimated ones of the fitted PDF so that the conclusion can be drawn that the assumption of fitting a log-normal is valid.

Binning the data, say per minute or per hour, which means scaling the distribution by a factor of 60 (60 seconds = 1 minute), will shift the parameter μ of the CDF towards the left by a factor of $\ln(60) \approx 4.09$. However, a remarkable property of the log-normal distribution is that the parameter σ will not change after linear scaling as shown in the Appendix 4.1.3.

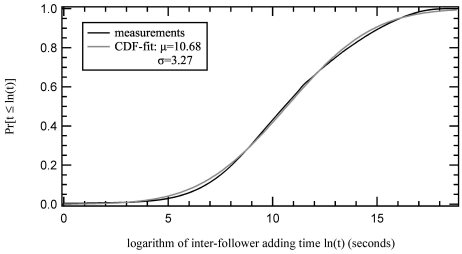


Figure 4.5: EDF of the inter-follower adding durations.

An important consequence of a binning operation is illustrated in PDFs in Figure 4.6 and Figure 4.7(a) and 4.7(b). These figures show the histogram of the durations to add followers $T_{follower}$ binned by minutes and hours, respectively. The parameter μ decreases, as predicted, by a factor of ≈ 4.1 and the parameter σ keeps its value, but the shape of the distribution changes. The distribution, binned per hour, is shown in Figure 4.7(a) where the distribution can be interpreted as a power-law distribution (with exponential cut-off). Moreover, this misinterpretation may be justified, because the exponent $\gamma = 1.2$ of a fitted power-law is close to the exponent $\gamma = 1.49$ found for the T_{friend} distribution.

In Figures 4.6,4.7(a) and 4.7(b), the data was artificially binned by hours, but the real problem occurs if the data provider (the OSN or web service) returns values per hour or even at larger scales. Then, the binning happens at the operator's side and it becomes impossible to distinguish between a power-law (with exponential cut-off) and a log-normal distribution. As often used to fit power-laws, in Figure 4.7(b) the same data is shown after using exponentially-sized bins and fitted to a power-law with the exponent being 1.81. But still, the visualization is misleading as the original distribution is, as shown earlier, a log-normal distribution.

The Enron data set shows this effect very nicely. The whole data set of emails covers the communication of all employees of the company for a duration of roughly 6 years, starting in January 1998 until February 2004. In the first few years, until May 2001, the time an email was sent is captured per minute, whereas afterwards (June 2001 - February 2004) the time is stored with an accuracy of a second. The resulting histogram of time differences between sent emails is drawn in Figure 4.8(a), where the black dots represent the measurements per second and the red ones the measurements per minute. The parameters of the fitted log-normal and power-law distribution lie in a similar region as the ones shown earlier for the durations for friends to be added to a user's account.

[2] The empirical distribution function is sometimes also called empirical cdf, as it is the cumulative distribution function based on an empirical measure.

[3] Fitting data to a EDF usually flattens interesting parts of a PDF, especially the tail of a distribution. Still, the benefit lies in the fact that binning is not needed and "raw" data can be fitted [144].

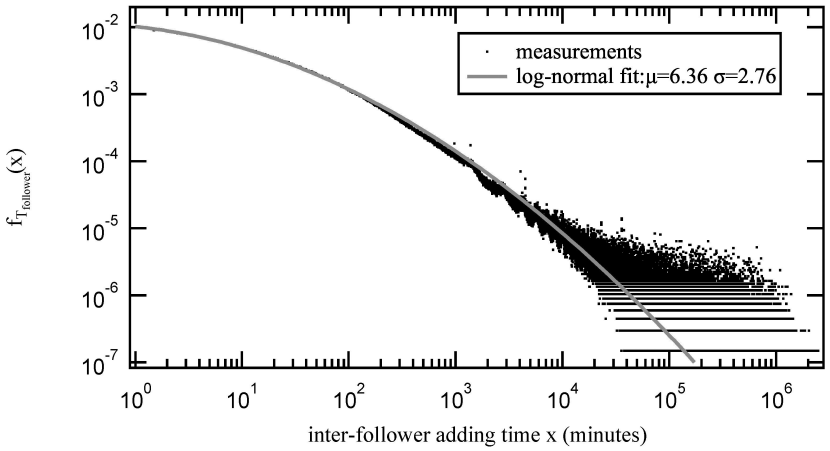


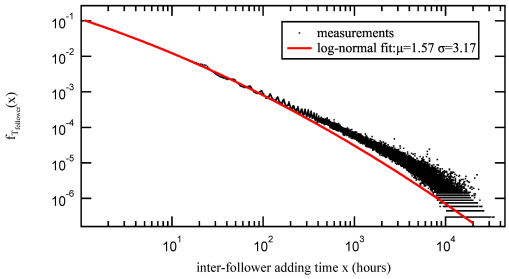
Figure 4.6: Probability density function (pdf) of the inter follower times binned per minute.

If one ignores the time difference between sending and receiving an email, one may analyze the time differences between receiving certain emails. This measurement, illustrated again by the histogram of inter-arrival times split in observations per minute (red) and per second (black) in Figure 4.8(b), can be compared to the durations of added followers, whereas the parameters are again in the same range as shown before. The EDF of the data shows that the assumption of a log-normal distribution is a valid one, as the fitted parameters match the ones of the PDF quite well.

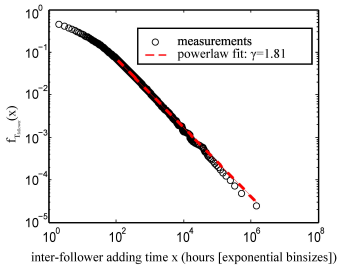
When artificially drawing numbers from a log-normal distributed random variable, exactly the same effect is observed as depicted in Figure 4.10. One million random numbers from a log-normal distributed random with parameters $\mu = 10$ and $\sigma = 2$ were generated as ground-truth. By binning (scaling) the distribution with larger binsizes, the upgoing regime disappears naturally and the observable part of the distribution "evolves" into a straight line on a log-log plot.

Exponents between 1 and 2 of power-laws are found in most data sets of human activity, whereas γ lies between 0.7 and 1.5. Table 4.1 lists reported interarrival times and the fitted exponents.

Certain methods of estimating and fitting parameters of power-law distributions are proposed. Clauset et al.[145] approached the problem of estimating the exponent and by testing different distributions. In their data, log-normals and power-laws were not clearly distinguishable either and a log-normal distribution actually achieved a higher p-values (goodness of fit) than power-laws for some tested data sets, but log ratio tests suggested that some of the tested distributions are closer to power-laws. By using the technique in [145] to fit the distribution of T_{friend}, a power-law having an exponent of $\gamma = 1.77$ with a reasonable p-value of 0.23 can be found. The distribution of $T_{follower}$ is most likely not a power-law because the $p = 0.0$. Table 4.2 lists the parameters of these two fits. As already mentioned, the original data (binned per second) of $T_{follower}$ is very likely not to be a power-law, whereas the data binned per hour seems to be quite a good

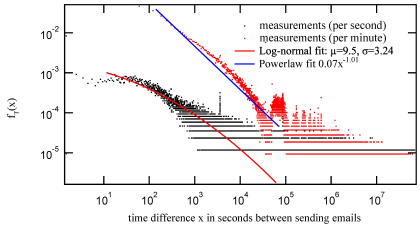


(a) Probability density function (pdf) of the inter follower times binned per hour with equally sized bins.

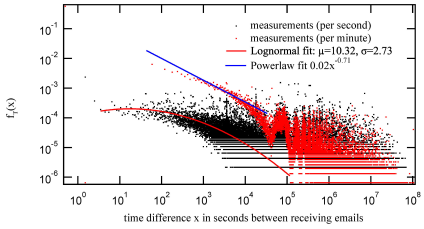


(b) Probability density function (pdf) of the inter follower times binned per hour with exponential sized bins.

Figure 4.7: Trajectories of a sample of 1000 users depicting the amount of friend- (a) and follower-relations (b), created during the time an account existed. Colors indicate different users.



(a)



(b)

Figure 4.8: Time difference between sending (a) and receiving (b) emails per person in the Enron data set.

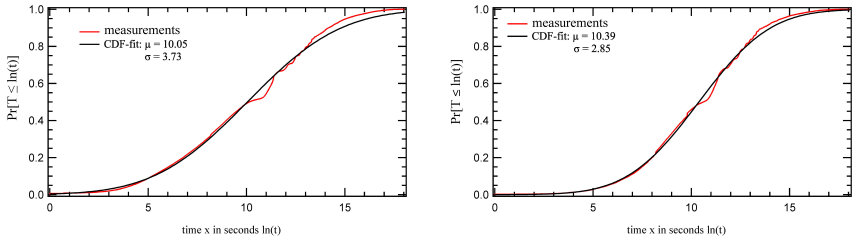
power-law with an exponent of -1.8.

One may observe in Table 4.2 that the larger the bin-size, exemplary for $T_{follower}$, the higher the p-value of the fitted power-law which depicts that the larger the bin-size the higher the probability a power-law is favored over a log-normal distribution.

The general question whether an observed straight line in a log-log plot is a power-law or a (truncated) log-normal, arises from the fact that both distributions are indistinguishable given certain parameters. Taking the logarithm of both sides of a log-normal density (4.2) results, as shown in the Appendix (in 4.15), in

$$\ln(f_X(t)) = -\frac{1}{2\sigma^2}\ln(t)^2 + (\frac{\mu}{\sigma^2} - 1)\ln(t) - \ln(\sqrt{2\pi}\sigma) - \frac{\mu^2}{2\sigma^2}$$

If $\sigma^2 \gg 1$, then the second term $(\frac{\mu}{\sigma^2} - 1)ln(t)$ dominates, which leads to $ln(f_X(t)) \approx -ln(t) + c$, a straight line in a log-log plot resembling a power-law with exponent $\gamma = 1$. On the other hand, if σ is rather small as in the measurements where $2 < \sigma < 3.3$, or if $\sigma^2 = \mu$, the first term $-\frac{1}{2\sigma^2}ln(t)^2$ dominates and a quadratic function appears in a log-log plot. Obviously a transition exists in which the second term "takes over", which denotes a



(a) (b)

Figure 4.9: EDF of the time difference between sending (a) and receiving (b) emails per person in seconds.

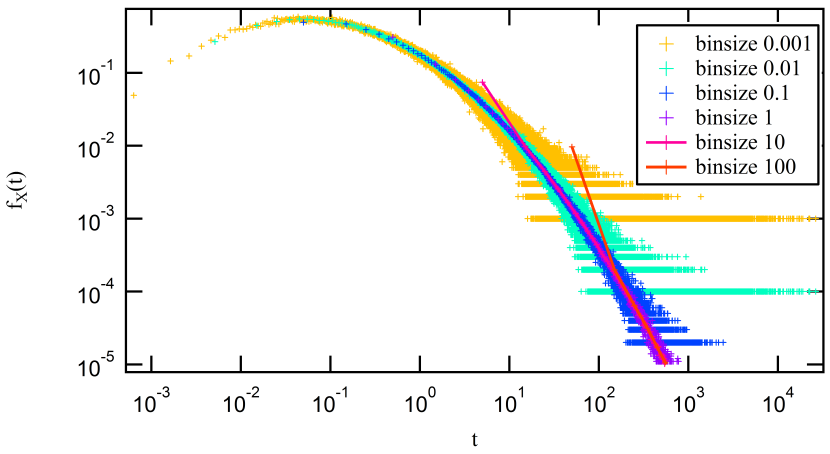


Figure 4.10: Effect of different binning on generated log-normal noise.

"flattening" of the tail of the observed distribution which, when fitted with a power-law, might show exponents larger than 1, but still reasonably large values for σ are necessary which are out of the range of the measured $2 < \sigma < 3.3$. As mentioned earlier, binning the data with larger binsize will decrease the parameter μ, even to the extent that μ may become negative. But, as the mode (the value at which the distribution has its maximum) of a log-normal distribution equals $e^{\mu-\sigma^2}$, when $\mu < \sigma^2$ the maximum of the log-normal tends to become zero (infinitely far to the left on a log-log scale). Therefore, just the decreasing part of the quadratic shape will be visible in a log-log plot. Additionally, increasing the binsize denotes summing all smaller values of the distribution which leads to an effect, nicely observable in Figure 4.10 by the points connected by violet lines (binsize 10), which nearly resemble a straight line having the first plotted point at position 5. The sum of smaller values in the first bin leads to an interesting effect, namely that the first point lies nearly exactly on an estimated straight line. Binning therefore obscures the quadratic shape the log-normal follows in a log-log plot. In the extreme case, plotted

Data set	Exponent (γ)	Reference
Time between sending Emails	1	Barabasi (2005) [131]
Time between sending Emails	1.2	Eckmann *et al.* (2004) [132]
Time between clicks in website usage	1, 1.25	Gonçalves and Ramasco (2008) [133]
Time between similar actions in web data	1.1, 1.2, 1.8	Radicchi (2008) [134]
Time between messages in instant-messaging	1.53	Leskovec and Horvitz (2008) [108]
Time between phone calls	0.9	Candia *et al.* (2008) [135]
Time between phone calls	0.7	Karsai *et al.* (2012) [136]
Time between sending Emails	1	Karsai *et al.* (2012) [136]
Time between sending short messages	0.7	Karsai *et al.* (2012) [136]

Table 4.1: Power-law exponents found for durations between technology related human dynamics.

Distribution	Exponent (γ)	p	xmin	gof
Adding friends	1.77	0.23	59	0.0142
Adding follower (binned per s)	2.03	0.0	12	0.0123
Adding follower (binned per m)	1.88	0.0	31	0.0189
Adding follower (binned per h)	1.81	0.37	116	0.0117
Adding follower (binned per 2h)	1.81	0.59	167	0.0135
Adding follower (binned per 6h)	1.81	0.96	192	0.0149
Sending emails (binned per s)	2.33	0.0	3	0.21
Sending emails (binned per m)	2.03	0.21	1	0.18
Receiving emails (binned per s)	2.01	0.0	1	0.019
Receiving emails (binned per m)	2.21	0.05	182	0.0221

Table 4.2: Estimated parameters using the method of Clauset et al. [145].

in Figure 4.10 in red (binsize 100) the first point may even lie above this imaginary line. When taking into account that a power-law is typically only fitted to the tail of a distribution means ignoring these first points which may lead again to the observation of a straight line in a log-log plot, starting at a certain t_{min} value.

THE HISTORICAL DEBATE OF POWER-LAW VERSUS LOG-NORMAL DISTRIBUTIONS

As mentioned earlier, power-law distributions with exponents smaller than two are often observed in interactivity durations of human behavior, as listed in Table 4.1, where the findings date back to 2004. However, log-normal distributions were measured for similar tasks as shown in table 4.3, which extends the collection of Limpert et al. [146]. The oldest analysis was conducted by Boag [147] in 1949. Based on the scaling invariance of σ as mentioned earlier and demonstrated in Appendix (4.1.3), the parameter σ can be compared over different measurements, whereas the parameter μ not, since it depends on the units in which the log-normal random variable is measured. Interestingly, all σ's in table 4.3 are within an small range of $0.35 \leq \sigma \leq 3.2$, which shows that the parameter σ only varies over one order of magnitude in different measured phenomena. Consequently, the rather small values of σ in Table 4.3 and $2.73 \leq \sigma \leq 3.24$ found in the previous sections contradict the common deductions made from (4.15), namely that only for large values of σ power-law and log-normal distributions are indistinguishable.

Table 4.3: Literature of log-normal distributions (excerpt).

mu	sigma	Process	Reference
5.547	2.126	Email forwarding	Iribarren and Moro[137]
≈ 8	≈ 2	Email forwarding	Stouffer *et al.* [141]
$\mu_1 = 1$ hour $\mu_2 = 2$ days		Email forwarding	Stouffer *et al.* [148]
2.47	0.38	Infection times	Nishiura [149]
14 days	1.14	Latency periods of diseases	Sartwell [150]
100 days	1.24	Latency periods of diseases	Sartwell [150]
2.3 hours	1.48	Latency periods of diseases	Sartwell [150]
2.4 days	1.47	Latency periods of diseases	Sartwell [150]
12.6 days	1.50	Latency periods of diseases	Sartwell [150]
21.4 days	2.11	Latency periods of diseases	Sartwell [150]
9.6 months	2.5	Survival times after cancer diagnosis	Boag [147]
15.9 monts	2.8	Survival times after cancer diagnosis	Feinleib and Macmahon [151]
17.2 months	3.21	Survival times after cancer diagnosis	Feinleib and Macmahon [151]
14.5 months	3.02	Survival times after cancer diagnosis	Boag [147]
60 years	1.16	Age of onset of Alzheimer	Horner [152]
4 days		Incubation periods (viral infections)	Lessler et al. [153]
3 to 5	≈ 2	Task completion	Linden [154]
0.5	1.4	Strike duration	Lawrence [155]
		Time of individual activities	Mohana et al. [156]
0.43	1.634	Call duration	Spedalieri et al. [157]
3.5	0.70	Message holding time	Barcelo and Jordán [158]
7.439	0.846	Transmission holding time	Barcelo and Jordán [158]
3.29	0.890	Channel holding time	Barcelo and Jordán [158]
3.3	0.89	Channel holding time	Barcelo and Jordán [158]
		Citations	Eom and Fortunato [159]
		Citations	Redner [160]
1 to 2	0.35 – 0.45	Citations	Stringer et al.[161]
	1.095	Citations	Radicchi et al.[162]
$\mu_1 = 3.7$ $\mu_2 = 5.6$	$\sigma_1 = 0.8$ $\sigma_2 = 3.1$	Retweeting behavior	Doerr et al.[143]
5.29	0.42	Distribution of votes on pages of Digg.com	Van Mieghem et al. [163]

Table 4.3 shows that the research on log-normal distributions has a long history and is often related to counting occurrences after an certain amount of time units. Still, the generating processes of most log-normals is not well understood. Mitzenmacher [164] gives an overview of processes leading to power-law and log-normal distributions, emphasizing that minor changes in the process will lead to either the one or the other distribution. He also mentions the work of Gabaix [165], who analyzed the size distribution of cities in the United States. Interestingly, Gabaix found that the city size distribution follows Zipf's distribution, which is similar to a power-law with an exponent of $\gamma = 1$. Gabaix argues that cities cannot become infinitely small, a fact that imposes a lower bound to the process. When modeling the size of cities as a Markov chain with a fixed number of cities, which grow stochastically as proposed by Gibrat [166], the steady-state of the Markov chain will follow Zipf's distribution with an exponent of $\gamma = 1$. If there would be no lower bound on the city size, then the distribution would become degenerated, leading to a log-normal distribution where most cities would have a infinitesimally small size. Gibrat [166], whose work is often associated with the law of proportionate

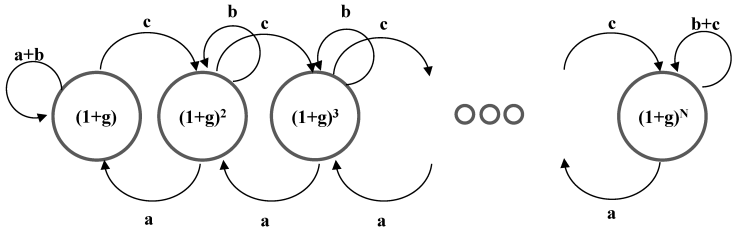


Figure 4.11: Discrete-time Markov chain representing the growth process as describes by Cordoba [168].

effect [163] or with Gibrat's law similar to preferential attachment, argues that the state of a random variable does not influence its transition probabilities or the variance of transition probabilities. Champernowne [167] and Cordoba [168] analyzed the income distribution of England and Wales and used a Markov chain to model a process that explains the distribution. Again, their crucial assumption is that incomes have a lower bound.

To clarify the mentioned models, Figure 4.11 depicts the discrete Markov chain used by Cordoba [168] describing a general random walk. The states represent different classes of exponentially increasing size as $x_i \equiv (1+g)^i$, where $1 \leq i \leq N$ and $g > 0$. The transition probabilities denote the probability that the size x_i represented by state i either stays the same (transition b), increases (transition c) or decreases (transition a). The steady state of this Markov chain is Pareto distributed with an exponent of $\gamma = 1$ [168].

The transition probabilities can be written as a matrix (see e.g. [23]):

$$\pi = \begin{bmatrix} a_0+b_0 & c_0 & 0 & 0 & \cdots & 0 & 0 \\ a_1 & b_1 & c_1 & 0 & \cdots & 0 & 0 \\ 0 & a_2 & b_2 & c_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & a_{N-1} & b_{N-1} & c_{N-1} \\ 0 & 0 & 0 & 0 & 0 & a_N & b_N+c_N \end{bmatrix} \tag{4.3}$$

Given $a_k = a$ and $c_k = c$ and $a > c$, one may define $\rho = \frac{c}{a}$. As shown by [167], the steady state of the discrete time Markov chain is then $\pi_j = M\rho^i = Mx_i^{-\delta}$, where M is a constant, $\delta = \frac{ln(a/c)}{ln(1+g)}$ and $\Pr[X_t \geq x_i] = \frac{1}{1-(c/a)} Mx_i^{-\delta}$ if $N \to \infty$. Simon [169] showed that Gibrat's law of the proportionate effect may lead to other heavy-tailed distributions as well.

Eeckhout [170] analyzed the distribution of city sizes by using accurate data from the US census in 2000 and found that the upper tail obeys Zipf's law. Nonetheless, the entire distribution is better described by a log-normal than a Pareto distribution. By comparing the census data from 1990 and 2000, Eeckhout shows that the growth of a city is independent of its size. The parameters of the log-normal distribution found by Eeckhout were $\mu = 7.28, \sigma = 1.75$.

The difference with the above mentioned Markov chain approach, lies in the fact that

Eeckhout modeled the process by a multiplicative process, which leads to a log-normal distribution. This multiplicative process, modeled by Kapteyn [171] in 1903 for the first time and later renown as the "Law of Proportionate Effect" by Gibrat [166], is based on the Central Limit Theorem applied to a multiplicative process, which therefore leads to log-normal distributed sizes [23]. Gibrat's growth process is defined as

$$S_{t,i} = a_{t,i} \times S_{t-1,i} \tag{4.4}$$

denoting that the size S of an element of interest i at state t is depending on the previous size S_{t-1} times a positive factor $a_{t,i}$, which is randomly distributed. By taking the logarithm of both sides in (4.4) and denoting $\xi_{t,i} \equiv \ln a_{t,i}$, one obtains

$$\ln S_{t,i} = \ln S_{t-1,i} + \xi_{t,i} = \ln S_{0,i} + \xi_{1,i} + \xi_{2,i} + \ldots + \xi_{t,i} \tag{4.5}$$

4

By the Central Limit Theorem [23], $\frac{\sum_{k=1}^{t} \xi_{k,i} - t\mu}{\sigma\sqrt{t}} \xrightarrow{d} N(0,1)$, one arrives at, for large t,

$$\Pr[\ln S_{t,i} \leq y] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{y-t\mu}{\sigma\sqrt{t}}} e^{-u^2/2}\, du$$

from which approximately

$$\ln S_{t,i} \cong t\mu + \sqrt{t}\sigma\xi \tag{4.6}$$

where μ denotes the mean and σ of the sequence $\{\xi_{k,i}\}_{1 \leq k \leq t}$.

The mentioned results about income and city size distributions are similar to the question asked in this thesis: Are the interactivity times log-normal or power-law distributed. The historical account demonstrates that the question is difficult to answer, because different ways of measuring the size distribution are possible and observations may be misleading since the tail of a Pareto and log-normal distribution may follow the same shape. Therefore, the important parts of the distributions are the upper tails. Malevergne [172] showed that, if $x \to \infty$, the tail of a log-normal is "shorter" than the tail of a Pareto distribution and will tend to zero faster than a power-law distribution. His reasoning is based on the fact that the Pareto distribution is slowly varying for $t > 0$

$$\lim_{x \to \infty} \frac{f(t \cdot x)}{f(x)} = t^a \tag{4.7}$$

The log-normal distribution on the other hand is not slowly varying. In the limit $x \to \infty$ a log-normal distribution will, for $t > 1$, always tend to zero:

$$\lim_{x \to \infty} \frac{f(t \cdot x)}{f(x)} = \lim_{x \to \infty} \frac{1}{t} e^{-\frac{(ln(t))^2}{2\sigma^2}} e^{-ln(t) \cdot \frac{ln(x)-\mu}{\sigma^2}} = 0 \tag{4.8}$$

In other words, the tail of a log-normal goes faster to zero than a Pareto distribution. This observation questions whether exponential bin sizes should be used in a log-log plot as these may hide the rapid decrease in the tail. On the other hand, it is questionable if data for high values is available in order to observe the tail behavior. Especially as log-normal and power-law distributions are not distinguishable for possibly long ranges of their tails. Malvergne [172] describes that the regime where the two distributions are

similar to each other by rewriting the the log-normal distribution as in (4.13) and (4.14). If the exponent (4.14) decreases slowly, the distribution may look constant over long ranges in t. Malvergne et al. [172] describe that for a parameter of $\sigma = 3.4$, the exponent varies no more than 0.3 units over a range of three orders of magnitude. As explained in the Appendix (4.13) details in which t-range the log-normal and power-law distributions are similar.

In experiments it may be difficult to measure high t values in order to observe this effect, especially if some other noise is affecting the data. Malevergne showed, for example, that the upper tail of the city size distribution is Pareto-distributed but for smaller cities the log-normal distribution gives a better model. In the presented measurements, the interactivity times are well defined and measurable but still, the upper tail of all the data shows large fluctuations so that the "drop" to zero is not visible.

Adding friends in online social media can only occur during the time a user is interacting with a service, which leads to a high number of short durations denoting the time between two consecutive actions. The distribution of durations to add friends follows a power-law with an exponent of $\gamma = 1.5$, whereas the durations to acquire followers is well described by a log-normal with $\mu \approx 10.5$ and $\sigma \approx 2.8$. Due to the impossibility of executing two tasks at the same time, one may claim that a log-normal distribution is a valid candidate, because it always has an up-going regime from zero on, but features a tail that follows a power-law distribution over multiple orders of magnitude.

Further on, binning of log-normal distributed data affects the perception of distribution functions, because the parameter μ shifts towards smaller values, but the parameter σ of a log-normal distribution does not change through a binning or scaling operation. In the extreme case, one may only observe the tail, which follows a power-law distribution with an exponent of $\gamma = 1$. Additionally, the fact occurring in any empirical measurement that events with very low probability are possibly not observed leads to the appearance of the tail of a log-normal as "nearly" a straight line in a log-log plot with an exponent larger than 1.

Another described concern is that a power-law with an exponent of smaller than 2 has no finite mean. Such a small exponent would indicate that with a small but non zero probability there is a chance that the time between adding friends will be larger than the lifetime of an individual. It is shown, however, that a log-normal distribution will converge to 0 faster than a power-law distribution so that extreme inter-activity durations possess a much smaller probability.

These observations and concerns, discussed for a long time in the literature of city size and income distributions, give reasons to argue that a log-normal distribution is fitting the whole data range better than power-laws. As reviewed in Section 4.1.2, similar probabilistic models produce either power-law distributions or, with minor changes, log-normals. These minor changes, defined through allowing different transition probabilities in a used Markov model may lead to a log-normal distribution whereas assuming all transition probabilities to be the same leads to a power-law distribution.

One may conclude: A power-law in empirical data with exponent smaller than 2 can actually be a log-normal.

4.1.3. THE LOG-NORMAL RANDOM VARIABLE AND DISTRIBUTION

As shown earlier in 4.2 a log-normal random variable is defined by its probability density function (pdf) of a log-normal random variable X as

$$f_X(t) = \frac{\exp\left[-\frac{(\log t-\mu)^2}{2\sigma^2}\right]}{\sigma t\sqrt{2\pi}} \tag{4.9}$$

with the parameters μ and σ.

The limit $\sigma \to 0$ reduces to a Dirac delta function at $t = e^{\mu}$, thus $\lim_{\sigma\to 0} f_X(t) = \delta(t - e^{\mu})$.

Given the mean and variance, the parameters of the log-normal are found as

$$\sigma^2 = \log\left(1 + \frac{\mathrm{Var}[X]}{(E[X])^2}\right) \tag{4.10}$$

and

$$\mu = \log E[X] - \frac{\sigma^2}{2} \tag{4.11}$$

Although $E[X] \geq 0$, one should remark that the parameter μ can be negative. Moreover, (4.10) and (4.11) shows that the scaled log-normal random variable $Y = bX$, where b is a positive real number, has mean $\sigma_Y = \sigma$ and $\mu_Y = \mu + \log b$. Hence, *scaling* by a factor b does not change the parameter σ, which has interesting consequences for binning and measured data: the unit (e.g. second versus hours) in which the random variable is measured does not alter the parameter σ, only the parameter μ.

The change of the argument $t \to e^u$ in $f_X(t)$ leads to

$$f_X\left(e^u\right) = e^{-\mu+\frac{\sigma^2}{2}} \frac{\exp\left[-\frac{(u-(\mu-\sigma^2))^2}{2\sigma^2}\right]}{\sigma\sqrt{2\pi}} \tag{4.12}$$

illustrating that the scaled log-normal pdf $e^{\mu-\frac{\sigma^2}{2}} f_X(e^u)$ is a Gaussian pdf $N\left(\mu', \sigma^2\right)$ with mean $\mu' = \mu - \sigma^2$. The maximum of $f_X(t)$ occurs at $t = e^{\mu-\sigma^2}$ and equals $\max_{t\geq 0} f_X(t) = \frac{e^{-\mu}e^{\frac{\sigma^2}{2}}}{\sigma\sqrt{2\pi}}$, which follows directly from (4.12). Moreover, it is easier to find from (4.12) than from (4.2) that $\lim_{u\to-\infty} f_X(e^u) = f_X(0) = 0$ and that $f_X'(0) = 0$. This means that any log-normal starts at $t = 0$ from zero, increases up to the maximum at $t = e^{\mu-\sigma^2} > 0$ after which it decreases towards zero at $t \to \infty$. Thus, the log-normal is bell-shaped, but, in contrast to the Gaussian, the log-normal pdf is not symmetric around its maximum at $t = e^{\mu-\sigma^2}$ and can be seriously skewed.

The expression for the log-normal pdf in (4.2) can be rewritten in a "power law"-like form as

$$f_X(t) = \frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} t^{-\alpha(t)} \tag{4.13}$$

where the exponent $\alpha(t)$ equals

$$\alpha(t) = 1 + \frac{\log t - 2\mu}{2\sigma^2} \tag{4.14}$$

which illustrates that a log-normal random variable behaves as a power-law random variable, provided the last fraction in (4.14) is negligibly small, say ε. The latter happens when $\left|\frac{\log t - 2\mu}{2\sigma^2}\right| < \varepsilon$. Thus, when $t \in [e^{2\mu-2\sigma^2\varepsilon}, e^{2\mu+2\sigma^2\varepsilon}]$, the pdf of a log-normal random variable is almost indistinguishable from the pdf of a power-law random variable with $\alpha \approx 1$. The t-interval $[e^{2\mu-2\sigma^2\varepsilon}, e^{2\mu+2\sigma^2\varepsilon}]$ exceeds the maximum $e^{\mu-\sigma^2}$ of the log-normal pdf and is clearly longer when σ is larger (as well as the tolerated accuracy ε increases).

Taking the logarithm on both sides of (4.2) shows as well that a log-normal may look like a straight line.

$$\ln(f_X(t)) = -\frac{1}{2\sigma^2}\ln(t)^2 + (\frac{\mu}{\sigma^2} - 1)\ln(t) - \ln(\sqrt{2\pi}\sigma) - \frac{\mu^2}{2\sigma^2} \tag{4.15}$$

If σ is large, then the second term $(\frac{\mu}{\sigma^2} - 1)ln(t)$ dominates, which leads to a straight line with in a log-log plot, which resembles a power-law with exponent $\gamma = 1$. On the other hand, if σ is small, the first, quadratic term in (4.15) dominates.

4.2. FAKE FOLLOWERS

Human behavior is not the only influence affecting the topology of an OSN. Since the number of friends or followers became some kind of measure of how "influential" or popular a person is, companies established how to sell followers or friends in OSNs. Such companies are typically operating websites which users of various OSNs (Twitter or Facebook for example) may log in to, using their corresponding login of the OSN.

These follower or friend markets typically offer two different ways of using them. Users may chose between free and paid usage. In case of the free usage, a user accepts the terms of usage and grants permissions to the market operator to post messages with the user's account and allows the market to create relations towards other users. In most cases, when having the permission, the company may use the user's account transparently with full functionality. In return the user receives between 20 to 100 followers (in the case of Twitter).

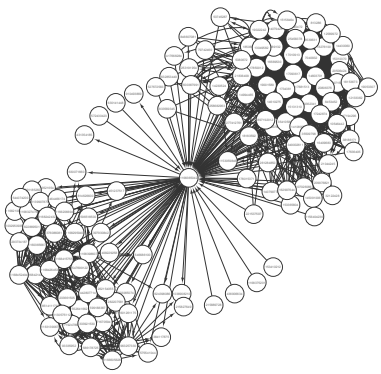


Figure 4.12: Example of the ego-centric network of a free market user.

In the second scenario, a user pays for a certain amount of followers, usually around 14$ for 1,000 followers, 70$ for 10,000 or up to 400$ for 100,000 followers. As a customer of a market, the user does not grant any permissions to the service. This denotes that a customer after paying the market just needs to supply the Twitter user name, which is a unique identifier sufficient to receive followers.

This means that a "free" user is actually at the same time the product of the market, whereas in most cases a "free" user only receives followers once but the market may con-

nect the user to others or post messages until access is revoked. Most services therefore offer to add 1-10 followers per day in order to keep permissions to facilitate the user's account.

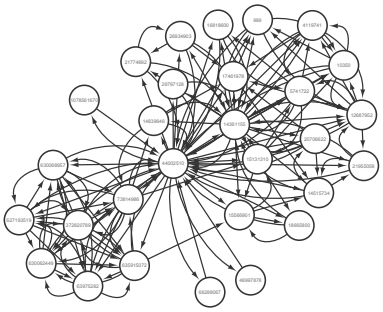


Figure 4.13: Example of the ego-centric network of a randomly chosen user.

Stringhini *et al.* [175] described economic effects of follower markets, and found techniques to identify "free" users of relationship markets by analyzing posted messages. If a person used a market for free, a tweet written by the user appears in the timeline, advertising the used market. Once knowing a set of "free" users, identifying market customers becomes possible because the "free" users will follow them. This means, if there is a significant number of free users in the list of followers of a user, who is not a free user, chances are quite high that the person bought followers. Also the ratio of friends to followers t_r indicates if a person bought followers, as markets sell batches of multiple thousands. However for famous personalities it is still possible to have a low ratio of t_r without having "fake" followers. Stringhini *et al.* [175] did not estimate the influence onto the friendship network, if markets, or robots employed by the markets, do connect users to each other, rather than the individuals themselves.

To estimate how the market robots connect user accounts, 20 Twitter accounts without any friends or followers were created, which attended a market as free users. After only one day, all accounts received between 50 to 185 followers but the number of friends per account grew by approximately the same number. Figure 4.12 shows the directed ego-centric network of a randomly chosen user account that attended in a market. Compared to Figure 4.13 showing the directed ego-centric network of a randomly selected Twitter user, one may observe the difference in the number of nodes and links.

The central nodes in both Figures 4.12 and 4.13 denote the selected ego. Apart from the size of the selected two networks, they visually appear similar. Both users seem to be connected to two larger communities, having some acquaintances which are only connected to the ego, but closer visual inspection shows that in the ego-centric network of the user account attending a marked the left community denotes accounts the ego is following and the right one consists of users that follow the ego. For the randomly selected node both communities consist of users that nearly equally follow or are followed by the user.

Figure 4.14 depicts more differences between randomly selected users (green) and user accounts facilitating a market (red). The figure shows box plots for certain metrics, calculated for the ego-centric networks of the aforementioned accounts attending a market, compared to metrics of randomly selected users.

The boxes from left to right in Figure 4.14 depict the logarithm of the number of nodes in the ego-centric networks, the logarithm of the number of links, the assortativity, the density, the diameter, the average path length, the transitivity (the global clus-

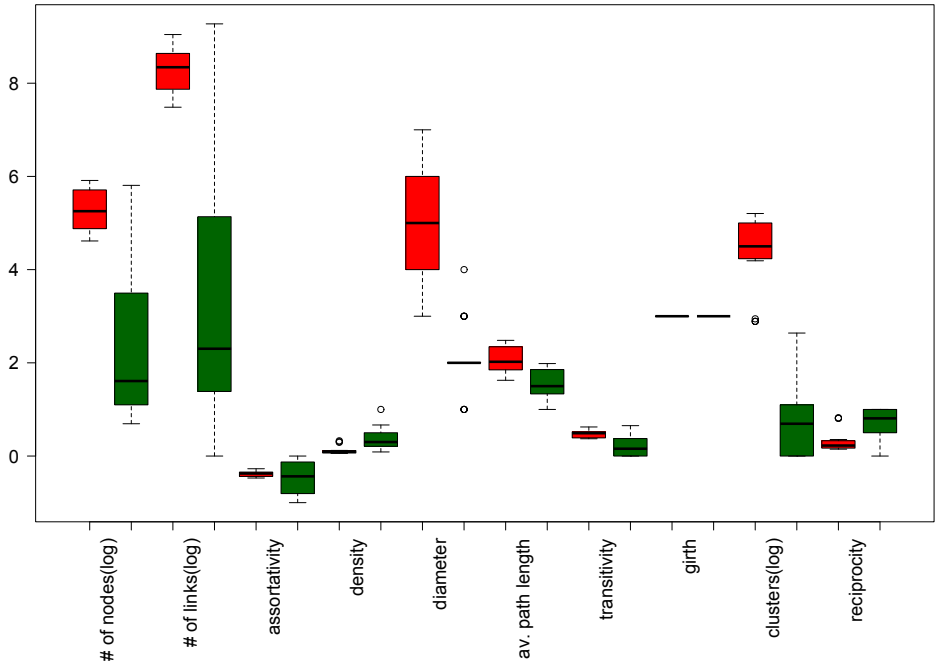


Figure 4.14: Analysis of 20 ego-centric networks of free market users(red) and randomly chosen users(green).

tering coefficient), the girth (the length of the shortest circle), the number of strongly connected clusters and the reciprocity. The boxes themselves are visualizing the median of all values by the black line inside the box. The top and lower line of the box denote the upper (Q_3) and lower quartile (Q_1) which means 50% of all values lie in the box. Whiskers denote the lower and upper interquartile range ($IQR = Q_3 - Q1$), which are calculated by $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$, of the data whereas outliers are plotted as individual circles.

One observes that the number of nodes in ego-centric networks of market users is higher, having a smaller variance, which arises from the fact, that all users received a similar amount of followers and friends which is rather high compared to random twitter users. The number of links in networks of market users is also higher as market users are connected to other market users, which are already connected to each other. In terms of the assortativity, less variance compared to randomly selected users can be seen whereas the average density of networks of market users is smaller because of the higher number of users in the friendship network. The diameter for networks of market users is higher which can be explained by the employed robots connecting user accounts. In a "real" friendship network friends usually know each other and so do friends of friends know each other. If a robot connects the useraccounts following the rule that a certain amount of users should connect to a user account, it would be computationally quite expensive to connect only pairs of friends. Therefore the assumption seems valid that users from the market are randomly chosen and connected to the ego, which denotes

an increment in the diameter (the longest shortest path in a graph). The same effect explains the shorter average path length for randomly selected users. In terms of the number of strong components and the reciprocity the "un-social" approach of connecting user accounts by a robot becomes even more visible. In "real life" friends do reciprocate a friendship quite often, if the follower is known to them whereas a robot does not consider such link- formation rules. Due to a low number of reciprocated links the number of strongly connected clusters is higher.

The aforementioned analysis was conducted on 20 randomly chosen accounts and 20 artificially created accounts that used the market. These numbers are rather small and cannot be used to measure the influence of markets onto the graph of 240 million nodes (users) of Twitter. As some markets like "hitfollow.info" or "followback.info", for example, display the user names of 10 to 20 users who joined their service in the past. By repeatedly monitoring the service of plusfollower.info, a list of 4,800 Twitter user names was obtained to further analyze the dynamics of the robots employed by follower markets. The 4,800 users directly connect to 1.755 million others weakly connected in one giant component. The number of strongly connected components is 415,539, where the largest strongly connected component contains 75% of all users. When comparing the egocentric networks of the 4,800 users of the market to randomly chosen ones the findings are similar as shown in Figure 4.15.

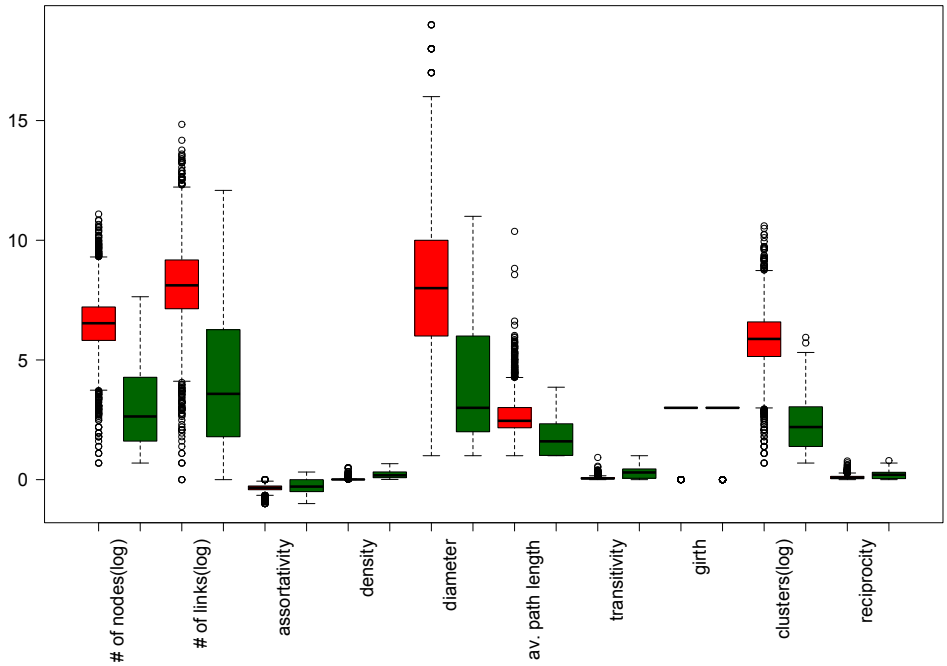
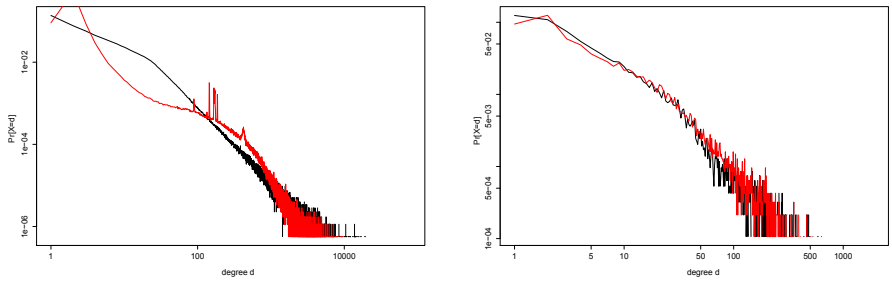


Figure 4.15: Analysis of 4800 ego-centric networks of free market users(red) compared to randomly chosen users(green).

From the calculated metrics one may assume that users who joined a market are

randomly connected to each other in a way which differs from the "natural", yet not well understood, way in which a OSN evolves. A comparison of the degree distributions of the union of ego-centric networks of market and randomly selected users furthermore depict the influence of market robots onto the friendship network of market users shown in Figure 4.16(a) and 4.16(b).



(a) Degree distribution of the union of egocentric networks of market users. in-degree (red), out-degree (black)

(b) Degree distribution of the union of egocentric networks of randomly selected users. in-degree (red), out-degree (black)

Figure 4.16: Degree distribution of the union of egocentric networks of market and randomly chosen users, in-degree (red), out-degree (black).

The degree distribution of the union of graphs of randomly selected users follows the "ground truth" presented in Figure 4.1 quite well, whereas the in-degree distribution of market users deviates quite heavily.

If the assumption, based on the previous analysis, that links are added randomly by the market robots to accounts of "free" users, the deviation of the degree distribution can be calculated.

If one considers for simplicity an undirected graph $G_0(N,L)$ with N nodes and L links, then the degree of a randomly chosen node in G_0 is denoted as D_{G_0}. Given a degree distribution $\Pr[D_{G_0} = j]$, the degree distribution of the graph G after randomly adding m links is denoted as $\Pr[D = k]$. Adding a link randomly, denotes randomly changing an element in the adjacency matrix to be one where the probability the position in the adjacency matrix is already set to one equals

$$p = \frac{m}{\frac{N(N-1)}{2} - L} \tag{4.16}$$

and the probability a random node has degree k given it's degree was $D_{G_0} = j$ is

$$\Pr[D = k | D_{G_0} = j] = \binom{N-1-j}{k-j} p^{k-j} (1-p)^{N-1-k} \tag{4.17}$$

because $N-1-j$ denote the values of the adjacency matrix having a 0 in the row corresponding to the randomly chosen node and $k-j$ are changed to 1 with probability p.

The remaining positions are not changed (with probability $1-p$) which leads to

$$\Pr[D=k] = (1-p)^{N-1-k} \sum_{j=0}^{N-1} \binom{N-1-j}{k-j} p^{k-j} \Pr[D_{G_0}=j] \tag{4.18}$$

The effects of randomly adding links to the degree distribution of a network that originally had a scale-free distribution is shown in Figure 4.17.

In Figure 4.16(a) and 4.16(b), the in- and out-degree distribution of randomly selected users follows a power-law convincingly whereas in Figure 4.16(a) the in-degree distribution deviates. The "lack" of users with an in-degree lower than 100 and the over representation of users with more than 100 followers shows that the market heavily influenced the distribution.

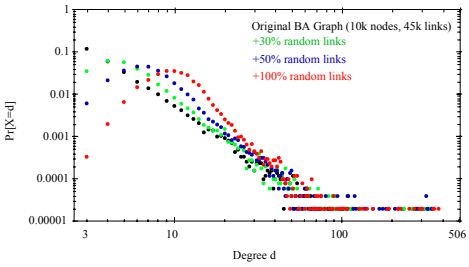


Figure 4.17: Degree distribution of a graph generated by using Barabási-Albert's model (black), with randomly added links (colors).

Another interesting fact of market users, and the reason the in-degree distribution in Figure 4.16(a) still looks different from the one presented in Figure 4.17, is the game market users are attending in. As it is important for market users to have a higher number of followers than friends, most individuals facilitate the market service, but manually remove friends they are connected to by the service. This behavior, an interesting game theoretical problem in it self assuming selfish behavior, may be the reason for the in-degree distribution depicted in Figure 4.16(a), but further research is necessary to prove this assumption.

However, in terms of content propagation, buying followers actually extends the number of individuals one may "reach" because chances are high, that the ones connected to a user account by a market would probably never connect to the account by themselves.

4.3. SATURATION EFFECTS

As mentioned before, human behavior is not the only but usually the main driver, affecting the topology of an OSN because automated services may create links and nodes as well. One may use the well known model of Barabási and Albert [91] to create graphs having similar topologies compared to social networks. As mentioned by Barabási and Albert, a power-law degree distribution only appears if preferential attachment occurs together with a steadily growing network.

In reality though, it may happen that an OSN saturates in terms of the number of registered users where these will stay active. Such a process, observed for the Dutch OSN Hyves.nl where the number of links within the network kept on growing because users befriended others whereas new users did not join the network anymore, states the reason that certain topological measures deviate from model parameters. Because of

the lack of "new" users with a low degree, i.e. a small number of friends, a skewed degree distribution, clearly not very well described by a power-law distribution, as shown in Figure 4.2(a) on page 81 evolves. The dates users registered to the OSN are depicted in Figure 4.18 showing that most individuals joined the service between 2008 and 2010.

Having a number of nearly 10.8 million users in 2010 who were mainly Dutch citizens, the OSN saturated because the total number of inhabitants of the Netherlands at this time was circa. 16 million. However, a high proportion of the users kept on using the service for more than two years which led to the fact that the network became quite dense. That means the average number of friends nearly doubled from 85 in 2008 to 177 in 2010. From a graph perspective the growing number of links has an influence on numerous other metrics, like a increasing density, shrinking diameter, average path length etc.

When comparing the other OSNs depicted in Figure 4.18 to the trajectory of Hyves.nl one may observe that a saturation is not visible. The degree distributions of these OSNs (Deviantart.com, Ratebeer.com, Digg.com and Twitter.com) are described by a powerlaw way better as shown in Figure 4.2(b) on page 81 for Deviantart.com, Figure 4.1 on page 80. For the network of Ratebeer.com it was not possible to estimate a degree distribution, because the friendship relations are not publicly available from the service.

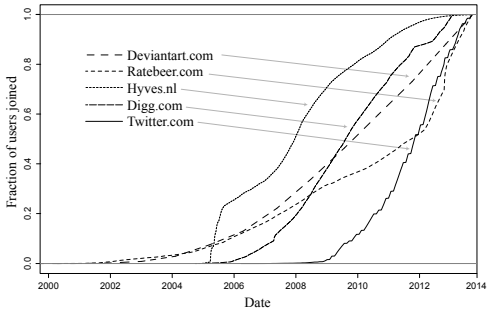


Figure 4.18: EDF depicting when users registered at Hyves.nl, Deviantart.com, Ratebeer.com, Digg.com and Twitter.com.

Similar effects are observed for other OSNs as well. Ugander *et al.* [3] for example, analyzed the friendship graph of Facebook.com, presenting a degree distribution of Facebook.com in 2011 with a similar shape as the one for Hyves.nl. Backstrom [176] showed that the average shortest path length between all active Facebook.com users ($\approx$ 721 million user accounts and $\approx$ 69 billion friendship relations) was 4.74 which denotes 3.74 intermediate persons in 2011. Following Milgram's experiment [1] where the average number of intermediates was between 4.5 and 5.7 which led to the famous assumption of the "6 degrees of separation", Backstrom claims that "the world is even smaller than expected" [176].

However, as mentioned before and as stated in Lescovec *et al.*[90] the fact that the average shortest path length in OSNs becomes shorter is a side effect of a network getting more dense. The statements by Backstrom that "the world is even smaller than expected" or conclusions stating that Facebook.com is connecting the world are therefore two-fold. On one hand, average shortest paths are not really reflecting the true distance between individuals because the strength of friendships should be considered to find a possible connection towards an other person. On the other hand it's not really clear what a distance > 2 actually means. It is possible that in social environments a distance > 1 already denotes that it is impossible to reach such a person within a short time. Further research

is needed to explore the meaning of metrics like the average shortest path length, but also others, like centrality and betweenness.

A different measure of saturation within OSNs is given through the number of "orphan" accounts, i.e. accounts that are not used anymore. If users decide not to use a certain OSN anymore they usually do not delete their account. In order to estimate if an account is not used anymore, operators may keep track of the last time a user logged in. However, as such data usually not available to researchers a stronger definition has to be taken denoting that an account is considered abandoned if no messages were sent for a long time and the number of friends and followers are below some threshold. Just analyzing the number of sent messages in order to identify inactive users would overestimate the amount to a large extent because a high fraction of Twitter users (ca. 43%) use the service merely to receive information (news and advertisements) without ever submitting a message. This amount also includes a high number of inactive accounts which can be estimated by analyzing the number of friends and followers. Out of all profiles in the used dataset 43% never submitted a single message whereas 39% of these accounts were registered more than half a year ago. Out of these 39%, 25.5% of profiles had no friends or followers at all, which denotes that at least 9% of all user accounts can be interpreted as "orphan" accounts, given the definition above. Of cause this number does not include accounts of users that were active once and left the service after some period of activity. The identification of users that are not using their account anymore after being active for a certain time, becomes difficult without knowing if a user is "passively" using a service.

These "abandoned" profiles account for a biased view when analyzing the network of an OSN in terms of topological saturation effects because their degree or the position within the network will not change, especially if the profiles of friends or followers are also not maintained anymore.

4.4. CHAPTER SUMMARY

In this chapter, the evolution of an online social network is described. The main factor for growing OSNs is described by activity patterns of users who may create friendship relations every time they are online. Therefore the duration users need between adding friends and receiving followers in a directed OSN are analyzed and it is claimed that this behavior might be well described by a log-normal distributions rather than power-law distributions. As a power-law is usually fitted to data with a lower and upper bound, the log-normal distribution approaches the measurements in a wider range without bounds.

Another influence on the topology of OSNs, given though markets which automatically connect users to each other is explained. Users of these markets are randomly connected to other participating users within short time. Due to the randomness of selected others within the markets, a process which differs from the one standard users would follow, the usage of markets can be detected by analyzing the degree distribution of nodes within the social graph. In terms of content propagation these markets might actually increase the potential of users to spread content, but further research is needed to quantify this claim.

5

TRACING CONTENT PROPAGATION

Users in Online Social Networks (OSNs) are sharing experiences, feelings and opinions about almost every aspect of life through status updates, direct messages and forum posts within different platforms. Therefore, one may call users of OSNs social sensors, because whenever something happens, someone gets sick, a new product is released etc. users will write about it immediately. This large repository of individual user data provides new and unique insights. For example by observing and analyzing broadcasted information, outbreaks of diseases can be detected earlier [28, 177, 178].

Every now and then, some piece of content is more actively passed along among users and spreads unusually wide, thus going "viral" and becoming an Internet "meme". One crucial part is to understand how such memes are formed and how exactly content makes the transition to "viral" spreading. Therefore a rigorous understanding how messages spread through an OSN is necessary. Chapter 5.1 explains a method to obtain such useful data. In Chapter 5.2, the relation towards epidemiological spreading will be drawn relating to the initially stated question, if "viral" spreading can be modeled by epidemiological methods. In this way, users become "infected" with information if they receive a message and infect others by forwarding messages. The time, users need in order to forward information may therefore relate to the spreading time within epidemiological models which is explained in Chapter 5.3.

5.1. DIFFUSION CASCADES

Messages are in most cases passed along the friendship relations of a particular user within an OSN. Typically OSNs (like Twitter, Weibo, Facebook, Digg, etc.) automatically notify the connected acquaintances about posts, status updates, or changes submitted by a user. The relations between users are directed in many OSNs: if a user A follows user B's activity, then B is called A's friend. From B's perspective, A is a follower, who will receive all messages written or forwarded by B. The friendships within Twitter.com or Digg.com for example, are formed without additional agreement of the friend, which explains why some users in Twitter have literally millions of followers. On the other hand it is not possible for users to create links towards followers themselves.

As already mentioned, content will traverse links in the opposite direction of the friendship relation. As OSNs are typically observed passively by listening and searching for content, it is not possible to distinguish whether a user who received some information actually read it or not. Therefore, the following assumptions enable the analysis of content propagation in OSNs using epidemiological models:

1. Common practice [137, 179] is to consider everyone who forwards information as infected. Users that receive a message without forwarding it are called informed or exposed.

2. A user may forward each message only once and he will always forward it to all of his followers at the same time. In many OSNs as well as Twitter, which is the focus of this section, status updates are automatically sent out and immediately appear in the inbox of all connected followers. In the case of Twitter, incoming messages can only be forwarded once through an action called "retweeting".

3. A person stays informed when he received a message and he cannot receive the message again. In Twitter, when a user receives multiple identical retweets by his friends, only the first copy is directly visible in the user interface.

A variety of epidemiological models have been proposed to capture the unique behavioral characteristics of a particular contagion. From this class of models, the *S*usceptible-*I*nfected-*R*emoved (SIR) model and its closely related *S*usceptible-*E*xposed-*I*nfected-*R*emoved (SEIR) model are the most suitable to accurately capture Twitter's information diffusion process. A user gets informed by one of his friends, is exposed to the message, and stays informed. By forwarding the message to his followers, he becomes "removed", because it is not possible for him to forward the information again. Therefore, he cannot become infected or exposed/informed with the same information again.

Certain metrics are used in order to describe epidemiological processes, such as the basic reproductive number R_0, commonly used to quantify the infectiousness of a virus. It is defined as the expected number of secondary infections an individual is causing during his entire infectious period given the population around him is susceptible [180]. It will be shown that messages and their subsequent diffusion cascades in Twitter can be interpreted as viruses having different basic reproductive numbers. A second key metric in epidemiological models is the transmissibility of a virus [181], which is described as the probability of a user to infect a neighbor, or in our analysis of Twitter the ratio of informed followers over the number of infected ones. The transmissibility in the case of Twitter is very small and close to a constant value.

Apart from epidemiology, epidemic models are widely researched in the fields of social sciences, physics and mathematics. Although the terminology changes for every discipline, the basic methods remain similar. A nice overview of the basis of epidemiological theory and its connection towards graph theory and different network types can be found in Keeling and Eames [182].

Most widely researched epidemic models are the SIR and Susceptible-Infectious-Susceptible (SIS) models. In the SIR model, every node of the graph can be in one of the following three states:

- *S*usceptible: the node is healthy and can be infected,
- *I*nfective: the node is infected and can pass the disease on to its neighbors with an infection rate β,
- *R*emoved: the node was infected, spread the infection on, cured with rate δ and is removed from the population.

In the SIS model, the first two states are the same as in SIR but an infective node may become susceptible again with a curing rate δ. This means that it can be infected with the same virus multiple times.

Two basic assumptions are usually taken in these mathematical models:

1. The distribution of infection times t_r, during which a node is infectious, is exponentially distributed to allow the use of Markov theory.
2. The network of nodes is homogeneous: each node is infected with a same strength given by a rate β and cured with a curing rate δ.

One of the most interesting questions is about the existence and the value of an epidemic threshold. In every disease or spreading process this threshold describes how an epidemic will develop i.e., if it will explode or die out. In viral marketing the epidemic threshold describes that a message having a smaller infection rate than the threshold will not spread very far whereas one above the threshold could reach a high number of individuals.

Pastor-Satorras and Vespignani [183] analyzed data from computer virus epidemics by using the SIS spreading process in scale free networks. Together with Boguñá [184], they claimed that a scale free network has no epidemic threshold if the number of nodes goes to infinity. In simulations where initially half of the nodes in a network were infected, a steady-state is established after a few iterations of the spreading process. In this state a constant average density of infected nodes was found that scaled with the effective spreading rate $\tau = \frac{\beta}{\delta}$, which means in return that even for a very small τ, the virus may survive a long time in the network.

Newman [185] describes solutions for the SIR model. He claims that the assumption of a random network as basis for the spreading process is rather unrealistic because large scale computer networks like the Internet are assumed to have a scale-free structure as well as human contact networks or OSNs. Another assumption considered unrealistic by Newman is that all individuals have the same probability of infecting a neighbor and stay infected for the same time. His main observation is that the SIR process is equivalent to generalized bond percolation processes where nodes can be seen as points in a network and the chance to infect a neighbor is given by the uniform probability a link towards a neighbor exists [186]. Newman defines the transmissibility T (in the range of $0 \leq T \leq 1$) as the probability a susceptible person will be infected by an infected neighbor. The scale free networks used during simulations are modeled by a truncated power-law with the following degree distribution.

$$P_k[D = k] = \begin{cases} Ck^{-\alpha}e^{-k/\kappa}, & \text{if } k \geq 1 \\ 0, & \text{else} \end{cases} \tag{5.1}$$

The degree distribution function (5.1) can model a power-law as well as an exponential distribution, where C is a normalization constant, α and κ are also constants. When simulating virus spreads on networks with degree distributions (5.1), Newman shows that an epidemic threshold exists. This threshold is significantly lower compared to the one predicted for random networks. If the degree distribution of a network follows a pure power-law then he claimed that an epidemic transition exists only for a small range of exponents α. If the exponent is in the range of $3 < \alpha < 3.4788$, the model has an epidemic transition. However, if the degree distribution is not a pure power-law but exhibit small changes like a cut in the tail, there will always be an epidemic transition at finite T.

Vazques *et al.* [187] analyzed the impact of non-Poissonian activity patterns on spreading processes by looking at data about email worms. The spreading process was defined by the SI-model. They observed that the response time to forward an email is a heavily tailed distribution which can be fitted by a power-law with an exponent of $\alpha = 1$. This means that the average time a virus "lives", compared to the standard model assuming a Poissonian process, takes 25 times longer.

Iribarren and Moro [137] analyzed data of a viral campaign of IBM in which emails were forwarded between persons. Approximately 31,000 individuals attended the experiment. The cascades formed by tracing emails are described by tree structures: the longest tree had a depth of 8 hops, containing 146 email accounts whereas the smallest ones contained only 2 individuals. Interestingly, the number of emails sent to already informed individuals was less than 1%, leading to a large number of trees describing the process. Due to this relatively small number of "back" links, the percolation process can be modeled by an Bellman-Harris branching process [188]. Bellman and Harris [188] described a process in which a node, "born" at time t_0 branches out into n nodes at time $t_1 > t_0$ with probability p_n. This process is non-linear in contrast to the linearity of Markov processes. By analyzing how messages were forwarded by the attendees of the campaign Iribarren and Moro observed that the infection times are actually log-normal distributed. As infection times are usually assumed to be exponentially distributed, this finding contradicts the assumption that information spreads in the Internet very fast.

Doerr et al. [143] observed that the infection time is actually distributed as the convolution of two log-normally distributed processes. This means that the time of two consecutive actions of forwarding is based on two processes called the observation and reaction time. In terms of epidemic models this translates in the SEIR model to the time a node is *E*xposed and the time it is *I*nfectious. As it seems to be unclear whether a power-law or a log-normal distribution actually fits measured data of the reaction time, research on how these two distributions are connected can be found in [146, 163, 164].

In a second paper, Iribarren and Moro [189] describe that although there is a lowered epidemic threshold, still most rumors, innovations or marketing messages do not reach a significant part of the population. One reason is the value of information which, if perceived as low is not transmitted further. Another reason could lie in the dynamics of human behavior. Again the Bellman-Harris branching process was used to model results from a large word-of-mouth experiment where users recommended a newsletter to their friends with the chance of being rewarded. The data showed a small number of triangles in the propagation patterns which means the process can be visualized again by tree structures. It is claimed that the response of human activity towards a certain task

cannot be described by homogeneous models: they found that a very small part of all individuals (2%) has a significantly higher spreading rate than all other nodes in the network, which is also observed by Doerr et al. [34]. Irribarren and Moro further describe that the infection time is not exponentially distributed which means that Markovian theory is not applicable.

Van Mieghem et al. [190] use Markov theory to model SIS virus spread in networks. Since there is an absorbing state, the steady-state for any finite graph will be this overall-healthy state. However, also a metastable state of a Markov chain exists above the epidemic threshold. In the NIMEA (N-intertwined mean-field approximation), the epidemic threshold is given through $\frac{1}{\lambda_1}$ as the reciprocal of the largest eigenvalue of the adjacency matrix A of the network:

$$max(\sqrt{d_{max}}, E[D]\sqrt{1+\frac{Var[D]}{E[D]^2}}) \leq \lambda_1 \leq d_{max} \tag{5.2}$$

As the largest eigenvalue is never smaller than the average degree of the network, any finite network including scale free networks has an epidemic threshold and is not as vulnerable to viruses as expected. Moreover, the mean-field epidemic threshold $t_c = \frac{1}{\lambda_1}$ is shown to be a lower bound in SIS and SIR epidemics in networks [191].

"Tweets" in Twitter, appear on the user's personal status page in chronological order These tweets are by default visible to the general public, unless a user marked his profile to be private, and can be retrieved when performing a general search query on Twitter. All status updates of a user are shown in the interface of each of his followers. This interface also lists replies to a message, the number of retweets (i.e. the number of users that forwarded the message to their respective followers), as well as a favorites counter indicating how many users marked the message as one they like.

Through the "sample stream" interface of Twitter, a random sample of 1% of all messages written within Twitter can be obtained. Starting from this initial, systematic sample, Twitter's search API was used to retrieve all potential retweets of a message.Through searching its possible to find every tweet written in Twitter, except for messages marked as private.

For the scope of this thesis, the epidemic spread of content across friendship networks, the data acquisition process was limited to include only messages "traveling" along friendship relations, eliminating messages containing news or advertisements, content broadcasted by newspapers, radio and TV programs. By filtering for tweets referencing a service called Twitpic, a service provided by Twitter to include and reference images in tweets, a commercial influence is minimized because images submitted to Twitpic are mainly photos taken by users with their smart-phones which are not published or advertised elsewhere. Twitter is then facilitated to publish a message containing a short text about the image and a hyperlink linking to the picture hosted by Twitpic. Twitpic itself uses only Twitter user accounts and holds no social (friendship) relations itself.

Over the course of one month 20,493,701 tweets referencing 8,181,998 different Twitpic images were observed. Out of those messages, roughly 1.7 million were retweeted at least once and 11.5 million messages were not forwarded at all. Figure 5.1 shows the distribution of how often the messages have been forwarded. The data is fitted by

a power-law with an exponent $\gamma = 2.12$. The most forwarded message in the dataset with 27,980 retweets contained a reference towards a still image taken from a popular teenager movie. To overlay the information spread with the social graph of Twitter.com, additionally the names of all followers of the 2.7 million users in the dataset were collected. The used dataset in this analysis is therefore given by approx. 284 million users and their 5.4 billion friendship relations.

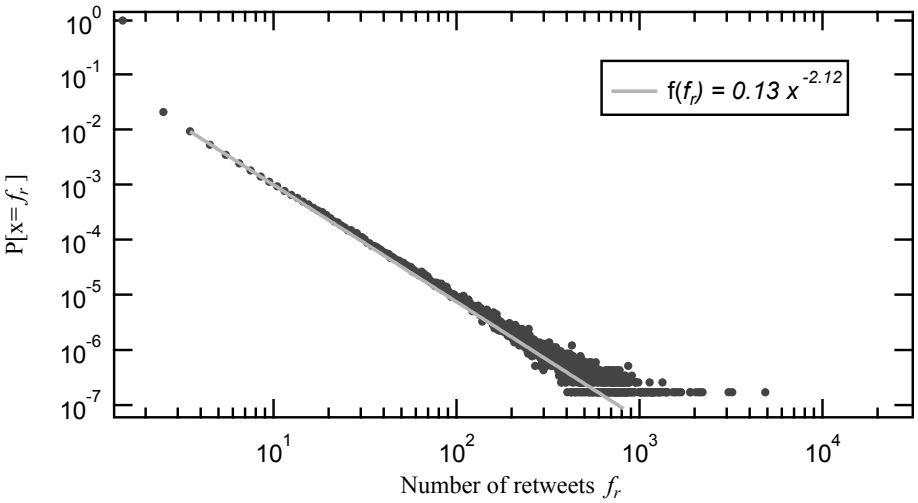


Figure 5.1: Distribution of the number of retweets.

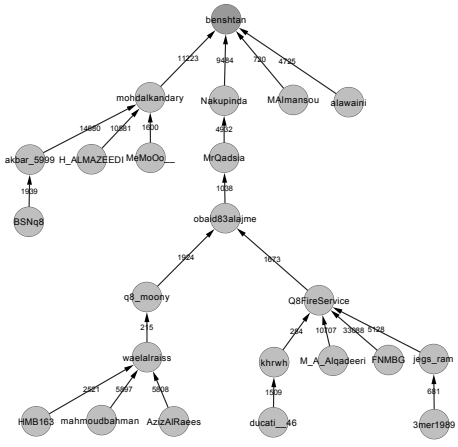


Figure 5.2: Randomly picked example of propagation trees.

The retweets of users can be assembled into individual trees, each describing the flow of one particular piece of information, together forming a forest of information diffusion across the microblog. Figures 5.2 and 5.3 show exemplary propagation trees from the analyzed dataset. The followers that forwarded the message are represented in green. The links are labeled with the time difference between receiving and forwarding the message. Only users that forwarded the message are included in the plot. As Twitter does not directly provide a trace of message propagation via its API, the trees are reconstructed from the set of individually observed retweets using the following algorithm which resembles the way Retweet trees were constructed in Kwak *et al.* [192]:

1. Tweets contain the unique username of the (retweeting) person and the username of the originator of the tweet, as well as a timestamp. Therefore, one does not directly know who informed a person.

2. In order to analyze which user was "infected" by whom, the friendship graph is used. All retweets of a message were ordered chronologically and it was checked if the person who retweeted is in the list of followers of the originator.

3. If the user who retweeted is a follower, then these two persons were connected.

4. If the user who was retweeting a message is not a follower of the originator then he must be a follower of a follower etc. or he was "activated" by factors, external to the friendship network.

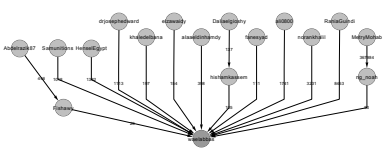


Figure 5.3: Randomly picked example of propagation trees.

As users may follow a person who retweeted and the originator at the same time, the third assumption could in principle lead to ambiguous reconstruction scenarios, if not for the specific behavior of the Twitter notification system. This can be demonstrated using a scenario of four test accounts. Consider the friendship graph between four users as depicted in Figure 5.4, where user 1 and 2 follow user 0 who was the originator of a message. The accounts 1 and 2 are informed by the Twitter interface that 0 wrote a message. In addition, suppose that user 2 is followed by both user 1 and user 3. After receiving the message from 0, 2 may retweet it and 3's interface will now list the original message from 0 as well as the information that 2 forwarded the status.

However, the interface of user 1 will still only show that the user 0 wrote the message, even though it was additionally forwarded by a second friend of his. Only if there is no direct connection to the originator of the message, like in the case of user 3, contextual information such as "retweeted by 2" is given through the user interface. In practice, this means that during the reconstruction of the information diffusion trees, the trigger for a retweet must be attributed to the highest common ancestor in the tree as subsequent duplicate notifications would not have been visible in the user interface, thereby eliminating this potential ambiguity.

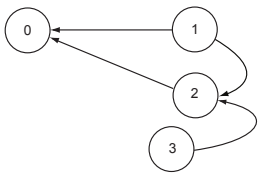


Figure 5.4: The friendship graph of 4 test accounts on Twitter.

To further demonstrate that a user in Twitter does not know how many of his friends already forwarded the message one may calculate if the number of friends who forwarded the message (and could have informed the user) has an influence on the probability to forward a message. In a real life epidemics this would translate to the probability an individual is infected given more of his friends are already infected.

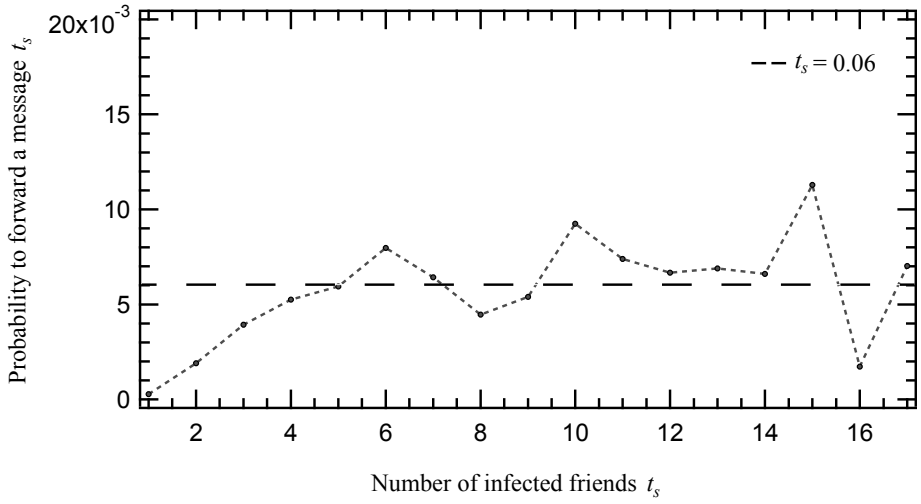


Figure 5.5: Probability to forward (retweet) a message given the number of "infected" neighbors.

Figure 5.5 depicts the marginal influence of the number of friends that retweeted a message on those that forward a tweet. Therefore one may conclude that Twitter's notification system, which is not informing about the number of friends that forwarded the message, is the main trigger used by individuals. This assumption is based on the fact that in different networks an effect of dependency can be observed. A few examples are:

- In real life, the more friends of a healthy person are infected by a contagious disease, the higher the probability the person gets infected.
- When buying new products most people tend to buy the items already owned by a number of friends [193].
- In online social news aggregators users tend to vote and read the stories, friends already read or voted [61].

A forest containing 1,713,624 trees having at least two nodes (a seed and one forwarding node) and 11.5 million seeds without any retweet within the observation period, each describing how one message was propagated, was constructed. The largest tree had 27,980 nodes and the deepest tree has a depth of 126 levels. The number of created trees provides a forest that allows to analyze the parameters of epidemiological processes.

5.2. A FOREST OF TWITTER CASCADES

The created forest of trees captures the process of content propagation of each individual message in Twitter. In the following, the epidemiological diffusion process of information using branching processes, starting from the most basic spreading process towards time-dependent propagation will be analyzed.

5.2.1. CASCADES DESCRIBED BY BASIC STOCHASTIC BRANCHING PROCESSES

The simplest possible view on viral spread is to ignore all contextual information about the propagation, such as the time it takes users to react to an incoming tweet, and only focus on the persons affected and involved in forwarding a particular piece of content. This most basic case of information diffusion is captured by the Galton-Watson branching process [194], originally intended to analyze the extinction of family names. Given that parents pass their family name only on to their sons, the process solves the size of the "infected" population (i.e. the number of families with a particular name) X_n after n iterations as $X_{n+1} = \sum_{j=1}^{X_n} \xi_j^{(n)}$. The probability for a family name to become extinct depends on ξ: clearly, for $n \to \infty$ the probability of extinction will be one if $E[\xi_1] \leq 1$ and less than one if $E[\xi_1] > 1$.

In the scenario of information propagation, $E[\xi_1]$ translates to the average number of infected neighbors who retweet a particular piece of content. Called the basic reproduction number R_0 in epidemics, it determines whether a contagious outbreak will die out ($R_0 < 1$) or spread ($R_0 > 1$). Another way of defining a continuous branching process is via the probability to infect a neighbor, called transmissibility [181].

Figure 5.6 shows the number of followers that retweeted a message versus the number of followers of users. Although a slightly positive trend can be observed in the plot (having a Pearson correlation of 0.35), a linear fit suggests that the slope is $9.56e^{-5}$ which is very low. The ratio of the number of activated neighbors to the total number of followers defines the percentage of activated followers (transmissibility), which varies between 1 and 3% with an average of 1.8%.

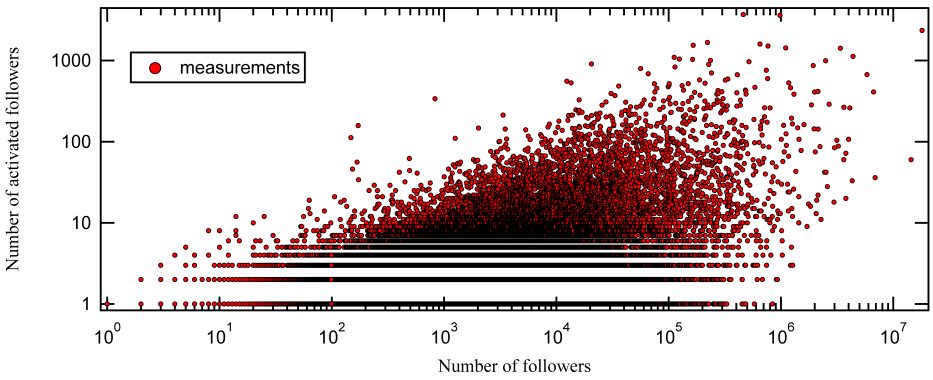


Figure 5.6: The number of followers that retweeted a message versus the total number of followers.

This low spreading ratio on Twitter is not surprising, as it has also been found in other OSNs: As reported by Doerr *et al.* [34] for the case of the social media aggregator Digg.com, only 2% of followers vote and thereby spread incoming recommendations from their friends. In contrast to Twitter, Digg users are able to see both the total number of users spreading a story as well as a full list of users and their friends doing so, which leads to a proportional amplification of a user's likelihood to follow suit described by Van Mieghem *et al.* [163]. This effect however requires a significant quantity of simultaneous triggers from within the friendship network to be robustly detectable shown by Doerr *et*

al. [61], which was typically not the case in the diffusion cascades on Twitter.

The observation of almost a constant infection probability strengthens the applicability of epidemiological models, as the basic *SIR* and *SIS* models assume a constant spreading rate, drawn from an exponential distribution, for every node in the network. As the relation to the so called epidemic threshold is not directly given by the aforementioned probabilities, the basic reproduction number R_0 given by the average number of infected neighbors for all messages can be calculated. It is found to be on average 0.9 in our dataset. Being smaller than one denotes that most Twitter messages, even if they are retweeted at least once, it will not reach a high proportion of users and become viral. This explains why only occasionally tweets gain sufficient critical mass to escape the Twitter ecosystem to become an Internet meme on their own or gain collective attention. Figure 5.7 shows the histogram of the basic reproduction numbers for all messages in our dataset in which every message can be interpreted as a virus of its own. The largest 20 values in terms of the basic reproduction number were between 7 and 40. For comparison, table 5.1 lists the basic reproduction numbers of a number of well-known infectious diseases [180], which do not die out and continuously resurface.

Disease	R_0
Measles	11 – 18
Pertussis	10 – 18
Mumps	7 – 14
Chicken pox	7 – 8
Scarlet fever	5 – 8
Diphtheria	4 – 5
Immunodeficiency virus	2 – 10

Table 5.1: R_0 Values of well known infectious diseases.

The basic reproduction numbers of tweets are estimated by taking only the first and second hop nodes in the propagation tree of a message, counting the average number of infected nodes one and two level below the spreader as suggested in Rothman *et al.* [195]. The relation of the lower bound of the epidemic threshold towards the transmissibility is given through the largest eigenvalue λ_1 of a network described in [190]. The lower bound is shown to be $\frac{1}{\lambda_1}$. However, as the whole network of Twitter is too large, the lower bound can be approximated by the average number of followers which is 763 and the fact, that λ_1 will be larger than this number. Therefore the lower bound of the epidemic threshold is given by $\frac{1}{763} \approx 0.00131$.

5.2.2. CASCADES DESCRIBED BY AGE-DEPENDENT STOCHASTIC BRANCHING

As discussed earlier, the Galton-Watson process is only a very fundamental approximation of the observed process. As every user needs time to infect his neighbors (by visiting the Twitter website, reading and reacting to incoming messages) and the spreading process dies out after a level of hops, the Bellman-Harris branching process provides a more realistic approximation [188]. The process describes an age-dependent stochastic

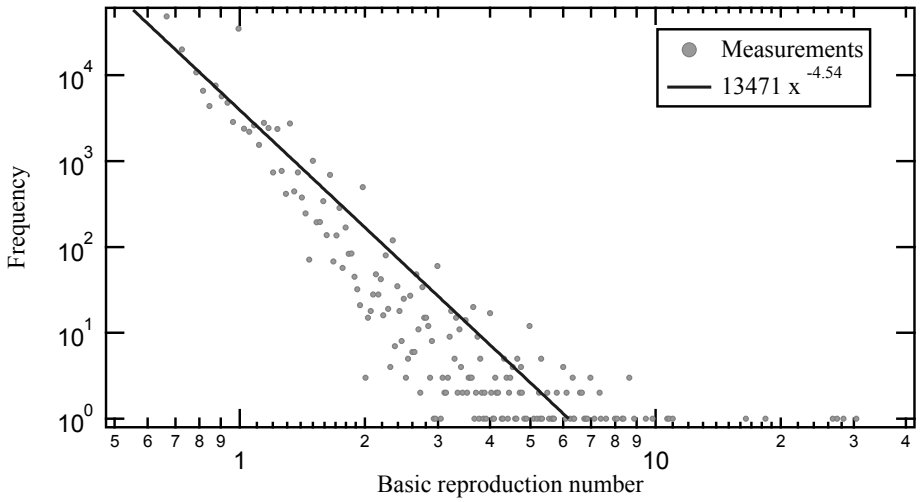


Figure 5.7: Histogram of the number of Tweets with a given basic reproductive number in our dataset of ca. 1.7 million messages.

branching process, where a node is split into n nodes at time $t > 0$ with probability p_n, under the assumption that the time to split is independent from the time of creation and the number of existing nodes.

Consequently, the time for an exposed user to infect his followers was analyzed. In order to apply Markov theory, this distribution should follow an exponential distribution. However as mentioned earlier, Vasquez *et al.* [187] found power-law distributions with an exponent close to $\gamma = 1$ for email worms and Iribarren and Moro [137] measured a log-normal distribution for the infection time. Similar values were found for the propagation cascades in Twitter, as shown in Figure 5.8. Here the fitted log-normal distribution has similar values as the ones found by Iribarren and Moro [137] who fitted their data with a log-normal with $\mu = 5.5$ and $\sigma^2 = 4.5$. However, the fit does not cover the whole range of the distribution. The reason lies in the fact that the time between two retweets was measured. This means that the measured data consists of the sum of the two distributions in the SEIR model of being exposed and infected. In terms of the action of forwarding, a follower who is *susceptible* becomes *exposed* once he receives a message. As a user needs some time to observe a message (read it), he is not *infectious* for this time-span. After reading the message the individual becomes *infectious* and may take some time to decide if and when to retweet.

The fitted distributions in Figure 5.8 approximate the data quite well which is noteworthy as Twitter and email communication (as in the viral campaign of IBM analyzed by Iribarren and Moro [137]) are technically quite different. In Twitter, once a person writes a new message, all followers are informed, whereas in an email one chooses who to inform. In addition, email communication one could actually forward information to peers at different times. In epidemiological terms this difference can be modeled, using an infection rate and a curing rate drawn from exponential distributions. In Twitter,

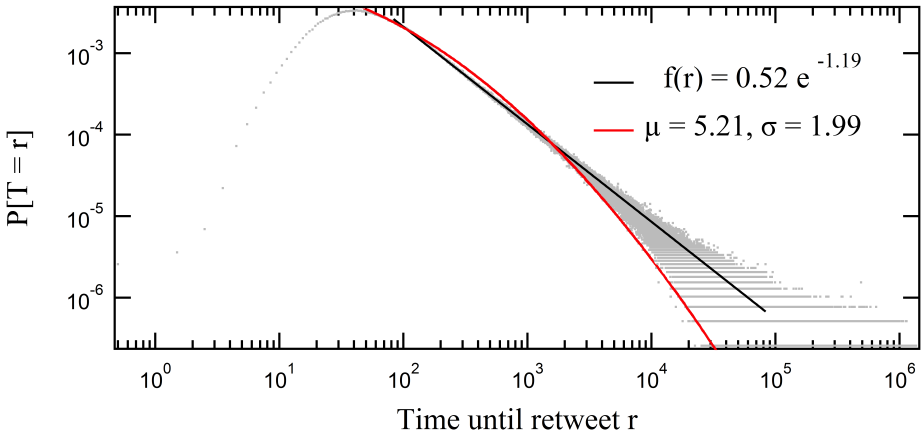


Figure 5.8: Distribution of durations until first hop retweet.

once the information is forwarded to all neighbors, the node gets removed immediately because it cannot get infected again with the same message.

The distribution of the depth of all trees, shown in figure 5.9, is well fitted by a power-law. In contrast, Figure 5.10 shows the distribution of nodes per level in the trees.

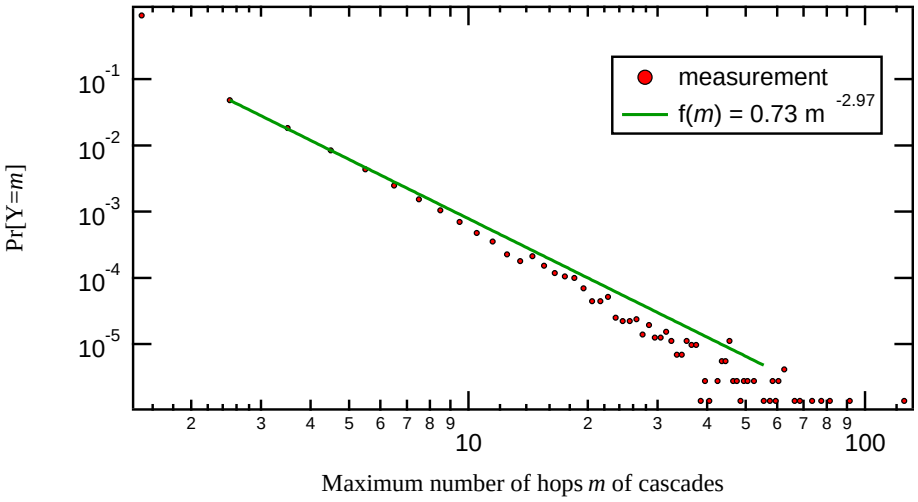


Figure 5.9: Distribution of the depth of all propagation trees.

In order to understand the distribution of the depth of trees, simulations were enforced in which undirected random (Erdős Rényi [196]) graphs with a size of 10,000 nodes were generated with different link connection probabilities between 0.001 and 0.01 in steps of 0.0001.

In every graph, a random seed node was selected to spread a message with given probabilities to infect the neighbors, between 0.05 and 0.3 in steps of 0.05. By keeping a list of infected and removed nodes, the spreading process was repeated until all infected nodes were removed. This means initially only a randomly selected seed node was infected, which spread to it's neighbors based on the infection probability and the procedure was repeated until no spreading is possible anymore. Additionally, the spreading process within one graph was repeated 10,000 times for all possible starting nodes.

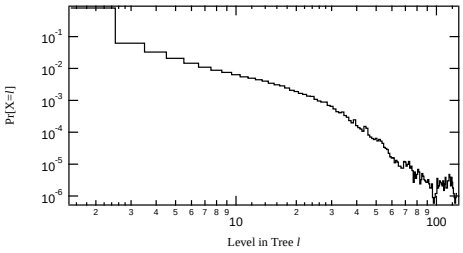
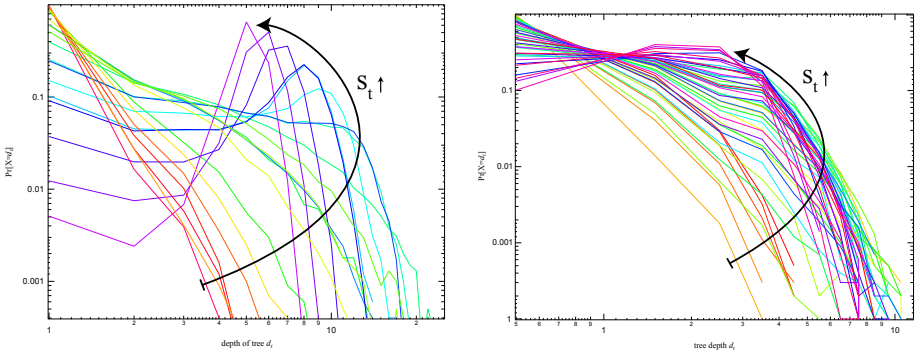


Figure 5.10: Distribution of the number of nodes per level of depth in the trees.

The Figures 5.11(a) and 5.11(b) suggest that the fraction s_t of the link density p_G of the graph to the transmissibility (infection probability) t, $s_t = \frac{p_G}{t}$ denotes the shape of the resulting distribution.



(a) Erdős Rényi graphs (b) Barabási Albert graphs

Figure 5.11: PDF of the depth of simulated trees.

As the spreading process is related to a discovery process along the graph, the total number of informed users per level in the tree, shown in Figure 5.12, follows a saturation curve.

The saturation effect occurs, on average, when approximately 450,000 users are informed, whereas the highest number of informed users is about 1.2 million. Given the self-reported size of Twitter at 241 million users [197], this saturation emerges surprisingly early on. Hypothetical explanations are:

1. Twitter consists of multiple communities that may overlap each other. As Twitter is used by people all over the world those communities are based on diverse spoken languages, the geographic location of users as well as different interests of users.

2. Timing issues related to the activity users and the probability that two friends are

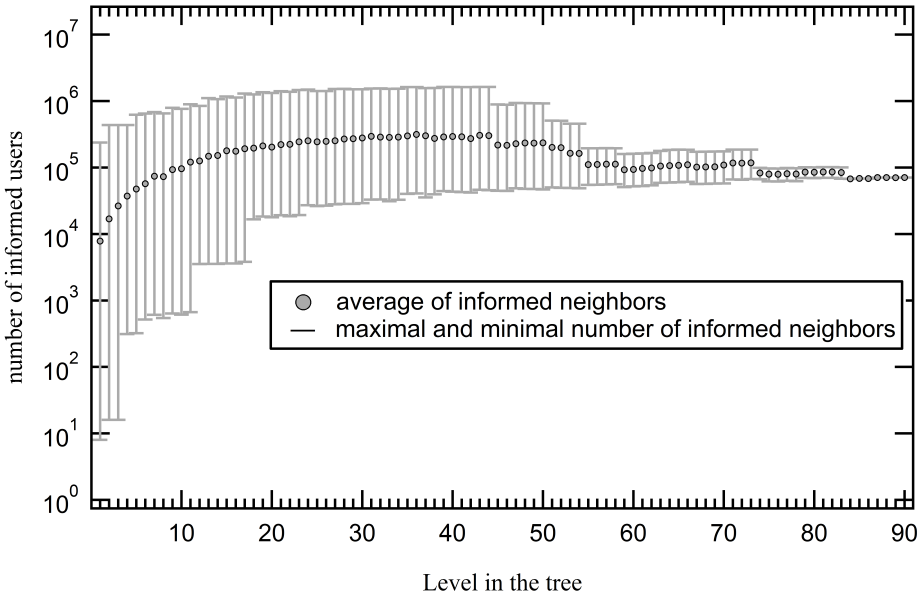


Figure 5.12: The average, maximum and minimum number of total informed users per level of trees.

online at the same time in order to forward the information may play a major role as found in Tang et al. [77]and shown before in section 3.4.

3. The directionality of the Twitter network might be a reason which increases the possible path length or may even lead to disconnected strong components.

5.3. INFECTION DURATION

The analysis of the observed temporal distributions of online human activity data such as the inter-arrival and forwarding times of email [137], the propagation time of microblog posts [198] or telephony holding times [158], have commonly been approached similar with the techniques established in topological network analysis, for example the approximation by power-law distributions. The applicability of fitting temporal behavioral data by a power-law has however been questioned [141, 142] and bears a number of complications. First, the approximation through a power-law primarily concentrates on the fat tail leaving the lower part of the observations unconsidered and provides no approximation or an over-estimation of the frequency of low values within the distribution (see the example fit in Figure 5.13).

Second, the relatively low power-law exponents found on temporal data militates against the presence of preferential attachment [199]. Third, while the process of preferential attachment provides an explanation for a scale-free degree distribution [91], this model cannot serve as a basis for and therefore does not provide any insight into the creation of propagation time distributions. There exists to this date no theoretical model able to explain the observed traces of online human behavior. In this analysis, a working

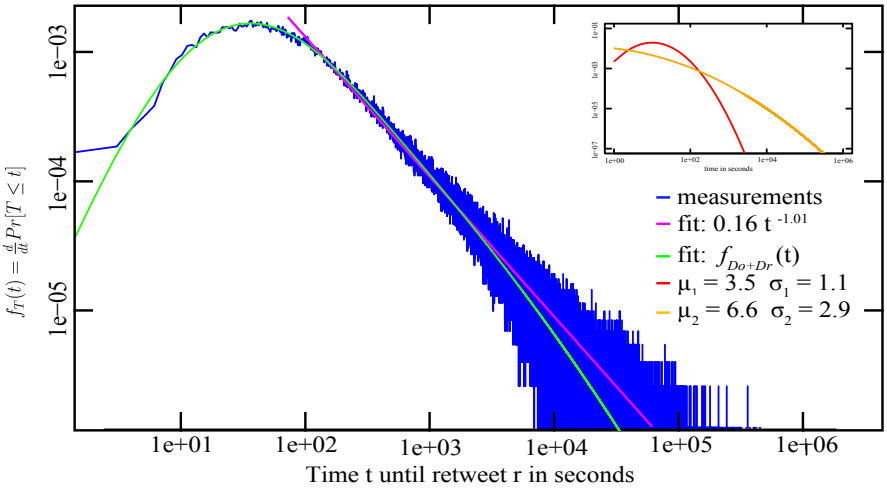


Figure 5.13: The tail of the probability distribution of the inter-arrival time T of retweets on the microblog Twitter.com, here tweets reporting the death of Steve Jobs, can be approximated by a power-law, the entire distribution of tweets is however better fitted by a convolution of log-normals.

5

hypothesis to explain the generation of this data will be put forward.

Log-normal distributions [200] have frequently been associated with human behavior, such as the time to complete tasks, duration of strikes, income frequencies, epidemic incubation times or marriage age [146, 154, 155]. Additional, the cumulative behaviors of many individuals interacting with each other has also been demonstrated to result in a log-normal distribution, one theoretical explanation being given for example by the law of proportionate effect [163, 166].

There are a few general processes that lead to a log-normal random variable. A log-normal random variable is defined [23] as $Y = e^X$, where X is a Gaussian or normal random variable. The corresponding probability density function of the log-normal random variable Y is:

$$f_Y(t) = \frac{\exp\left[-\frac{(\log t-\mu)^2}{2\sigma^2}\right]}{\sqrt{2\pi}\sigma t} \tag{5.3}$$

While a normal distribution describes well-behaved events, characterized by a mean and a standard deviation, such as the height of men and the temperature at noon in summer, the log-normal distribution exponentially blows up the relatively controlled deviations around the mean. Second, as a consequence of the Central Limit Theorem (CLT), the product of n independent identically distributed (i.i.d.) random variables tends for large n to a log-normal distribution for a sum of the logarithms of i.i.d. random variables. Slightly more general multiplicative processes (under certain conditions), such as the law of proportionate effect and geometric Brownian motion, give rise to log-normal behavior. Third, Marlow [201] demonstrates that, if a properly scaled sum of n random variables tends to a normal distribution $N(0,1)$ for large n, the logarithm of the sum (also

properly scaled) *also* tends to $N(0,1)$. In other words, both the scaled sum and the scaled logarithm of the sum converge to a same Gaussian $N(0,1)$, though at different rate. Yet differently stated, if the scaled sum of random variables tends to a Gaussian (by virtue of the CLT), then that same sum, though differently scaled, tends to a log-normal. Marlow's theorem proves the observation in radio communication that a sum of log-normal random variables also tends to a log-normal. These limit laws illustrate why log-normals may appear relatively frequently.

5.3.1. THE DISTRIBUTION OF THE SPREADING TIME T

In order to spread information, such as forwarding a message or news item on a social media platform, three consecutive processes take place as shown in Figure 5.14: First, after an e-mail, tweet or message has been sent by a person, the information is processed and physically copied across the network of servers of the large social media sites and is delivered into the inboxes and queues of the receiving users where they compete for the user's attention [5]. This first action requires D_n time units, called the network propagation time. Second, users have to become aware of the content, for example by logging into the platform. The time between delivery and observation is denoted as the observation time D_o. Third, users decide to actively spread the information, for example through actions such as "retweets", "likes", "diggs", thereby effectively forwarding the message to their connected friends and/or followers. The time between observation and passing a message is the reaction time D_r. The overall person-to-person forwarding time T is the sum of these three time components: $T = D_n + D_o + D_r$.

In the following discussion time measurements from the microblog service Twitter and the (former) social news aggregator Digg.com are used to explore the spread of information on online social media. For the case of Twitter one may measure T as the time to forward ("retweet") messages by the followers of tweet originators, on Digg.com T is denoted as the time between a user's recommendation (called a "digg") and the resulting diggs of a person's followers. For both these services, users see a summary of their friends' activities after visiting and logging into the website [61], and are presented with the opportunity to "retweet" or "digg" next to a particular piece of information.

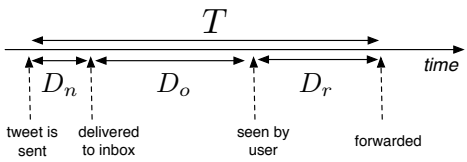


Figure 5.14: The three components of network processing, user observation and user reaction time together form the measured inter-arrival time T.

Measurements of the network time D_n indicated that D_n is about two orders of magnitude smaller than the observation and reaction times. The network effect can therefore be neglected and the overall person-to-person forwarding time can be approximated by $T \approx D_r + D_o$. Hence, T is, to a good approximation, determined by the habits and behaviors of the human participants.

As follows, two basic hypotheses are used:

1. The random variable D_r is independent of D_o, i.e. the time to react to a message does not depend on the observation time, which allows the probability density function of $T = D_o + D_r$ to be expressed as a convolution of those of D_o and D_r:

$$f_{D_o+D_r}(t) = f_{D_o}(t) * f_{D_r}(t) \tag{5.4}$$

$$= \int_{-\infty}^{\infty} f_{D_o}(x) f_{D_r}(t-x)\,dx \tag{5.5}$$

2. The hypothesis is based on previous findings ([137, 149, 154] that the observation and reaction times are log-normally distributed, $D_o \overset{d}{\sim} logn(\mu_o, \sigma_o)$ and $D_r \overset{d}{\sim} logn(\mu_r, \sigma_r)$, so that with (5.3),

$$f_{D_o+D_r}(t) = \frac{1}{2\pi\sigma_o\sigma_r} \int_0^t \frac{e^{-\frac{(log(t-x)-\mu_o)^2}{2\sigma_o^2}} e^{-\frac{(log(x)-\mu_r)^2}{2\sigma_r^2}}}{(t-x)x}\,dx \tag{5.6}$$

A maximum-likelihood estimation parameters of (μ_o, σ_o) and (μ_r, σ_r) for the measured spreading times on Twitter and Digg indeed generates a very good fit of the experimental data based on the two log-normally distributed time distributions.

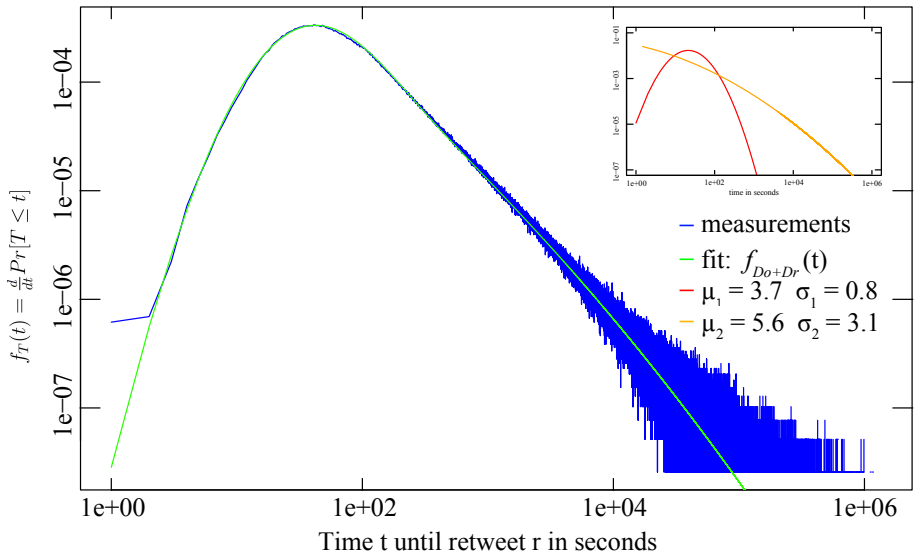


Figure 5.15: Distribution of spreading times on Twitter.

Figure 5.15 and 5.16 depict the experimental data and the maximum-likelihood fit of $f_{D_o+D_r}(t)$ for the retweet time on Twitter and upvoting time on Digg respectively. The

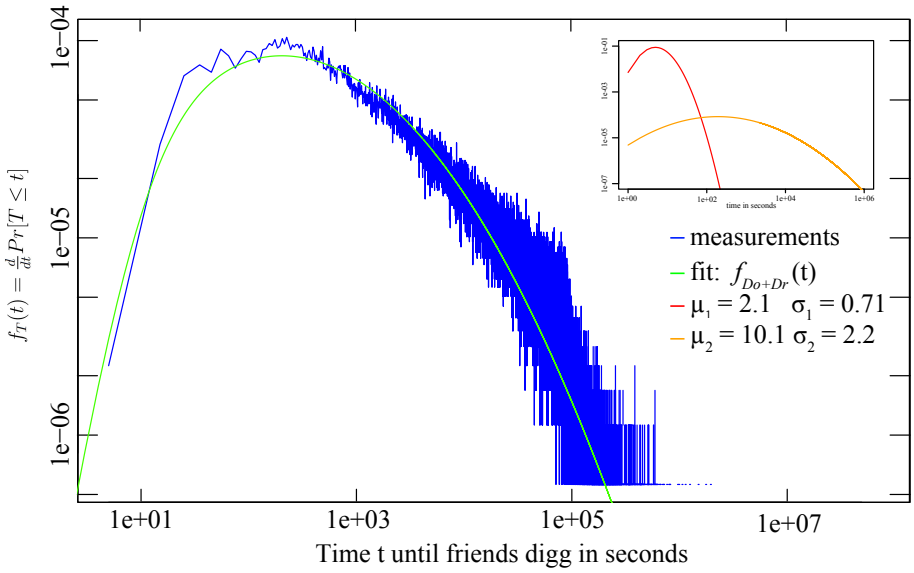


Figure 5.16: Distribution of spreading times on Digg.

figure insets show the underlying log-normal distributions. The ML-fit is based on a log-squared error function which provides comparable attention to the entire distribution and does not overemphasize its tail, as discussed in [202]. The ML-fits are conducted using a set of 20.5 million tweets and 310 million diggs.

As the convolution is based on four variables, one would assume that a broad range of data could be fitted with a four-parameter function. It should be noted that the convolution of a wide and narrow log-normal distribution is highly sensitive and only a combination of very small parameter ranges results in a good match with the observed Twitter data. Figure 5.17(a)-(d) depicts a mapping of all evaluated parameter combinations in the 4-dimensional fit optimization into four individual plots. The parameter combinations performing within 5% of the optimum are color-coded in yellow to red based on the goodness of fit and depicted in this color across all subfigures to be able to re-identify a particular scenario across the plots. As can be seen in Figures 5.17(a) and (b), even small variations of μ_1, σ_1 have a profound impact on the fit with a sub-exponential slope around the optimum, variations in μ_2, σ_2 (Figures 5.17(c) and (d)) have an attenuated but still significant impact as this distribution is very broad and mainly responsible for the tail of the convolution. As a result, only a few parameter combinations in a narrow region of each parameter space result in a good quality fit of the data, moving out of the optimum in any of the four dimensions significantly degrades the goodness of the fit.

Similar patterns can be found across independent data measurements. Figure 5.13 shows analogously to figure 5.15 the distribution of retweet times for the spread of the news that the CEO of Apple has died. Although this propagation is different from regu-

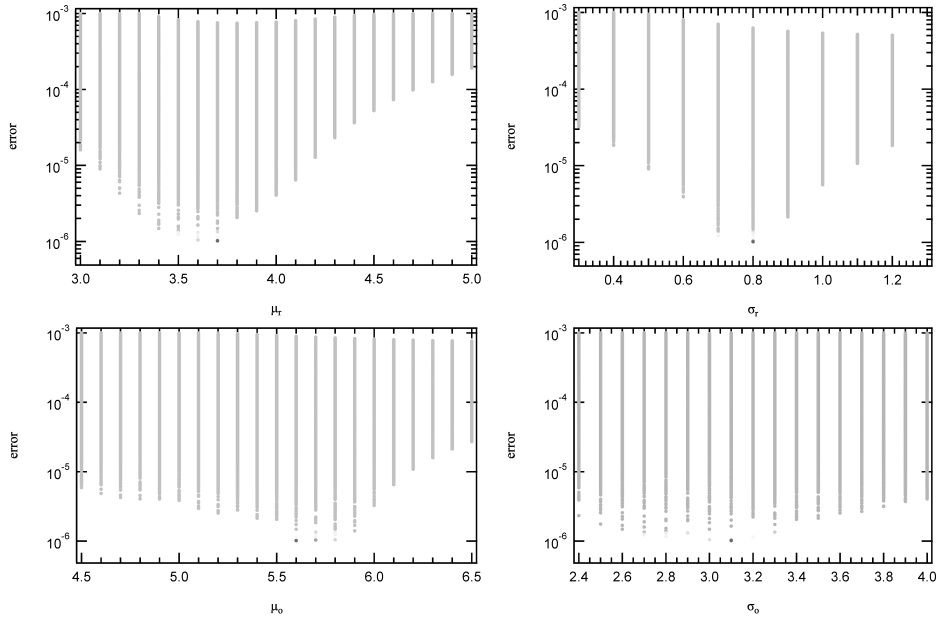


Figure 5.17: Fit error of the log-normal fit.

lar Twitter conversations shown in Figure 5.15 in that most participants were likely triggered by external sources (traditional media coverage) and the high timely relevance of this news let the content spread an order of magnitude faster than for un-influenced retweets, the same process of two convoluted log-normal distributions fits the observed data well.

The accuracy and the high sensitivity of the fit leads to the conclusion that the two assumptions (independence and log-normal distribution of the time components D_o and D_r) are realistic.

5.3.2. IDENTIFYING OBSERVATION AND REACTION TIME

As a convolution of two functions is commutative, $f * g = g * f$, it is not possible at this point to identify which of the two log-normal distributions describes observation and reaction time respectively. This section will therefore use empirical data from Digg.com and Twitter to make this differentiation. As directly obtaining measurements to assess the observation and reaction time is not feasible, these times will be indirectly inferred by monitoring a particular web site, and determine from the publicly visible record of status updates, friend requests or comments when a person is active or not.

For the the differentiation in the case of Digg.com, data is utilized on the times that their registered users upvoted or commented on a story published within the social news aggregator that until the sale of Digg.com in July 2012 could be publicly downloaded via its API. From this timestamped record of 310 million individual upvotes, the approxi-

mate coherent time periods a particular user is active on the website and thereby may spread information (Figure 5.18) was reconstructed. As the epidemic spread of information assumes the diffusion along social relationships, this simple view was augmented through information about the unidirectional follower and bi-directional friendship relations (i.e. reciprocal following) that users are forming on the Digg network. When "following" a user, the follower receives notifications about the other person's activities with the possibility to upvote and thereby spread the information in turn to his own followers and friends. The observation time in an epidemic spreading process across these social relationships is the time between a particular user has upvoted a news item, and this recommendation will be displayed on the follower's user interface on his next login on the website. The reaction time equals the interval between the notification and the receiver's upvote (if applicable).

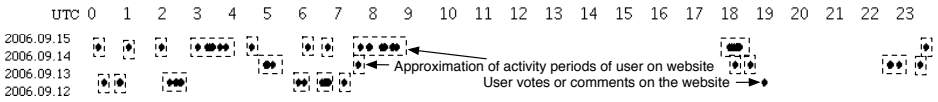


Figure 5.18: The record of a user's votes on the Digg.com website, each indicated by a dot on the timeline, provides a lower bound estimation of the user's presence on the website, as well as an upper bound on this user's observation times, exemplary for four consecutive days in September 2006.

Since the public record of diggs and comments does not directly list the instances a user has consulted his inbox to see the incoming notifications during a particular visit on the website, this monitoring procedure will only create a general bound of the observation and reaction times (see Figure 5.19). As the notification page is only visible after a login (and the login event must have been completed by the time the first digg was observed), the time between the original digg and the approximated login time of the follower will be lower bound of the observation time, provided that a particular person could have checked the incoming recommendations at any time point after the login. Correspondingly, the time between the login event and the person's own digg will therefore be *an upper bound of the reaction time*, as the notification triggering the digg could have been read anytime between the login and the actual action. The same approach is followed in case of Twitter, where the timestamps of tweets will be used to establish an estimate on when a user is present on the social network, leading upper and lower bounds of reaction and observation time respectively.

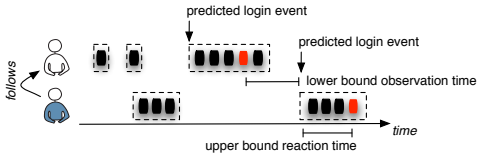


Figure 5.19: A lower bound estimate of the observation time and an upper bound of the user's reaction times can be determined when augmenting the voting behavior with the friendship network.

The analysis of the experimental data shows that both lower and upper bound of observation and reaction time are following a log-normal distribution, thereby confirming our initial hypotheses for the spreading time T. Table 5.2 list the ML-fit based on (5.6)

against the experimentally obtained parameters of the underlying distributions. The two distributions from the maximum-likelihood fits also line up on the correct side and neighborhood of the empirical lower and upper bounds, thereby making it possible to identify μ_1, σ_1 as the reaction time, and μ_2, σ_2 as the observation time. The measured and fitted observation time parameters μ_o, σ_o match well those predicted from our earlier maximum-likelihood fit. The mean of the measured reaction time also falls within the expected interval, between 0 and the computed upper bound, although the bound is less tight.

Table 5.2: Comparison of observation and reaction times of ML-fit and derived bounds from Digg and Twitter dataset.

	Digg		Twitter	
Measurement	μ	σ	μ	σ
Log-normal D_o (ML-Fit)	10.2	2.3	5.6	3.1
Lower Bound of Observation Time	9.48	2.75	5.5	2.9
Log-normal D_r (ML-Fit)	2.1	0.71	3.7	0.8
Upper Bound of Reaction Time	6.26	2.02	6.92	1.32

5.3.3. DISCUSSION

The hypothesis analysis and the experimental data point to log-normal distributions as an explanation of the temporal distributions observed in epidemic spreading online. This finding is not per se surprising as log-normal distributions have for a long time been observed in and connected to human behavior, which is driving the spread of information and innovation online.

The analogue to the reaction time D_r is, for example, in classical epidemiology referred to as the incubation period. This incubation time in epidemic processes between the infection of an individual and the first symptoms or active transmission of the disease is frequently found to adhere to a log-normal distribution, for example in the cases of chicken pox, hepatitis, or salmonellosis [146, 150]. In an overview of incubation periods of 86 diseases, Nishiura concludes that 70.9% can be accepted as log-normal at a 5% level of significance [149].

Similar log-normal patterns have been discovered across domains for the equivalent of the observation time D_o. Barcelo [158] reports the channel holding time on cellular networks, the duration of calls and thus the time of the transmission process to be log-normally distributed.

This general notion of the log-normally distributed duration of human activities and associated spreading tasks are also repeatedly found across other domains, for example in the duration of strikes [155] or the time to complete tasks on a test [154]. This same pattern seems to extend into the online domain for spreading times, for example in the response times of viral marketing campaigns as reported by Iribarren and Moro [137].

The log-normal distribution regularly appears in the broader context of universal human activities and behavior. When tracing the mobility of cell phone handsets, Barcelo and Jordan [158] find that the time a person stays at a certain location (and is associated with a particular cell phone cell) also follows a log-normal distribution, which at near 100% penetration in many countries can be seen as a good proxy to measure human mobility. Radicchi, Fortunato and Castellano [162] for example report that the citation counts of academic articles universally follows a log-normal distribution when normalized by the relative citation habits in a scientific field. A similar phenomenon can also be observed in the law of proportionate effect [166], demonstrating that the publicly visible total number of recommendations changes the voting behavior of individuals in Digg and leads at a population-level to results best characterized by the log-normal distribution [163].

The SEIR (*S*usceptible - *E*xposed - *I*nfectious - *R*ecovered) model fits the process of content propagation in Twitter quite well. However the applicability of epidemic processes is only possible to a limited extend, due to the specific behavior of the Twitter messaging system: retweets are immediately sent out to all followers, and users can retweet a message only once. The measurements give further insights into the distribution of certain important factors, like the distribution of the infection time, the basic reproductive number and the number of infected/exposed individuals in an online social network. As the distribution of infection times is given through the convolution of two log-normal distributions standard Markovian differential equations used in basic epidemic models cannot be applied. Therefore, the simpler model of branching processes namely the Bellman-Harris branching process seems to be appropriate to model content propagation.

5.4. CHAPTER SUMMARY

In this chapter, diffusion cascades were analyzed under the assumption that epidemiological models may describe content propagation in OSNs. It is shown that within this terminology some messages are as infective as real-world viruses but will not reach a large fraction of all registered users. This fact may be explained by the limited activation ratio of the friendship network as explained in chapter 3.4, the fact that Twitter's friendship network is a directed one or the existence of communities defined through different languages or interests. Further research is needed to estimate the influence of these factors.

If every message is interpreted as an individual virus, tree-structures can be used to describe the propagation process because the "susceptible-exposed-infected-removed" (SEIR) model might describe the process. In this model users receive information and become exposed if a message arrived but is not read, infected once they read the message and if they decide to forward it to all of their peers they become removed. Interestingly the time a user needs to forward information follows a convolution of two log-normal distributions. The time an individual is exposed is called the observation time and the second period denotes the time a user is infected, called the reaction time. It is found that the observation time on average will always be longer than the reaction time.

6

Analysis of the Content of Online Social Networks

Apart from the analysis of attributes of individuals as mentioned in Chapter 2.1 the published content of messages in OSNs provide further insights into population statistics as well as content propagation. Chapter 2.1 explained the analysis of user data towards different interests and the similarities of egos to their alters based on profile information of users. This chapter will describe examples of how the content provided by users, mainly messages, can be analyzed using graph theory in order to estimate opinions and population level statistics.

6.1. Dutch Twitter Users Mobility Patterns

As described in Chapter 2.1.1 users within Twitter may attach their current GPS position to a tweet. To attach the current geographical position, a person just needs to enable the location services on the used mobile device and grant the messaging application rights to attach the current location on sent tweets. Once this setting is made and the device is able to estimate its location every message sent via Twitter's network will include the current location as GPS coordinates.

Other OSNs like Foursquare exist, which are mainly based on location information. In such services, users may register ("check in") to geographical places, shops or museums, called venues, if they visited the place. The number of times a person "checks in" to a place is counted, whereas people with the highest number of points will be awarded with titles like "mayor" of a place etc. One may create friendships in order to get informed about venues visited by friends or badges and rewards a peer receives. Foursquare itself had 45 million users (January 2014) occurring for over 5 billion check-ins per day [204]. Additionally, users may connect their accounts in other OSNs like Twitter or Facebook in order to publish their Foursquare check-in in these OSNs.

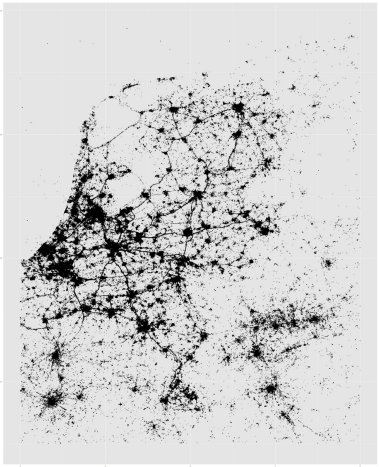


Figure 6.1: Positions of Twitter messages within the Netherlands.

In the used Twitter dataset only users of the service within the Netherlands were selected as explained in chapter 2.1, with 844,200 selected user accounts who accounted for 28.2 million messages including a GPS position within a period of one year. Out of these users and messages approximately 9 million messages refered to check-ins from Foursquare. Figure 6.1 depicts the position of all these messages, resembling the shape of the Netherlands including cities quite well.

If a person published his/her position more than once per day, one may assume that the user traveled between multiple locations. One may now interpret a location as a node in a graph where links are defined by mobility of users and two nodes are connected by a link if a person traveled, and tweeted, from the first towards the second position. The resulting graph will be directed and weighted as the link weight is defined by the number of individuals moving from one position to another one. This technique is similar to the one presented in Ratti *et al.* [205] where the graph of connected locations was based on phone calls. Ratti *et al.* analyzed a dataset of phone calls in Great Britain, connecting the locations of the origin of a phone call with its destination. In order to optimize the number of nodes in the graph the map of Great Britain was rasterized into a grid of 3,042 squares with a size of 9.5×9.5 km and links were estimated between these squares (denoting nodes) if a person called another one in a different square.

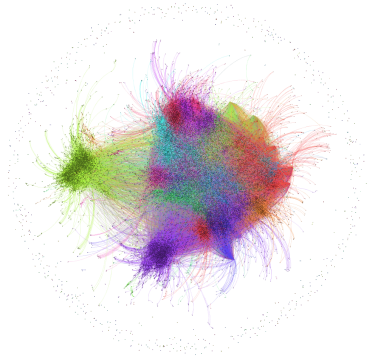


Figure 6.2: Graph of connected locations. Colors denote different communities.

In terms of the mobility of Dutch users the map of the Netherlands was also rasterized into squares of 7×7 km and the graph was constructed in the same way as in Ratti *et al.* [205] with the difference that self loops are allowed. This denotes that every node may have an edge towards itself if a person moved within the square. The resulting graph having 44,946 nodes and 308,071 links is depicted in Figure 6.2.

By using the community detection technique explained in Blondel *et al.* [73], one may estimate communities within this directed weighted graph. The partitioning having the highest modularity of 0.817 consisted of 706 communities. The high number of communities results from the fact that the graph contained 674 weakly connected components. Figure 6.2 depicts the discovered communities by different colors. The largest 26 communities contained 90.85 % of all nodes of the graph.

When mapping the nodal position back onto a map of the Netherlands, communities appear as regions which denote higher mobility inside than between these regions. Figure 6.3 depicts the different modular structures appearing when constructing the graphs from geo-locations observed during week days compared to weekends.

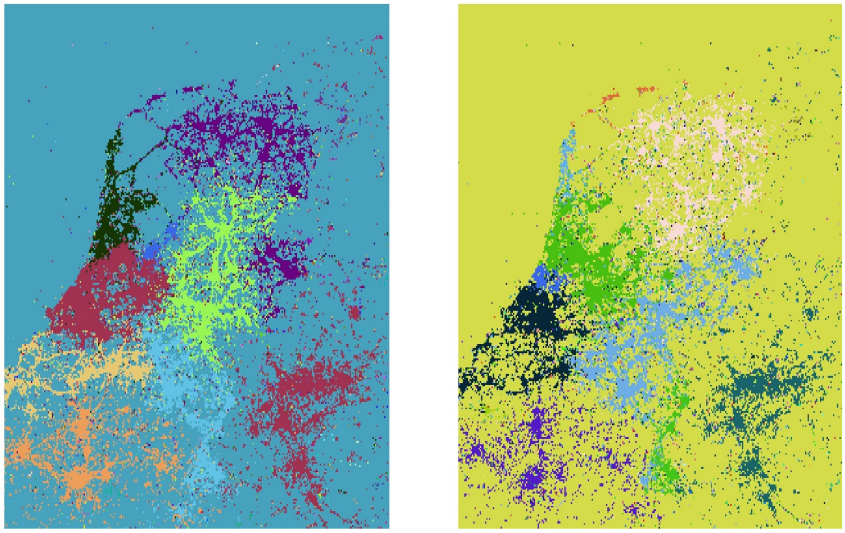


Figure 6.3: Map of the Netherlands depicting community structure during weekdays and weekends.

Given the sampled locations and mobility traces are reflecting real mobility patterns within a country, such analysis may help to improve urban mobility planning. However, one has to consider that sending messages within OSNs is often event driven, i.e users write about events they consider as interesting which denotes that complete traces of mobility are not possible to obtain by passively "listening" to data streams of OSNs. On the other hand, a high number of "status" updates sent within the OSN of Twitter.com are based on updates of the location of a user, which is why further research is needed to estimate the quality of inferred mobility traces.

6.2. SENTIMENT ANALYSIS

Research questions, such as the analysis of how individuals are influencing opinion formation in groups, require an automatic assessment of the sentiment of user statements, a challenging task further aggravated by the unique communication style used in online social networks.

Sentiment analysis, i.e. the extraction of an opinion's overall polarization and strength towards a particular subject matter, is a recent research direction [206, 207], and typically approached from a statistical, or machine-learning angle. Attention has been given particularly in the domain of movies [15, 208], by analysis of social media data, as reflection of common opinion. It is found that prices of the movies industry have a strong

correlation with observed outcome frequencies, and therefore they are considered as a good indicator of future outcomes. Most recently published work either perform unsupervised learning on a provided corpus of perceived positive and negative texts such as product reviews [209, 210], or use a set of curated keywords with positive or negative connotations to classify input [206, 211].

Another common approach [209, 210] is measuring sentence similarity between given data input and texts of specific polarity, which explores the hypothesis that opinion sentences will be more similar to other opinion sentences than to factual ones. Additionally, work of [212, 213] was focused on learning extraction patterns associated with objectivity (and subjectivity) in order to be used as features of objective/subjective classifiers. It is shown that this approach achieves higher recall and comparable precision than previous techniques. Apart from that, recent publications [214, 215] introduced the use of Natural Language Processing modules in order to extract concepts from the processed text and eventually derive sentiment out of them. In the very recent past, several of these general approaches have been specifically extended towards the mining of sentiments from online social media sources, in particular the microblogging platform Twitter [197]. In this study, the classification accuracy is compared with the following classifiers:

Twitter Sentiment The bulk classification service available on the Twitter Sentiment website [216] was used in order to classify a test-set. This tool attaches to each tweet a polarity value: 0 for negative, 4 for positive and 2 for neutral- therefore one may consider the first to describe subjective tweets, while neutral is for objective tweets. The main idea behind Twitter Sentiments approach is the use of emoticons as labels for the training data which is shown that it increases the accuracy of different machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM). It is noted that the web service of Twitter Sentiment uses a Maximum Entropy classifier.

Tweet Sentiments The test-set was also tested through the API of TweetSentiments [217], a well known tool for analyzing Twitter data which provides sentiment analysis on tweets. TweetSentiments is based on Support Vector Machines (SVM) and is using the LIBSVM library developed at Taiwan National University. It classifies tweets as positive, negative or neutral and these values are treated as stated previously.

Lingpipe The sentiment analysis tool of the LingPipe [218] package was used as well, which focuses on the subjective/objective (as well as positive/negative) sentence categorization especially on the movie-review domain. This approach uses the usual machine learning algorithms (Naive Bayes, Maximum Entropy, SVM) and a Java API of the classifier is available online. Even though it comes with its own training set, half of a hand-classified set was used to train the classifier in order to improve results. The other half was used as test-set and results were compared to the corresponding hand-classified tweets.

To evaluate the performance of established sentiment classifiers and create a benchmark for a newly developed solution, a set of some 1,000 publicly readable messages from the microblogging platform Twitter was randomly sampled. Prime use cases for sentiment analysis are for example research questions revolving around the spread of

information, opinion formation and identification of influential relationships in social networks, and such processes are typically believed to be present in discussions around product and media such as music, book or movie reviews.

For evaluation a data-set of 1,073 randomly chosen tweets related to the five most popular films of the 83rd Academy Awards in 2010 was collected. The language detection library of Cybozu Labs [219] was used, in order to eliminate tweets written in another language other than English, while advertising tweets were mainly manually removed from the set. Multiple retweets of the same text were also removed to prevent performance over- or underestimation, as well as unnecessary tokens like link urls, "@" tags for mentioning a user, 'RT' tags etc. Each tweet of this test-set was classified by hand before the begin of the evaluation into an objective or subjective statement; this corpus was used throughout this study as a reference benchmark.

Starting point for any sentiment analysis (as shown in Figure 6.4) is the detection of any form of opinion in written text. If the author expresses some form of judgment, the input can be considered a subjective statement, otherwise the data is classified as an objective claim. Example cases distinguished by such test are for example "I liked *The King's Speech*" versus, "*The King's Speech* was a really long movie", respectively. Another case is given by subjective messages where the sentiment is not based on the relevant reference point (the title of the movie). For example the tweet "I like you, even when watching *The King's Speech*" is a positive tweet, but its not the opinion about the movie that is positive. Once the existence of a sentiment has been established, typically a classification step is performed to determine whether the speaker is expressing a positive or negative opinion over a particular subject matter.

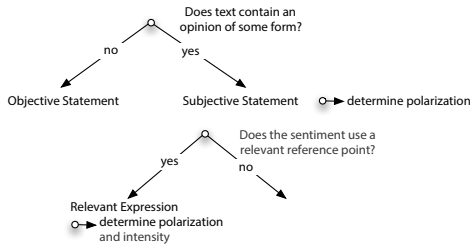


Figure 6.4: Sentiment classification involves a multi-stage process, in which not only the existence of a sentiment should be checked, but also the reference point and strength should be assessed. (Contributions of this work are indicated in gray.)

In many types of inputs, and specifically in micro-texts such as tweets or chats, however a problem arises: conversations are highly abbreviated. As tweets offer only 140 characters of payload, messages are reduced to a bare minimum and several different thoughts - for example reactions to previous incoming messages - frequently are abbreviated and intertwined: "Watched King's Speech today in class. I love the end of the term." This results in a very small footprint on which sentiment analysis can be conducted, at least compared to the essay- and article-type classification previously used for polarization analysis. Previously established approaches, which for example operate using a statistical word-frequency analysis, are therefore less suited, as the low quantities of text and the high concept compression ratios are resulting in very high statistical fluctuations and noise during the detection. For this reason, a different approach was pursued in this work and the grammatical structure of messages was analyzed. By detecting which concepts a particular sentiment is referencing to, one can make more fine-grained decisions

and consequently achieve a higher prediction accuracy especially in dense, intertwined texts. In the example above this concept is the end of the term (and therefore potentially a looser schedule) rather than the movie itself.

Finally, many applications and research hypotheses can be better served if not only the existence of a sentiment and its general polarization is known, but an absolute notion of "how positive" or negative a particular opinion can be derived. This would allow a better assessment of how opinions are propagated and adopted, as a person with a strong negative attitude towards a particular concept is first expected to become less and less negative before developing a positive sentiment if at all. Without a quantification of polarization such trends would go unnoticed. Additionally, a quantitative measurement of attitude would allow of differentiation between alternatives, which in sum are all considered positively, but in a pairwise comparison are not equal.

If considered in previous work, this aspect is typically approached using manually curated word lists, as for example in [207]. Following this strategy in the application context of micro-messages however discovered two fundamental difficulties: 1) Users utilize a rich set of vocabulary to describe their opinions about concepts. Capturing and maintaining an accurate ranking of evaluative comments would require a significant effort in a practical setting. 2) Expressions indicating positive and negative sentiments and their relative differences are neither stable over time nor between different people, thus a method to re-adjust and "normalize" sentiment baselines over time or between say generally very positively oriented, neutral or pessimistic speakers will provide an advantage.

To evaluate the performance of existing sentiment classifiers, the set of available classifiers was used to analyze and distinguish a reference body of tweets. As most methods do not allow for a sentiment quantification, this evaluation was limited to only a general polarization detection which is supported by all systems. Comparing the output against the previous human classification, the overall accuracy of the automatic classifiers in distinguishing subjective from objective statements was measured, as shown in Figure 6.5(a). Figure 6.5(b) shows the overall performance in correctly and incorrectly classified statements.

As can be seen in the figure, the classification accuracy of all statistical sentiment analyzers is between 55 and 60%, whereas the proposed statistical-grammatical hybrid approach yields a correct classification accuracy of about 85%, a 40% gain over previous work. Note also that the accuracy of existing system also varies significantly between the type of input data: Twitter Sentiment [216] for example is much stronger identifying objective statements compared to subjective ones, while Tweet Sentiment [217] shows exactly the opposite behavior. The proposed hybrid solution on the other hand does not show any significant bias.

As shown in Figure 6.4, the general task of sentiment analysis can be conducted in two general phases, first the detection of opinions in general (yielding to a categorization in objective and subjective text), after which a quantification of the polarity can be attempted. The following discussion mirrors these steps.

6.2.1. GRAMMATICAL SENTIMENT CLASSIFICATION

In order to classify a given message into subjective and objective we analyze the grammatical structure of a tweet. Subjectivity is mostly based on adjectives or verbs express-

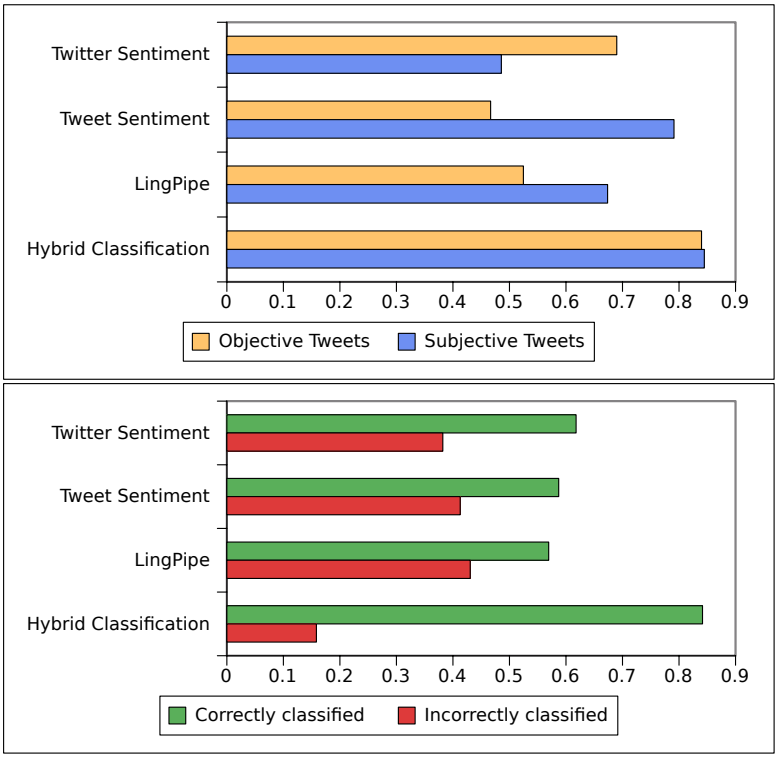


Figure 6.5: Classification accuracy of different sentiment analyzation methods.

ing the polarity related to the subject of the message. This means if directly expressing an emotion one usually uses a verb like “like/love/hate" whereas expressing the mood about something usually contains an adjective. Consider the examples: *I’m feeling sick today*, *I liked the movie*.

To determine the existence of a sentiment, the grammatical structure of a given text was investigated to detect the presence of verbs carrying an emotional meaning or to find the adjectives associated with the keywords of interest, which are used to perform the sentiment analysis on, in this scenario the titles of movies. Grammatical structure analysis is a mature research area, and in in this Klein and Manning’s lexicographical parser [220] was used to determine the structure of English texts of tweets. For a given text, this tool is estimating the grammatical structure. An example for the tweet “I liked the movie" is given in the following:

```
[I, liked, the, movie]
(ROOT
  (S
    (NP (PRP I))
    (VP (VBD liked)
```

```
      (NP (DT the) (NN movie)))))

nsubj(liked-2, I-1)
det(movie-4, the-3)
dobj(liked-2, movie-4)
```

Here, the parser is reasoning that "I" is the nominal subject of "liked", and "movie" is the direct object of the verb "liked". As mentioned, most tweets expressing a sentiment have this kind of structure. If an adjective is referring to a subject the likelihood is quite high that this tweet expresses the mood about something. Note however that it is in general possible that the speaker was using sarcasm or irony, which could not be detected from a grammatical viewpoint.

In a second step, one needs to cross-check whether the word referring to the subject of a tweet is an adjective. This can be done using either a lexical database such as WordNet [221] or a part-of-speech Tagger [222], which will be used further in this discussion. Through such a tool, every word in a given sentence can be annotated with a tag identifying its purpose in the sentence [223], so the example "I liked the movie" is marked as:

```
I/PRP liked/VBD the/DT movie/NN
```


The part of speech tagger tells that "I" is a personal pronoun, "liked" a verb in past tense, "the" a determiner and "movie" a noun. By connecting the so gathered information of a message, simple rules were built to detect if a message is a subjective statement whereas all the others by inversion have to be objective:

1. if an adjective is referring to the subject
2. if an verb out of a list is referring to the subject
3. adjective + [movie, film]
4. [movie, film] is/looks [adjective]
5. love/hate + [movie, film].

6.2.2. AUTOMATIC POLARITY ESTIMATION

After the existence of a subjective component has been established, it is necessary to determine the overall polarity of the sentiment and if possible also the magnitude of the sentiment. In order to estimate the general polarity direction of words in the corpus, an unsupervised approach was used, based on word correlations. This approach is inspired by the way a person is learning to judge which words have a positive or negative meaning, which is essentially a result of a lot of exposure to speech and written text, from which the learner infers which words appear in a positive or negative context.

The same basic principle, inferring which words appear together in a positive or negative context, can however be easily mirrored in a machine as well. Here, a computer would simply need to count how often a particular adjective has been encountered with

a positive meaning compared to the frequency it has been observed with a negative connotation. To begin such an automatic classification, some notion of what is deemed positive or negative will be necessary. In this work, two general options were explored. First by manually specifying a set of keywords one would associate with positive expressions such as “fantastic”, “incredible”, “amazing”, which can read from existing databases such as [221], and second by looking at the most basic positive/negative expression commonly used in online messages such as chats, emails or microblogs: a positive smiley :-) and a negative smiley :-(.

From a list of one million tweets, all tweets containing a positive smiley :-), :) or =) were found – in the following referred to as *positive keywords* – and a list of texts containing at least one of those symbols was created. Similarly, a list of all tweets containing a negative smiley such as :-(, :(or =(– deemed *negative keywords* – was prepared. Using the techniques discussed above, all individual statements were dissected and the number of co-occurrences between every detected word and the positive or negative keywords was counted. To correct for differences in length of those two lists – as users typically write more positive than negative statements –, the two values are then normalized by the number of words in the list of messages containing positive words and the list of messages containing negative words. This results in a relative assessment of a particular word to appear in a positive or a negative context, where context is defined by the positive and negative keywords, respectively.

To arrive at a relative polarity of a particular word between the two extremes “positive" and “negative", it is now simply enough to subtract the relative frequencies. This number is positive if the word is typically used within a positive context and negative if the word typically occurs with a negative meaning. As this number is biased by the number how often a word is used in general, a final correction step is executed in which each rating is multiplied by the term frequency measured in all tweets: the emphasis of frequently observed words is therefore reduced, and the value of unusual ones is lifted.

Figure 6.6 shows the output of this simple procedure conducted over a body of one million tweets, and using just positive and negative smileys as corresponding positive and negative keywords. As can be seen, this unsupervised process leads to a clear and meaningful ranking of adjectives. This analysis was repeated over a selection of data sets of different duration and verified the automatically generated polarity estimation against those done by a human. Typically, less than 7% of the words were considered wrongly placed in the overall order; out of a list of 30 adjectives between one and two placement were deemed higher or lower in the ranking by a human observer than by a machine. As this method can be executed without manual intervention, this new methodology can be continuously conducted to detect the general development of sentiments in entire online communities, as well as to identify whether the polarity of certain words shift in strength over time.

A more comprehensive analysis is needed to see if the sentiment of adjectives changes over time. However taking Twitter data of durations of one week, one and two months the polarity seems to stabilize the longer the duration of observation.

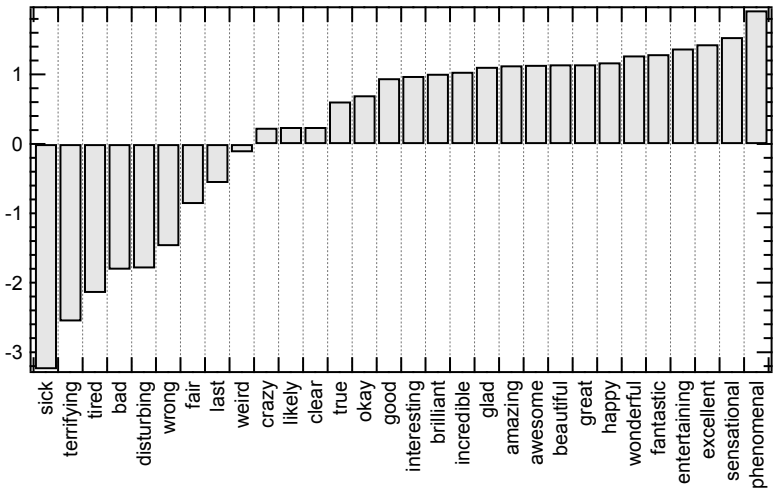


Figure 6.6: Automatic polarity ranking of adjectives based on general Twitter messages.

6.2.3. DETECTING NETWORKS OF CONCEPTS

This general method is however not limited to determine the polarity strength of words in general, but can be used to detect and identify common concepts and their associated sentiments in general. To do so, it is simply necessary to swap out the two sets of keywords (which in the last section were *:-), :), =)* and *:-(, :(, =(*, respectively), and replace them with those terms and synonyms relevant to a particular study.

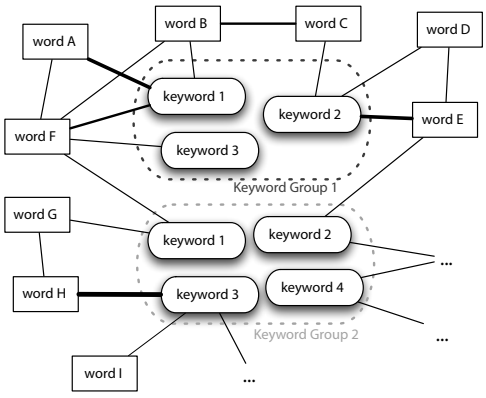


Figure 6.7: The technique can be broadened to determine the concepts commonly associated with any keyword as well as the particular strength of the association.

Consider for example a situation where one would to determine the associations made with the brands and products of two hypothetical tea manufacturers: *McArrow's orange-peppermint tea* and *DrBrew's strawberry-melon tea*. Here, one would populate *keyword group 1* with words from the first area, i.e. McArrows, orange-peppermint, etc. and analogously *keyword group 2* with words such as DrBrew, strawberry-melon, etc., and by the same means described above, this method would derive the set of words frequently used in combination with any of those keyword terms as well as the strength of their typical common appearance as shown in Figure 6.7.

The words $A - I$ discovered to be co-

located can however be themselves further interpreted, for example, 1) depending on how positively/negatively they are (as discussed above), or 2) which general topic areas or word field the concepts come from. Imagine for example words *A* and *F* in Figure 6.7 to be "taste" and "flavor", while words *D* and *E* are "packaging" and "price". Clearly, such combined word co-localization, polarization and word-field analysis will provide a significant insight to our hypothetical tea manufacturer, which can also be easily repeated over time to track its overall development, but also to the researcher interested in how particular opinions form, are spread and change over time.

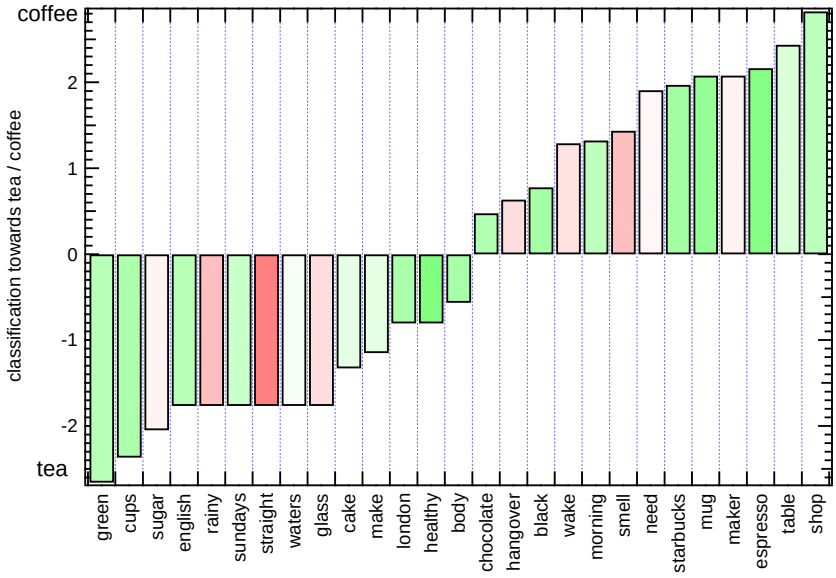


Figure 6.8: Polarization analysis of commonly used words with coffee and tea based on general Twitter messages.

As a general example, this techniques was applied towards the keywords "coffee" and "tea", and the system arranged the resulting associated words according to their overall polarity strength. The overall abstract network can then be drawn in a similar manner as Figure 6.6, and Figure 6.8 shows the 27 strongest connections commonly made with the terms coffee and tea. All bars indicating the affiliation towards the two beverages are colored by their polarity value. Red indicates a negative, green a positive and white a neutral polarity.

The presented method describes an alternative to determine the existence and strength of subjective opinions in short colloquial text, an application domain where existing approaches do not yield a high detection accuracy. The proposed system works through a combination of grammatical analysis with traditional word frequency analysis, does not need supervised training and improves the accuracy of previous work by about 40%.

6.3. CHAPTER SUMMARY

In this chapter the content, usually messages, propagating through an OSN is analyzed where it is shown that the identification of mobility patterns is possible. When applying a graph topological approach to these patterns it becomes possible to estimate regions of mobility. These findings might enable better planning of public transportation or network planning for telecom operators.

The sentiment of a message plays an important role in content propagation as measurements showed that positive messages propagate faster than negative ones and are more often forwarded compared to negative ones. A unsupervised method of estimating the sentiment of short textual messages is proposed in Chapter 6.2, which outperforms existing solutions by being applicable to multiple languages as well as enabling the estimation of a polarity value, i.e. a quantization of how positive or negative a concept is perceived by the users of OSNs. This technique, also allows the estimation of networks of concepts by creating graphs of connected words. Further research in terms of an automatic classification of products or brands and the estimation of how these concepts are perceived in the environment of an OSN are possible through this technique.

7

Conclusion

This thesis summarized work on online social network (OSN) analysis in the context of content propagation. Content in OSNs, typically given by short texts or messages originating by a user is propagating if friends or followers of the originator forward the information to their own friends or followers whose peers might follow suit and distribute information further. The network based upon friendship relations defines the distribution channel for emails, short messages, notifications but also for opinions, recommendations, images and videos. That is also the reason that the term viral spreading is often used in conjunction with content distribution in OSNs because the information relates to a virus, distributed via contacts between individuals whereas the messages once spreading might not be stopped by the originator.

7.1. Main Contributions

Due to the high number of registered users in current OSNs and the ever-growing amount of information publicly shared by them within the services, predictions solely based on information in OSNs are increasing in number and frequency. The second chapter of this thesis shows that sampling users from an OSN in order to obtain a random subset of OSN users from the population of a country is only possible to a limited extent. In terms of the distribution of family names, first names and the location of users a subset can be found, but in terms of the age of all users the sample will be biased. For other important factors that should be addressed in a random sample, like the profession, income, size of the household or education, to name a few, it's not possible to directly obtain information from an OSN like Twitter, which states a problem, if data from such services is used in order to predict future events. It is also shown that different widely-used methods of obtaining data from OSNs are not able to provide good samples of inhabitants of a country.

Using data gathered from OSNs enables researchers to analyze commonalities, differences and interests of individuals as shown in chapter 2.1. Recommendation and feedback systems benefit from available information, but at the same time privacy concerns arise. It is shown that in terms of an individuals privacy the common approach of

offering tools to a user to hide his information is not sufficient because of similarities between friends in general referred to as "homophily". Friends have similar interests, age, hometown etc. which enables attackers to infer private attributes of users from public information shared by their friends. Even further, the number of friends with a public profile does not necessarily need to be high, as a small number of friends with public viewable profiles, compared to the average number of friends in the OSN of Hyves.nl, was enough to reconstruct up to 86% of the hidden attributes of a user. The sheer knowledge that some individuals are befriended with the "targeted" person is therefore enough which leads to the problem, that the privacy of individuals who are not even registered to OSNs is affected. On the other hand, when interpreting the strength of a friendship as the number of common attributes, the spread of opinions, diseases even habits or feelings can be modeled because close friends will use their friendship relation more often. Such analysis, despite the negative connotation, will improve recommendation systems dramatically.

The connections of users to direct neighbors, also called alters, and the links between alters define so called ego-centric networks. These typically rather small networks represent the OSN from the perspective of an ego. Therefore one may analyze personal and group communication of an ego, asking the question if a person is part of a large group or connecting groups by being positioned "in between" them. The theory of "structural holes" by Burt [49] assumes that users "bridging" groups are more effective and independent because of their position in the ego-centric network. Structural holes refer to "missing links" between groups which "strengthen" the position of an ego. In graph theoretical terms an ego "bridging" groups, would have a high betweenness which indicates that most of the communication will pass through the node who may decide to alter or not to forward the content. Therefore it is more "powerful" than other nodes. When analyzing communities within the graph of alters without the ego, further insights about the groups, the ego is part of, can be estimated. In this way one is able to identify overlapping communities in higher resolution than before.

Obtaining topological data from an OSN might be a difficult task, as standard crawling techniques like breath or depth first search obscure intermediate results in terms of network metrics. Nearly 70% of a social network need to be crawled in order to estimate the final value of most common network metrics as shown in chapter 3.1. By changing the traversal to follow more "human" like discovery patterns the crawling procedure converges faster, at least for certain metrics. The more "human" like procedure, called *Mutual Friend Crawling* is based on a reference score which directs the crawling procedure to stay inside communities of users by estimating the next user profile to visit based on the number of references already discovered. This reflects the way individuals would employ social search, namely by asking friends if they know some person with certain qualification and picking the one most often mentioned. An interesting side effect of this crawling technique is given by the fact that communities can be detected during the procedure as explained in chapter 3.2.

Once a large fraction of the network of an OSN is obtained one may search for influential users which are commonly assumed to have a high centrality and high betweenness, measures basically stating that users which are embedded in the core of the network have higher abilities to reach out to a high number of others. Apparently for em-

pirical data sets like the friendship network of Digg.com it was not possible to identify individual influential users based on the structure of the friendship network as no relationship was found between the link betweenness and the actual amount of content propagating along these links. Other metrics such as coreness or eigenvector centrality did not show significant relationship to the propagation of information either. However, additional information apart from measures of topology of the friendship network is relevant for successful spreading of messages, such as the timely alignment of friends, denoting that friends or followers of a spreader should be online at the same time or the impact of influential groups, i.e. groups of users that coordinate their voting or spreading behavior, which are typically well embedded into the network, undiscoverable through standard topological metrics as shown in chapter 3.4.

In terms of the evolution of OSNs, when new users register to the service or create links, online social networks are increasing in terms of their link density and the amount of communication increases. This process leads to the fact that OSNs might experience a saturation effect in terms of the number of registered users, whereas the number of links, given the users are actively creating relations, is increasing. Obviously this process will lead to a higher average degree of registered users. Additionally the perception of certain users that a high number of followers or friends denotes some measure of influence led to the existence of services that sell friendship relations. Currently[1] reports show that on Twitter, most celebrities' set of followers contain quite a high number of fake users [224]. As the whole process of "buying" followers is rather negatively connotated as it states an act of artificially increasing measures of influence, it actually also increases the reach of these users. In a naturally grown system, only users which are interested in particular persons would chose to follow them, whereas a politician for example, does not need to convince individuals who are already interested in him or her but the ones which are not. One of the easiest ways to connect to the ones who are not is through utilization of follower markets.

As initially stated, once content spreads through the network of an OSN the spreading procedure is often referred to as viral spreading because individuals who create or receive information "infect" their followers when forwarding content to them, who might chose to forward again. A procedure similar to epidemics because if every user infects 1 or more of his followers to forward the content again, the whole population of users will receive the information in a short time. Luckily as shown in chapter 5, in real-world measurements of viral spreading most information proved to be less infective, because otherwise, given the high number of sent messages, every user of Twitter would be overloaded with information. Further on it is shown that the "SEIR" epidemic model might capture the spreading procedure because susceptible users become exposed once they receive a message, turning into infective, after they read the message and will be removed from the process after spreading the message on. However the commonly assumed exponentially distributed infection times can not be confirmed in empirical data. The distribution of the duration a follower is exposed until getting infected, called the observation time, and the distribution of the durations it takes until the content get forwarded again, called the reaction time are both log-normal distributed where the observation time is on average longer than the reaction time.

[1] March 2014

When analyzing content propagation only in terms of users which forward information, tree structures can be used to capture the spreading dynamics. Where it was found that the fraction of friends who will forward information is nearly a constant around 1.8%. Also the number of potential receivers of information was found to be way smaller than the size of the network in terms of the number of nodes. A fact that might be based on the directionality of the friendship graph and clusters of users formed by different languages, religions or interests or a timely dependence.

7.2. FUTURE WORK

The analysis of online social networks might enable one to understand basics of the most important factor of most systems, the individual or groups of individuals and the relationship between them. However the famous statement "correlation does not imply causation" should always be kept in mind when drawing conclusions from empirical data. Considering Occam's razor ("Pluralitas non est ponenda sine necessitate", i.e., "Plurality is not to be posited without necessity") stating that among multiple hypotheses, the simplest one should be selected, this statement should not mistakenly be interpreted in the wrong way leading to the assumption that correlation implies causation.

This rather strong introduction to the future work section is based on the observation that multiple publications exist in which correlation is mistakenly interpreted as causation. Predicting the outcome of future events like elections or the stock marked based on data from social media seemed to be possible, whereas the results showed in this thesis depict that the used sample of data probably does not represent large fractions of a population. Future work should therefore focus on finding causation, trying to explain why certain predictions were correct.

When analyzing data from online social networks one does not necessarily observe the opinion of a high number of users, but also the attempts of influential individuals trying to form opinions. Therefore the problem becomes two-fold. Given one observes the opinion of users, it might be possible to predict events, whereas every prediction might on it's own influence the future event and the result. On the other hand, when observing mainly data from individuals trying to influence users, one could also predict the outcome by merely quantifying the success rate of influential users. To which extent these two groups are represented within the landscape of social media has to be quantized.

In terms of users' privacy, techniques and methods are needed to protect individuals from possible attackers by still being able to build useful recommendation systems, a task that seems to be difficult in the first place, but on a second thought, a recommendation system does possibly not need to know "everything" about a user as a system that is too effective would only recommend products, apps or friendships to users that would buy the product or connect to the proposed users anyway. It is also possible that trustful third-parties could be employed in between users and products in order to ensure a maximum amount of privacy. Such services could be operated by telecom operators for example, which are already "in between" the user and social media, maintaining the distribution channel while having privacy regulations. Given the amount of data, generated by social media, efficient systems to store and effective algorithms to analyze "Big Data" are needed. Another question concerning these huge databases is, if one really needs to store all information or is it possible to invent systems, that forget about outdated in-

formation. Most individuals seem to be concerned that content once published in the Internet will possibly never be removed again. Especially private information should have an “best-before date”. To this extent, a content-centric like approach, as proposed for the Internet, in which content may expire could define a promising path.

In terms of the detection of influential users it is shown that common topological metrics are not sufficient. Therefore the influence of overlapping communities, timely alignment of peers, the amount of communication and the content propagating through an OSN should be taken into consideration. Widely used topological metrics like betweenness and centrality metrics only consider the position of a user within the complete topology of a network, whereas a user who is not identified as influential by using node betweenness might indeed be influential in it’s own neighborhood. Such discrepancies should be considered and analyzed in further research.

The aforementioned need for influence metrics also arises from the analysis of content propagation. Whether propagation as measured in online social networks reflects the process of real-life propagation needs to be estimated because if messages, opinions, feelings or diseases propagate in a similar manner in the physical world, one would be able to immunize the right persons in order to protect a population from hazardous viruses. From the measurements in this thesis the classical epidemiological models could benefit because if the measured heavy tailed (log-normal) distribution models a real-life virus spread then the virus might survive and spread longer. Within online social networks the limited reach of content needs further attention, which might be caused due to the directionality, different communities (language or interest based) or aging effects of the content.

A

APPENDIX

A.1. DATA SETS

> You can know the name of a bird in all the languages of the world,
> but when you're finished,
> you'll know absolutely nothing whatever about the bird.
> So let's look at the bird and see what it's doing
> – that's what counts.
> I learned very early the difference between knowing the name of something
> and knowing something.
>
> Richard Feynman (Physicist (1918 - 1988) Nobel Prize in Physics in 1965)

As the quote of Richard Feynman states, one can know the name or the concept of something in all languages in the world but one may not know something about the subject at all. This concept holds not only for physical systems or real life objects, but also for data sets used in scientific research. Understanding processes behind data sets of online social networks always needs a preliminary understanding of the OSN and, sometimes more important, the process of how the data was gathered. The crucial part of any analysis is to understand and describe the used data sets completely in order to keep up with "statistical laws" like the randomness of samples, possible influence of different processes etc. (i.e. one should not try to find prove of a certain assumptions in a dataset, as one may always find, after applying enough "'filters"' and "'cleaning"', what one is looking for. Understanding what happens "'between the lines"' and causation is often more important.). This section will describe the data sets used in this thesis.

A.1.1. ARXIV COAUTHORSHIP NETWORK

The ArXiv co-authorship data set containing publications of the subjects of "General Relativity and Quantum Cosmology" and "High Energy Physics - Theory" in the period from January 1993 to April 2003, collected by Leskovec *et al.*[90]. The data set contains 29,555 papers defining nodes of the created co-authorship graph. Nodes are connected by a directed edge, if a paper cites another one leading to 352,807 links. Multiple other ways

of creating graphs out of such a data set are possible. A bipartite graph describes which authors participated in the created in the creation of a publication. Therefore publications are one type of nodes whereas individuals are another type. Edges in a bipartite graph only connect nodes of different type. Two projections of such a graph into simple graphs are possible whereas on one hand, used in this thesis, if authors are nodes in the simple graph who are connected by undirected edges if two individuals contributed to a publication. On the other hand if publications should be the nodes of the desired graph. Two publication would be connected if one author coauthored both texts.

A.1.2. DEVIANTART

DeviantArt operates an online service for artwork created by its users. Launched in August 2000 it became the largest internet art platform and the 13th largest online social network with more than 31 million user accounts and 283 million submissions [225]. Users upload mostly images, but also poems, short movies, animations and photos, which are shown in an section called “newest”. The community of registered users votes, and comments on uploaded artwork, which if it received enough attention will be displayed on sections called popular, the main section of focus for most deviant users.

Users may befriend or watch other’s activity, whereas befriending denotes that the user will be informed about all activity of the friend while watching another user implies that the user is only informed about new submissions of the “watched” person.

The data set of deviantArt used in this Thesis was obtained through multiple perspectives defined by sections of the service. As already mentioned, every time a user uploads an image the submission will appear on the “newest” section of the service which was continuously observed for a period of 6 months. The found usernames were then used as seed-points for a breadth first search following both, the friendship relations as well as the “watchers”-relationships. The total number of gathered friendship relations is: 225,456,955 and for the “watchers”-relationship: 266,125,047 based on 11,005,894 profiles. Every user profile contains the username, real name, the gender and location, the registration date, the number of submissions, comments, pageviews, the birthday as well as categories for favorite interests. Users are also free to join or create certain groups. The number of groups in the used data set is 246,787.

A.1.3. DIGG

The news portal “Digg.com" is a social content website founded in 2004, which according to the rating provided by Alexa Internet, Inc. [125] in 2010 belonged to the top 120 most trafficked websites in the Internet. A total of 2.2 million users were registered on the webpage, submitting between 15,000 to 26,000 stories per day to the system. Out of those submitted stories, approximately 180 stories per day were voted to become “popular”. The collected corpus contained the activities of users and the content of more than 10 million stories in total, 200 000 of which achieved critical mass.

Within a social media aggregator such as Digg.com, registered users are able to participate by submitting, commenting and voting on content they like or dislike. Users can send in news or blog articles, images and videos by submitting a link to the web page where the information can be found, together with a title and brief description of the media item. Entries in Digg are categorized in 10 main topics (Business, Entertainment,

Gaming, Lifestyle, Offbeat, Politics, Science, Sports, Technology, World News), each further divided into a total of about 50 special interest areas. Registered users and visitors to the site can browse the collection for example by category, submission time or through a recommendation engine, thus, Digg also acts as a online social bookmarking site.

New submissions to the system are enqueued in a special section of the web site called “upcoming”, where entries are staying for a maximum of 24 hours. If an item generates enough attention and positive recommending votes, an activity called “digging”, within this time period, the story is tagged as “popular" and “promoted” to the “front page”, which is the main home page immediately visible to anyone navigating to the Digg.com website. Thus, once promoted to the front pages, a story generates a lot of attention and traffic from registered users and casual visitors. The concentrated, sudden instream of users following the link from a promoted story is often so large to frequently overload remote web servers, referred to in the community as “the digg effect”.

On the Digg website, users also engage directly with each other and can create friendship connections to other users in the network. These connections can either be one-directional or two-directional, in which case the user is either a fan or a confirmed mutual friend with another person. Fans and friends are notified by the friends interface of digg if their contact has “digged" or submitted a story. It should be noted at this point that the semantics of a friend in Digg (obtaining information) are certainly different from a friendship in Facebook (personal acquaintance) or LinkedIn (business contact) [36], as also the main function differs between these social networks.

While most social network traces are crawled using friendship relations, e.g. [107] and [226], the Digg dataset was obtained by a simultaneous exploration of the network from four different perspectives, as shown in figure A.1. By using the Digg Application Programming Interface (API) and direct querying of the website, it as possible to explore the aforementioned four perspectives (from bottom to top in Figure A.1) during data collection:

- **Site perspective**: The Digg website lists popular and upcoming stories in different topic areas. Every hour, the frontpages with all popular stories (for all topics) that are listed on Digg were retrieved. Every four hours, all upcoming stories (for all topics) are collected. All discovered stories are added to an “all-known story” list.
- **Story perspective:** For each story that has been retrieved, a complete list of all activities performed by different users (who digged on the story) is collected. Any user that is discovered will be added to the “all-known user” list for future exploration.
- **User Perspective:** For each user discovered within the Digg OSN, the list of their activities, such as submitting and digging on stories, is retrieved. Occasionally, a previously unknown story is discovered (this is typically the case for older stories before the collection started). For such a story, the entire (digging) activities of users are retrieved for that story.
- **Social Network Perspective**: Each registered user can make friends with other Digg users. In the crawling process, a list of friends is retrieved for every user. If a friend is a previously unknown user, this user is added to the data discovery

process, and a list of all his/her friends and his/her public user profile information are retrieved. This procedure is continued in a breath-first search (BFS) manner until no new user can be found. The process is periodically repeated afterwards to discover new friendship relations that have been formed after the last crawling pass through the data.

By using the above crawling methodology, it was possible to collect nearly the entire information about friendships and activities of users and the published content in the Digg network. This is a significant and important distinction as traditional crawling techniques exploring a social network based on the friendship graph will only discover those users which are engaging in active community building and are also part of the (giant) connected component of the social graph. By exploring all four dimensions simultaneously, the used data collection was able to identify any user that was either (a) digging or commenting on a story, (b) submitting a story, or (c) made at least one friendship with any other user (even outside the connected component).

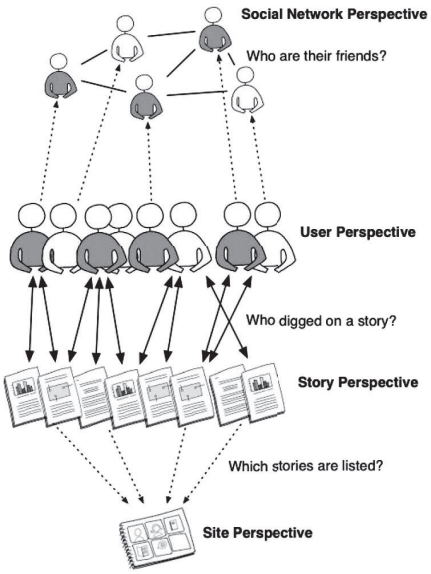


Figure A.1: The four components of the Digg crawling process.

A.1.4. ENRON

The Enron data set contains nearly all e-mails (619,446 messages in total) sent and received by 158 employees of the company Enron originally published by the Federal Energy Regulatory Commission collected by the CALO Project (A Cognitive Assistant that Learns and Organizes) described by Klimt and Yang [227]. The data set includes all attributes of an email like the text, subject, recipients, time the message was sent etc.

A.1.5. HYVES

Hyves.nl used to be the largest Dutch Online Social Network founded in 2004 containing nearly 10.6 million accounts in 2011.
Given the total population of the Netherlands (ca. 16.5 million), a large fraction of the inhabitants are registered. However, the total number of user accounts of Hyves.nl includes duplicates and orphan accounts as well as commercial pages. In December 2013 Hyves.nl changed from being a online social networking service and became a online gaming platform.

The dataset used in this thesis was obtained by screen scraping Hyves.nl using multiple parallel breadth first searches.It contains 2,971,261 user profiles. Out of those roughly

one third are public viewable profiles. On a profile page, users have to define a username and a real name and they may provide birthday, age, hometown, relationship status, living situation, address, phone number and their email address.

Additionally, users may join a large selection of groups. Those groups could be real world communities like sport clubs, schools or companies, famous people, bars and restaurants, books, movies etc. Groups are organized in 19 topics namely: brands, hangouts, school, college, club, company, TV shows, books, food, film, gadgets, games, famous people, media, music, traveling, sport, TV programs and others. It is possible to join groups without invitation and a user may create a new group. Groups are displayed on the user's profile page ordered by their topic.

Every group has its own page, listing all members of this group, additional information like events, addresses or opening times. Friendship relations are set up by sending a friendship request via Hyves.nl to a user. If the request is accepted the two users are mutual friends. The average number of friends in our dataset equals 127. Users may also upload photos and tag people in these photos. Also, 446,868 images and people tagged in them were crawled, resulting in 624,478 user names 1,311,423 relations.

In terms of privacy control, Hyves.nl allows a user to change privacy settings to display each attribute to the public, viewable for everyone registered at Hyves.nl, friends of friends or only friends. Nearly one third of all profiles, are publicly viewable, which means that the real name, groups, age, hometown and the list of friends is displayed. If a user has a private profile page, the real name, if entered by the user, is still displayed.

A.1.6. MOVIE ACTOR NETWORK

Internet Movie Database contains movies, series and video productions as well as information about involved persons. It states the largest (free available) database of movies, reviews and critiques. The data of IMDB movie actors collaboration network contains 127,823 movies, 392,340 nodes (actors) and 13,738,786 links, obtained directly from the web-page of IMDB.com (`http://www.imdb.com/interfaces`). The initial graph, stated as bipartite graph in which actors are nodes of type one and movies are nodes of type two, whereas links only connect nodes of different type was projected in order to create a simple graph of actors, who are connected if they participated in the making of the same movie.

A.1.7. SOURCEFORGE

SourceForge offers a web-based project repository assisting programmers to develop and distribute open source software projects. SourceForge facilitates developers by providing a centralized storage and tools to manage the projects. Each project has multiple developers. The hyper-graph used in this thesis was constructed by completely scraping all project pages of SourceForge.com, taking software projects as nodes and the developers as hyper-edges. A hyper-edge is incident to a node if the corresponding developer participates in the corresponding software project. The used SourceForge software collaboration network has 259,252 software projects and 161,653 developers.

A.1.8. RATEBEER

Ratebeer.com is a webservice designed to rate and share opinions on beers. Registered users are requested to rate on beers in the categories: aroma, appearance, taste and palate. Additionally a written rating in form of a text is needed. Beers are categorized by the corresponding brewer, the style (92 different ones like stout, cider, ale, pilsener etc.), the location of the brewery, calories, alcohol by volume and the usual way how it is served (like English pint, dimpled mug, flute, tulip etc.). Additionally the users may rate stores, bars, breweries or restaurants where they bought the beer. The dataset used in this thesis contains: 4,465,714 beer ratings of 191,542 beers by 254,431 users and 147,296 place ratings of 24,198 places.

A.1.9. TWITTER

Twitter.com is the operator of a "microblogging" service created in 2006. Registered users write short messages of up to 140 characters, called "tweets", which appear on the user's personal status page in chronological order. These tweets are, by default, visible to the general public, unless a user has marked his profile as private, and can be retrieved when performing a general search query on Twitter that matches a particular tweet. According to Alexa.com [125], Twitter is the 8th most visited webpage worldwide as of December 2012. Per day an average number of 140 million messages are written by the 241 million users registered on the microblog [197].

The friendship graph of Twitter is based on follower relationships which are created if a user chooses to connect to another one. Unlike in other OSNs the befriended user does not need to approve this friendship request. Once befriended, all status updates of a friend are shown in the interface of the user in chronological order. This interface also lists replies to a message, the number of its retweets (i.e., the number of users that forwarded the message to their respective followers), as well as a favorites counter indicating how many users marked the tweet as one they like. Messages only contain text but can be annotated by location information or links.

The "sample stream" endpoint of Twitters API allows to receive a random sample of 1% of all messages written within Twitter, leading to an average of 17 tweets per second. It is possible through searching to find every tweet written in Twitter, except for messages marked as private.

Every tweet contains all information about the user who wrote the message, like the userid, username, real name, location, a description, the number of friends and followers, the number of sent messages, the timezone of the user, the account registration date and time, and profile settings like the background colors and images. Additionally, information about the status (message) is provided, like the messageid, username of the author, the text, the time the message was sent, if it is a reply or a retweet, i.e a forwarded message, to or of another message and if so, all information about the replied or forwarded tweet and the GPS location of the user when writing the message. The location of a message only contains coordinates, if the user enabled the GPS hardware on the used device or provided a location himself.

The data set used in this thesis was collected through continuously listening to the "samplestream" for a period of 3 years from 2011 to 2014 and by querying other endpoints of the Twitter API in order to sample data matching specific topics. Additionally

the friendship relations of observed users were obtained as well through Twitters API.
The total data set contains:

- 529,692,345 user profiles,
- 4,701,571,902 messages,
- 14,436,364,697 friendship relations and
- 826,154,616 GPS positions.

SUMMARY

This thesis presents methods and techniques to analyze content propagation within online social networks (OSNs) using a graph theoretical approach. Important factors and different techniques to analyze and describe content propagation, starting from the smallest entity in a network, representing a user-account, up to complete friendship graphs and traces of content are described.

All individuals and their attributes are stating the basic elements for statistical analysis of user behavior and individuals interests. When trying to identify the opinion of the population of a country for example, a random sample or data from everyone within the population is needed, a task which is not trivial because of different activity patterns and the fact that individuals may either do not provide information about themselves or obscure their data by supplying bogus information. This thesis shows that obtaining a random sample of the population of the Netherlands is possible in terms of certain parameters like the location, family and first names of users. Such a sample is likely not to be "random" in terms of the age of inhabitants and the usage of gathered data in order to predict the outcome of elections may be questioned.

The representation of an individual's view onto an OSN is called an ego-centric network. It contains all friends and relations between friends of an ego within a sub-graph. Within such graphs, the influence between friends can be estimated improving the usability of recommendation systems which also raises concerns about the privacy of users. This thesis describes possibilities to reconstruct private information of a user if only a few friends of the individual share their data publicly because most friendships are created between persons having similar interests. Therefore the current way of dealing with privacy concerns, by enabling users to protect their data, is not sufficient. The structure of ego-centric networks also unveils the ability of egos to spread and control the spread of information as a person completely embedded in a group has less control over disseminating content than a person connecting multiple groups.

A snapshot of a whole network of an OSN includes all user-accounts (nodes) and friendships (links) at a certain point in time. But as OSNs may contain millions of nodes the process of obtaining data by crawling is likely to be skewed depending on the used method and duration. Therefore a new way of traversing the graph called "Mutual Friend Crawling" is proposed in which certain network metrics converge faster to the final value by also detecting communities of users while traversing the graph.

When analyzing the diffusion process of content in multiple OSNs, only a limited fraction of the neighbors of a user (i.e. friends) are "useful" in terms of spreading content to their peers. Commonly used network metrics which reflect the centrality of a node are shown to have no correlation with the ability to repeatedly succeed in passing messages to a high number of users. The reason lies in the fact that the whole network of friends contains inactive or abandoned user accounts and a critical dependency to the time a message was sent exist. This denotes that friends of a user that forward a message have

to be available or online at the time they are "needed" in order to forward content. On the other hand, influential groups might exist which act together in order to spread content with the help of each other. These groups might organize themselves via external communication channels, shown by the example of a famous group, the "Digg Patriots", where members of the group cannot be found through purely topological measures.

A similar time dependency exist in terms of the evolution of OSNs, because users can only forward information or befriend others when they are online. The interactivity durations of these actions are shown to be log-normal like distributed rather than exponential or power-law as assumed in multiple previous publications. The argumentation for such an assumption is based on the fact that power-law and exponential distributions would indicate most interactivity durations to be very short whereas individuals always need some time to complete tasks. However, it is shown that the time-scale of observations is crucial, because log-normal and power-law distributions with a small exponent $\gamma < 2$ might look the same in a log-log plot if the chosen bin-size is too large.
Another process involved in the structural evolution of a friendship network is given by markets that sell friendship relations in OSNs. These markets are accounting for quite a high number of friendship relations whereas their usage has usually a negative connotation. But in terms of content propagation they might be beneficial because, for example politicians, "buying" followers are able to reach users which would otherwise not connect to them.

The term viral spreading is often used in combination with content propagation within the network of an OSN. Therefore certain parameters of epidemiology are compared to "viral spreading" in Twitter. It was found that most messages had a low basic reproductive ratio < 1, a ratio depicting the infectious a virus, whereas few messages were highly infectious because a high number of users forwarded them. Interestingly even these popular messages were not able to spread to a large fraction of the total number of Twitter users. When trying to use epidemiological theory the "Susceptible-Exposed-Infected-Removed" model seems to be applicable to content propagation exhibiting the complication that the distribution of the duration a user is "exposed" and "infected" seems to be log-normal distributed. The distribution of these durations, also called observation and reaction duration denotes that Markov theory cannot be applied to model the "epidemics". Another more general approach is therefore given by a Bellman-Harris branching process.

The content, propagating though a network can be analyzed using graph theory as well in order to get insights into population statistics. The example of mobility pattern was chosen to depict the "meaning" of community detection within graphs created out of locations of Twitter users. The detected patterns allow better planning of transportation services, depicting in which areas of the Netherlands people are most frequently traveling during the working weekdays and weekends.
Analyzing the most common type of content, short colloquial text, using a new unsupervised way of estimating the sentiment of messages enables the analysis of graphs in which words are denoted as nodes and links describe the co-occurrence of words. These graphs reflect which words are related to concepts and their sentiment allowing to infer the perception of products and concepts within the population of OSN users.

SAMENVATTING

Dit proefschrift biedt methoden en technieken om de verspreiding van informatie in online sociale netwerken (OSN) te analyseren door middel van een graaf theoretische benadering. Belangrijke factoren en verschillende technieken om informatieverspreiding te analyseren en te beschrijven, van gebruikersaccount, het kleinste element, tot volledige vriendschap netwerken en sporen van inhoud, zijn belicht.

In principe zijn er altijd twee mensen in een OSN nodig om een relatie te creëren, in het geval van een gerichte relatie voldoet een niet-wederzijdse interesse van één gebruiker in een ander. Alle gebruikers en hun eigenschappen bevatten belangrijke elementen voor statistische analyse van gebruikers gedrag en individuele interesses. Om bijvoorbeeld de opinie van een bevolking van een land te identificeren, is het noodzakelijk data van een verzameling willekeurige gebruikers of van de gehele bevolking te hebben. Een taak die niet triviaal is omdat er verschillende activiteitspatronen van gebruikers zijn en het feit dat personen hun informatie niet delen ofwel verbergen door het bewust onjuist invullen van gegevens. Dit proefschrift toont dat het mogelijk is om een willekeurige verzameling van de Nederlandse bevolking te verkrijgen op basis van locatie, familienaam en de voornaam van gebruikers. Het blijkt omstreden of een dergelijke verzameling ook de leeftijd van de inwoners willekeurig weergeeft.

De vertegenwoordiging van het gezichtspunt van een individu op een OSN wordt een egocentrisch netwerk genoemd. Deze bevat alle vrienden en relaties tussen vrienden van de individu in een subgraaf. Binnen deze grafen is het mogelijk de invloed tussen vrienden in te schatten om de bruikbaarheid van aanbevelingssystemen te verbeteren. Gelijktijdig ontstaan zorgen omtrent de privacy waarborging van gebruikers. Dit proefschrift beschrijft mogelijkheden om particuliere informatie van een gebruiker te reconstrueren. Doordat de meeste vriendschappen tussen mensen uit gelijkaardige belangen bestaan, is reconstructie al uitvoerbaar indien slechts een beperkt aantal vrienden hun gegevens in het openbaar delen. Hieruit blijkt dat de huidige manier waarop wij met privacy om gaan, door de gebruiker zijn instellingen te laten wijzigen, niet voldoet. De structuur van egocentrische netwerken toont tevens de invloed van een individu om informatie te verspreiden en beheersen. Iemand die volledig in een groep ingebed is heeft minder controle over de verspreiding van inhoud dan iemand die meerdere groepen verbindt.

Een momentopname van de graaf van een heel OSN omvat alle knopen en takken bestaand op een zeker tijdstip. Maar omdat OSNs miljoenen knopen kunnen bevatten is het proces om data via crawling te verkrijgen waarschijnlijk niet precies. Nauwkeurigheid varieert afhankelijk van de gebruikte methode en looptijd. Daarom is er een nieuwe methode, “Mutual Friend Crawling”, in dit proefschrift voorgesteld om een graaf te doorkruisen. Met deze methode convergeren netwerk eigenschappen sneller naar de definitieve waarden door gemeenschappen van gebruikers te detecteren.

Uit analyse van het diffusie proces van inhoud in meerdere OSNs verschillende methoden blijkt slechts een beperkte fractie van de buren van een gebruiker (d.w.z. de

vrienden) “nuttig” te zijn in termen van het doorsturen van boodschappen naar hun vrienden. De gebruikelijke netwerkeigenschappen die de centraliteit van een knoop weerspiegelen tonen geen correlatie met het vermogen om meermaals succesvol informatie te verspreiden. Dit komt door het feit dat een netwerk van vrienden ook inactieve of verlaten gebruikersaccounts bevat welke versterkt wordt door cruciale afhankelijkheid van de tijd sinds het bericht verstuurd is. Dit geeft aan dat de vrienden van een gebruiker die een boodschap stuurt op dat moment beschikbaar of online moeten zijn om inhoud door te sturen. Daarom is het mogelijk dat groepen verschijnen die samenwerken om inhoud te verspreiden en elkaar helpen. Deze groepen kunnen zich via externe communicatie kanalen organiseren, zoals de bekende groep de “Digg Patriots” wiens leden niet door puur topologische eigenschappen gevonden kunnen worden.

Een vergelijkbare tijdsafhankelijkheid bestaat in termen van de evolutie van OSNs, omdat gebruikers informatie alleen mogen doorsturen of anderen kunnen bevrienden wanneer ze online zijn. De interactiviteits-duur van deze handelingen zijn lognormaal verdeeld, dit in tegenstelling tot de in eerder uitgebrachte publicaties aangenomen exponentiële of power-law verdelingen. De argumentatie voor deze aanname is gebaseerd op het feit dat power-law en exponentiële verdelingen duiden op veel voorkomende korte interactiviteits-sessies terwijl personen altijd enige tijd nodig hebben om een taak af te ronden. In dit proefschrift is verduidelijkt dat de tijdschaal van observaties heel belangrijk is aangezien lognormale en power-law verdelingen met een kleine exponent $\gamma < 2$ gelijk zijn in een log-log plot als de gekozen klassengrootte te groot is.
Een ander proces betrokken bij de structurele evolutie van een vriendschapsnetwerk is gegeven door markten die vriendschap relaties in OSNs verkopen . Deze markten brengen een groot aantal vriendschap relaties in terwijl het gebruik ervan een negatieve connotatie heeft. Qua inhouds-verspreiding kan een markt toch voordelig zijn bijvoorbeeld omdat politici die volgers “kopen” de mogelijkheid hebben om gebruikers te bereiken die anders geen relatie hadden gecreëerd.

De term virale verspreiding wordt vaak gebruikt in combinatie met inhouds-verspreiding binnen het netwerk van een OSN. Daarom zijn bepaalde epidemiologische parameters berekend, onder andere de hypothese dat boodschappen in Twitter verschillende virussen weerspiegelen. De meeste berichten hadden een basis-reproductieve ratio < 1, een ratio die beschrijft hoe aanstekelijk een virus is, terwijl sommige berichten heel infectieus zijn omdat veel gebruikers deze berichten door gestuurd hebben. Vreemd genoeg blijkt dat zelfs deze populaire berichten niet naar het overgrote deel van Twitter gebruikers werd verspreid. Vanuit de epidemiologische theorie lijkt het erop dat het “Susceptible-Exposed-Infected-Removed” model geschikt is voor inhouds-verspreiding. Hierin zijn de verdelingen van de tijdsduur dat een gebruiker “blootgesteld” en “geïnfecteerd” is lognormaal verdeeld. Deze verdelingen, waaronder observatie en reactie duur vallen, betekenen dat Markov theorie niet aangewend kan worden. Een meer algemeen model wordt gegeven door het Bellman-Harris vertakkingsproces.

De inhoud die zich door het netwerk verspreidt kan ook via grafentheorie geanalyseerd worden om inzichten in bevolkingsstatistieken te krijgen. Een voorbeeld van mobiliteits-patronen werd gekozen om de betekenis van groep detectie in grafen gegenereerd uit de locaties gedeeld door Twitter gebruikers aan te tonen. De gedetecteerde patronen maken het mogelijk om vervoersdiensten te verbeteren, doordat het mogelijk

is buurten te identificeren waaruit inwoners van Nederland vaak tijdens het weekeinde en werkdagen reizen.

Het analyseren van de meest gebruikelijke soort van inhoud, zijnde korte teksten, via een nieuwe methode welke zonder toezicht het gevoel van berichten schat, maakt het mogelijk om grafen te analyseren waarin knopen woorden en links gelijktijdig gebruik van deze woorden representeren. Deze grafen weerspiegelen welke woorden verwant zijn met concepten en welke emotie mensen met deze concepten associëren, hieruit kun je de waarneming van producten en concepten binnen een bevolking van OSN gebruikers afleiden.

REFERENCES

[1] S. Milgram, *The small-world problem,* Psychology Today **1**, 61 (1967).

[2] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks,* Nature **393**, 440 (1998).

[3] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, *The anatomy of the facebook social graph,* CoRR **abs/1111.4503** (2011).

[4] R. Dunbar, *Neocortex size as a constraint on group size in primates,* Journal of Human Evolution **22**, 469 (1992).

[5] B. Gonçalves, N. Perra, and A. Vespignani, *Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number,* PLoS ONE **6**, e22656 (2011).

[6] The economist, *Primates on Facebook,* (2009).

[7] N. A. Christakis and J. H. Fowler, *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives – How Your Friends' Friends' Friends Affect Everything You Feel, Think, and Do* (Little, Brown and Company, 2011).

[8] P. S. Levy and S. Lemeshow, *Sampling of Populations: Methods and Applications*, 3rd ed. (Wiley-Interscience, 1999).

[9] Nationalencyklopedin, *Nationalencyklopedin, the most comprehensive contemporary Swedish language encyclopedia,* (2013).

[10] N. Shuyo, *Language Detection Library for Java,* (2010).

[11] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, *Predicting elections with twitter: What 140 characters reveal about political sentiment,* in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (2010) pp. 178–185.

[12] E. T. K. Sang and J. Bos, *Predicting the 2011 Dutch Senate Election Results with Twitter,* in *Proceedings of the Workshop on Semantic Analysis in Social Media* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2012) pp. 53–60.

[13] A. O. Larsson and H. Moe, *Studying political microblogging: Twitter users in the 2010 Swedish election campaign.* New Media & Society **14**, 729 (2012).

[14] J. Bollen, H. Mao, and X. Zeng, *Twitter mood predicts the stock market,* Journal of Computational Science **2**, 1 (2011).

[15] S. Asur and B. A. Huberman, *Predicting the Future with Social Media,* in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10 (IEEE Computer Society, 2010) pp. 492–499.

[16] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, *Understanding the Demographics of Twitter Users,* in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)* (Barcelona, Spain, 2011).

[17] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series,* in *Proceedings of the International AAAI Conference on Weblogs and Social Media* (2010).

[18] A. Jungherr, P. Jürgens, and H. Schoen, *Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment,* Soc. Sci. Comput. Rev. **30**, 229 (2012).

[19] D. Gayo-Avello, P. T. Metaxas, and E. Mustafaraj, *Limits of Electoral Predictions Using Twitter.* in *ICWSM*, edited by L. A. Adamic, R. A. Baeza-Yates, and S. Counts (The AAAI Press, 2011).

[20] *The Meertens Institute,* Website (2012), http://www.meertens.knaw.nl/cms/en/meertens-institute.

[21] *Centraal Bureau voor de Statistiek,* Website (2012), http://www.cbs.nl/nl-NL/menu/home/default.htm.

[22] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, *How Old Do You Think I Am?: A Study of Language and Age in Twitter,* in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, ICWSM 2013 (2013).

[23] P. Van Mieghem, *Performance Analysis of Communications Systems and Networks* (Cambridge University Press, 2006).

[24] C. B. Department and C. Borgelt, *Efficient Implementations of Apriori and Eclat,* (2003).

[25] E. Locard and W. Finke, *Die Kriminaluntersuchung und ihre wissenschaftlichen Methoden* (Kameradschaft Verlags-Gesellschaft, 1930).

[26] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, *Unique in the Crowd: The privacy bounds of human mobility,* Scientific Reports **3** (2013).

[27] E. H. Simpson, *The interpretation of interaction in contingency tables,* Journal of the Royal Statistical Society, Series B (Statistical Methodology) **13**, 238 (1951).

[28] N. A. Christakis and J. H. Fowler, *Social Network Sensors for Early Detection of Contagious Outbreaks,* PLoS ONE **5**, e12948 (2010).

[29] A. Miller, *Untangling the social web,* The Economist (2010).

[30] L. Scism and M. Maremont, *Insurers Test Data Profiles to Identify Risky Clients,* Wall Street Journal (2010).

[31] N. Singer, *Face Recognition Makes the Leap From Sci-Fi,* New York Times (2011).

[32] M. McPherson, L. Smith-Lovin, and J. M. Cook, *Birds of a Feather: Homophily in Social Networks,* Annual Review of Sociology **27**, 415 (2001).

[33] N. Blenn, C. Doerr, N. Shadravan, and P. Van Mieghem, *How Much do your Friends Know about You?: Reconstructing Private Information from the Friendship Graph,* in *Proceedings of the Fifth Workshop on Social Network Systems*, SNS '12 (ACM, New York, NY, USA, 2012) pp. 2:1–2:6.

[34] C. Doerr, S. Tang, N. Blenn, and P. Van Mieghem, *Are Friends Overrated? A Study for the Social News Aggregator Digg.com,* in *NETWORKING 2011, Part II*, Lecture Notes in Computer Science 6641, edited by J. D.-P. et al. (IFIP International Federation for Information Processing, 2011) pp. 314–327.

[35] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, *You Are Who You Know: Inferring User Profiles in Online Social Networks,* in *Proceedings of WSDM* (2010).

[36] K. Raynes-Goldie, *Pulling sense out of today's informational chaos: LiveJournal as a site of knowledge creation and sharing,* First Monday **8** (2004).

[37] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, *All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks,* in *18th International World Wide Web Conference* (2009) pp. 551–551.

[38] R. Gross and A. Acquisti, *Information revelation and privacy in online social networks,* in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, WPES '05 (ACM, New York, NY, USA, 2005) pp. 71–80.

[39] B. Krishnamurthy and C. E. Wills, *Characterizing Privacy in Online Social Networks,* in *Proceedings of the first workshop on Online social networks* (2008).

[40] B. Krishnamurthy and C. E. Wills, *On the leakage of personally identifiable information via online social networks,* SIGCOMM Comput. Commun. Rev. **40**, 112 (2010).

[41] I. Bywater, *Aristotelis Ethica Nicomachea* (Cambridge University Press, 2010) p. 129.

[42] J. He, W. Chu, and Z. Liu, *Inferring Privacy Information from Social Networks,* in *Intelligence and Security Informatics*, Lecture Notes in Computer Science, Vol. 3975 (Springer-Verlag, Berlin/Heidelberg, 2006) pp. 154–165.

[43] E. Zheleva and L. Getoor, *To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles,* in *18th International World Wide Web Conference* (2009) pp. 531–531.

[44] J. Bonneau, J. Anderson, R. Anderson, and F. Stajano, *Eight Friends are Enough: Social Graph Approximation via Public Listings,* in *Proceedings of Second ACM Workshop on Social Network Systems* (2009).

[45] GeoNames, *The GeoNames geographical database,* (2013), http://www.geonames.org/.

[46] D. Jurgens, *That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships,* in *ICWSM* (2013).

[47] M. S. Granovetter, *The Strength of Weak Ties,* American Journal of Sociology **78**, pp. 1360 (1973).

[48] M. Mcpherson, L. Smith-lovin, and M. E. Brashears, *Social isolation in America: Changes in core discussion networks over two decades,* American Sociological Review **71**, 353 (2006).

[49] R. S. Burt, *Structural holes: The social structure of competition* (Harvard University Press, Cambridge, MA, 1992).

[50] *Key Facts - Facebook's latest news, announcements and media resources,* http://newsroom.fb.com/Key-Facts (2014).

[51] *Twitter,* about.twitter.com/company (Feb 20, 2014).

[52] T. Cormen, *Introduction to algorithms*, MIT electrical engineering and computer science series (MIT Press, 2001).

[53] L. A. Goodman, *Snowball Sampling,* The Annals of Mathematical Statistics **32** (1961).

[54] J. Leskovec, J. Kleinberg, and C. Faloutsos, *Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations,* in *KDD* (2005).

[55] C.-H. Lee, X. Xu, and D. Y. Eun, *Beyond random walk and metropolis-hastings samplers: why you should not backtrack for unbiased graph sampling,* in *SIGMETRICS* (2012).

[56] N. Blenn, C. Doerr, B. Van Kester, and P. Van Mieghem, *Crawling and Detecting Community Structure in Online Social NetworksUsing Local Information,* in *NETWORKING 2012*, Lecture Notes in Computer Science, Vol. 7289, edited by R. Bestak, L. Kencl, L. Li, J. Widmer, and H. Yin (Springer Berlin / Heidelberg, 2012) pp. 56–67.

[57] M. Kurant, A. Markopoulou, and P. Thiran, *On the bias of BFS (Breadth First Search),* in *Teletraffic Congress (ITC), 2010 22nd International* (IEEE, 2010) pp. 1–8.

[58] S. L. Feld, *Why Your Friends Have More Friends Than You Do,* American Journal of Sociology **96**, 1464 (1991).

[59] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, *Walking in Facebook: A Case Study of Unbiased Sampling of OSNs,* in *INFOCOM* (2010).

[60] L. Lovasz, *Random Walks on Graphs: A SURVEY,* Combinatorics **2**, 1 (1993).

[61] C. Doerr, N. Blenn, S. Tang, and P. Van Mieghem, *Are Friends Overrated? A Study for the Social News Aggregator Digg.com,* in *Computer Communications*, Vol. 35 (2012).

[62] N. Litvak and R. van der Hofstad, *Uncovering disassortativity in large scale-free networks,* Phys. Rev. E **87**, 022801 (2013).

[63] J. Kleinfeld, *Could It Be A Big World After All? The "Six Degrees of Separation" Myth,* Society (2002).

[64] A. Clauset, M. E. J. Newman, and C. Moore, *Finding community structure in very large networks,* Physical Review E **70** (2004), 10.1103/PhysRevE.70.066111.

[65] S. Trajanovski, H. Wang, and P. Van Mieghem, *Maximum modular graphs,* The European Physical Journal B - Condensed Matter and Complex Systems **85**, 1 (2012).

[66] M. Latapy and P. Pons, *Computing communities in large networks using random walks,* (2005) pp. 284–293.

[67] D. Lai, H. Lu, and C. Nardini, *Enhanced modularity-based community detection by random walk network preprocessing,* Phys. Rev. E **81**, 066118 (2010).

[68] J. Reichardt and S. Bornholdt, *Statistical mechanics of community detection.* Phys Rev E Stat Nonlin Soft Matter Phys **74** (2006), 10.1103/PhysRevE.74.016110.

[69] M. Girvan and M. E. J. Newman, *Community structure in social and biological networks,* Proceedings of the National Academy of Sciences **99**, 7821 (2002).

[70] M. E. J. Newman and M. Girvan, *Finding and evaluating community structure in networks,* Phys. Rev. E **69**, 026113 (2004).

[71] U. N. Raghavan, R. Albert, and S. Kumara, *Near linear time algorithm to detect community structures in large-scale networks,* Physical Review E **76**, 036106+ (2007).

[72] N. Nguyen, T. Dinh, Y. Xuan, and M. Thai, *Adaptive algorithms for detecting community structure in dynamic social networks,* in *INFOCOM, 2011 Proceedings IEEE* (2011) pp. 2282 –2290.

[73] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Fast unfolding of community hierarchies in large networks,* CoRR (2008), abs/0803.0476.

[74] S. Fortunato and C. Castellano, *Community Structure in Graphs,* (2007).

[75] S. Van Kester, *Efficient Crawling of Community Structures in Online Social Networks* (PVM 2011-071, Tu Delft, 2011).

[76] W. W. Zachary, *An Information Flow Model for Conflict and Fission in Small Groups,* Journal of Anthropological Research **33** (1977), 10.2307/3629752.

[77] S. Tang, N. Blenn, C. Doerr, and P. Van Mieghem, *Digging in the Digg Social News Website,* IEEE Trans. Multimedia **13**, 1163 (2011).

[78] S. Fortunato and M. Barthélemy, *Resolution limit in community detection,* Proceedings of the National Academy of Sciences **104**, 36 (2007).

[79] S. Fortunato, *Community detection in graphs,* Physics Reports **486**, 75 (2010).

[80] P. Van Mieghem, X. Ge, P. Schumm, S. Trajanovski, and H. Wang, *Spectral graph analysis of modularity and assortativity,* Physical Review E **82**, 056113+ (2010).

[81] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, *Uncovering the overlapping community structure of complex networks in nature and society,* Nature **435**, 814 (2005).

[82] T. S. Evans and R. Lambiotte, *Line graphs, link partitions, and overlapping communities,* Phys. Rev. E **80**, 016105 (2009).

[83] A. McDaid and N. Hurley, *Detecting Highly Overlapping Communities with Model-Based Overlapping Seed Expansion,* in *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '10 (IEEE Computer Society, 2010) pp. 112–119.

[84] P. Pollner, G. Palla, and T. Vicsek, *Preferential attachment of communities: The same principle, but a higher level,* EPL (Europhysics Letters) **73**, 478 (2006).

[85] R. Toivonen, J. Onnela, J. Saramäki, and J. Hyvönen, *A model for social networks,* Physica A: Statistical Mechanics and its Applications (2006).

[86] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, *Random graphs with arbitrary degree distributions and their applications,* Physical Review E **64**, 026118 (2001).

[87] S. Lattanzi and D. Sivakumar, *Affiliation Networks,* in *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, STOC '09 (ACM, New York, NY, USA, 2009) pp. 427–434.

[88] P. Van Mieghem, *Graph Spectra for Complex Networks* (Cambridge University Press, New York, NY, USA, 2011).

[89] D. Cvetkovic, P. Rowlinson, and S. K. Simic, *Eigenvalue bound for the signless laplacian,* Publications de l'institute mathematique, nouvelle serie **81**, 11 (2007).

[90] J. Leskovec, J. Kleinberg, and C. Faloutsos, *Graph Evolution: Densification and Shrinking Diameters,* ACM Trans. Knowl. Discov. Data **1** (2007).

[91] A. L. Barabási and R. Albert, *Emergence of scaling in random networks,* Science **286**, 509 (1999).

[92] J. C. Nacher, T. Yamada, S. Goto, M. Kanehisa, and T. Akutsu, *Two complementary representations of a scale-free network,* Physica A: Statistical Mechanics and its Applications **349**, 349 (2005).

[93] A. Manka-Krason, A. Mwijage, and K. Kulakowski, *Clustering in random line graphs.* Computer Physics Communications **181**, 118 (2010).

[94] J. Surowieck, *The Wisdom of Crowds* (Anchor, 2005).

[95] M. Richardson and P. Domingos, *Mining knowledge-sharing sites for viral marketing,* in *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2002) pp. 61–70.

[96] W.-S. Yang, J.-B. Dia, H.-C. Cheng, and H.-T. Lin, *Mining social networks for targeted advertising,* in *Proceedings of the 39th Hawaii International Conference on System Sciences* (2006).

[97] J. Davitz, J. Yu, S. Basu, D. Gutelius, and A. Harris, *ilink: search and routing in social networks,* in *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (2007) pp. 931–940.

[98] D. M. Boyd and N. B. Ellison, *Social network sites: Definition, history, and scholarship,* Journal of Computer-Mediated Communication **13**, 210 (2007).

[99] E. Katz and P. F. Lazarsfeld, *Personal Influence* (Free Press, 1955).

[100] D. Krackhardt and R. N. Stern, *Informal Networks and Organizational Crises: An Experimental Simulation,* Social Psychology Quarterly **51**, 123 (1988).

[101] M. T. Hansen, *The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunits,* Administrative Science Quarterly **44**, 82 (1999).

[102] G. C. Homans, *Social Behavior: Its Elementary Forms* (New York: Harcourt, Brace & World, 1961).

[103] J. S. Coleman, E. Katz, and H. Menzel, *Medical Innovation* (New York: Bobbs-Merrill, 1966).

[104] R. S. Burt, *Social Contagion and Innovation: Cohesion Versus Structural Equivalence,* American Journal of Sociology **92**, 1287 (1987).

[105] W. Tsai, *Knowledge transfer in intraorganizational networks: Effects of network position and absorptive capacity on business unit innovation and performance,* Academy of Management Journal **44**, 996 (2001).

[106] R. Albert, H. Jeong, and A.-L. Barabási, *Error and attack tolerance of complex networks,* Nature **406**, 378 (2000).

[107] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, *Measurement and analysis of online social networks,* in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (ACM, 2007) p. 42.

[108] J. Leskovec and E. Horvitz, *Planetary-scale views on a large instant-messaging network,* in *Proceeding of the 17th international conference on World Wide Web* (ACM, 2008) pp. 915–924.

[109] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, *Group formation in large social networks: membership, growth, and evolution,* in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2006) p. 54.

[110] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, *Characterizing user behavior in online social networks,* in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference* (ACM, 2009) pp. 49–62.

[111] *Orkut,* http://www.orkut.com (2010).

[112] *MySpace,* http://www.myspace.com (2010).

[113] *Hi5,* http://hi5.com (2010).

[114] *LinkedIn,* http://www.linkedin.com (2014).

[115] M. Cha, A. Mislove, and K. P. Gummadi, *A measurement-driven analysis of information propagation in the flickr social network,* in *Proceedings of the 18th international conference on World wide web* (ACM, 2009) pp. 721–730.

[116] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, *Analyzing the video popularity characteristics of large-scale user generated content systems,* IEEE/ACM Transactions on Networking (TON) **17**, 1357 (2009).

[117] G. Szabo and B. A. Huberman, *Predicting the popularity of online content,* Commun. ACM **53**, 80 (2010).

[118] E. B. Keller and J. Berry, *The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and what to Buy* (The Free Press, 2003).

[119] D. Fono and K. Raynes-Goldie, *Hyperfriends and Beyond: Friendship and Social Norms on LiveJournal.* Internet Research Annual: Selected Papers from the Association of Internet Researchers Conference **4** (2006).

[120] T. L. Chartrand and J. A. Bargh, *The chameleon effect: The perception–behavior link and social interaction,* Journal of Personality and Social Psychology **76**, 893 (1999).

[121] A. Goolsbee and P. J. Klenow, *Evidence on Learning and Network Externalities in the Diffusion of Home Computers,* Journal of Law & Economics **45**, 317 (2002).

[122] P. Van Mieghem, N. Blenn, and C. Doerr, *Lognormal distribution in the digg online social network,* Eur. Phys. J. B **83**, 251 (2011).

[123] J. P. Scott, *Social Network Analysis: A Handbook* (Sage, 2000).

[124] N. B. Ellison, C. Steinfield, and C. Lampe, *The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites,* Journal of Computer-Mediated Communication **12**, 1143 (2007).

[125] Alexa, *http://www.alexa.com/,* (2013).

[126] G. U. Yule, *A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, F.R.S.",* Philosophical Transactions of the Royal Society of London B **213**, 21 (1925).

[127] P. Csermely, *Weak Links: Stabilizers of Complex Systems from Proteins to Social Networks* (Springer Berlin, 2006).

[128] J. Deese and R. Kaufman, *Serial effects in recall of unorganized and sequentially organized verbal material,* J Exp Psychol **54** (1957).

[129] D. Jeffrey, M. Feng, N. Gupta, and R. Gupta, *BugFix: A Learning-Based Tool to Assist Developers in Fixing Bugs,* in *17th IEEE International Conference on Program Comprehension* (2009).

[130] J. Halliday, *Digg investigates claims of conservative 'censorship',* guardian.co.uk (2010).

[131] A.-L. Barabási, *The origin of bursts and heavy tails in human dynamics,* Nature **435**, 207 (2005), cond-mat/0505371 .

[132] J.-P. Eckmann, E. Moses, and D. Sergi, *Entropy of dialogues creates coherent structures in e-mail traffic,* Proceedings of the National Academy of Sciences of the United States of America **101**, 14333 (2004).

[133] B. Gonçalves and J. J. Ramasco, *Human dynamics revealed through Web analytics,* CoRR **abs/0803.4018** (2008).

[134] F. Radicchi, *Human activity in the web,* Phys. Rev. E **80**, 026118 (2009).

[135] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási, *Uncovering individual and collective human dynamics from mobile phone records,* Journal of Physics A: Mathematical and Theoretical **41**, 224015 (2008).

[136] M. Karsai, K. Kaski, A.-L. Barabási, and J. Kertész, *Universal features of correlated bursty behaviour,* Scientific Reports **2**, 397 (2012).

[137] J. L. Iribarren and E. Moro, *Impact of Human Activity Patterns on the Dynamics of Information Diffusion,* Physical Review Letters **103**, 038702+ (2009).

[138] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A. L. Barabási, and J. Saramäki, *Small but slow world: How network topology and burstiness slow down spreading,* Physical Review E **83**, 025102 (2011).

[139] E. Cator, R. van de Bovenkamp, and P. Van Mieghem, *Susceptible-infected-susceptible epidemics on networks with general infection and cure times,* Phys. Rev. E **87**, 062816 (2013).

[140] P. Van Mieghem and R. van de Bovenkamp, *Non-Markovian Infection Spread Dramatically Alters the Susceptible-Infected-Susceptible Epidemic Threshold in Networks,* Phys. Rev. Lett. **110**, 108701 (2013).

[141] D. B. Stouffer, R. D. Malmgren, and L. A. N. Amaral, *Comment on The origin of bursts and heavy tails in human dynamics,* arXiv:physics/0510216 (2005).

[142] R. Malmgren, D. Stouffer, A. Motter, and L. Amaral, *A Poissonian explanation for heavy tails in e-mail communication,* Proceedings of the National Academy of Sciences **105**, 18153 (2008).

[143] C. Doerr, N. Blenn, and P. Van Mieghem, *Lognormal Infection Times of Online Information Spread,* PLoS ONE **8**, e64349+ (2013), 1305.5235 .

[144] P. Van Mieghem, *Performance Analysis of Complex Networks and Systems* (Cambridge University Press, 2014) (*to appear*).

[145] A. Clauset, C. R. Shalizi, and M. E. J. Newman, *Power-Law Distributions in Empirical Data,* SIAM Rev. **51**, 661 (2009).

[146] E. Limpert, W. A. Stahel, and M. Abbt, *Log-normal Distributions across the Sciences: Keys and Clues,* BioScience **51**, 341 (2001).

[147] J. W. Boag, *Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy,* Journal of the Royal Statistical Society. Series B (Methodological) **11** (1949), 10.2307/2983694.

[148] D. B. Stouffer, R. D. Malmgren, and L. A. N. Amaral, *Log-normal statistics in e-mail communication patterns,* (2006).

[149] H. Nishiura, *Early efforts in modeling the incubation period of infectious diseases with an acute course of illness,* Emerging Themes in Epidemiology (2007).

[150] P. E. Sartwell, *The distribution of incubation periods of infectious disease,* American Journal of Hygiene **51**, 310 (1950).

[151] M. Feinleib and B. Macmahon, *Variation in the duration of survival of patients with the chronic leukemias.* Blood **15** (1960).

[152] R. D. Horner, *Age at onset of Alzheimer's disease: clue to the relative importance of etiologic factors?* Am J Epidemiol **126**, 409 (1987).

[153] J. Lessler, N. G. Reich, R. Brookmeyer, T. M. Perl, K. E. Nelson, and D. A. Cummings, *Incubation periods of acute respiratory viral infections: a systematic review.* The Lancet infectious diseases **9**, 291 (2009).

[154] W. J. van der Linden, *A Lognormal Model for Response Times on Test Items,* Journal of Educational and Behavioral Statistics **31** (2006).

[155] R. J. Lawrence, *The Lognormal Distribution of the Duration of Strikes,* Journal of the Royal Statistical Society. Series A **147** (1984).

[156] S. Mohan, M. Gopalakrishnan, H. Balasubramanian, and A. Chandrashekar, *A lognormal approximation of activity duration in PERT using two time estimates,* Journal of the Operational Research Society **58**, 827 (2006).

[157] A. Spedalieri, I. Martín-Escalona, and F. Barceló, *Simulation of teletraffic variables in UMTS networks: impact of lognormal distributed call duration.* in *WCNC* (IEEE, 2005) pp. 2381–2386.

[158] F. Barcelo and J. Jordan, *Channel holding time distribution in public telephony systems (PAMR and PCS),* IEEE Transactions on Vehicular Technology (2000).

[159] Y.-H. Eom and S. Fortunato, *Characterizing and Modeling Citation Dynamics,* PLoS ONE **6**, e24926 (2011).

[160] S. Redner, *Citation Statistics from 110 Years of Physical Review,* Physics Today **58**, 49 (2005).

[161] M. Stringer, M. Sales-Pardo, and L. Amaral, *Effectiveness of journal ranking schemes as a tool for locating information,* PLoS One **3**, e1683 (2008).

[162] F. Radicchi, S. Fortunato, and C. Castellano, *Universality of citation distributions: Toward an objective measure of scientific impact,* PNAS **105** (2008).

[163] P. Van Mieghem, N. Blenn, and C. Doerr, *Lognormal distribution in the digg online social network,* The European Physical Journal B - Condensed Matter and Complex Systems **83**, 251 (2011), 10.1140/epjb/e2011-20124-0.

[164] M. Mitzenmacher, *A Brief History of Generative Models for Power Law and Lognormal Distributions,* Internet Mathematics **1** (2003).

[165] X. Gabaix, *Zipf's Law For Cities: An Explanation,* Quarterly Journal of Economics **114**, 114 (1999).

[166] R. Gibrat, *Les Inégalités économiques [The Economic Inequalities]* (Librairie du Recueil Sirey, 1931, French).

[167] D. G. Champernowne, *A Model of Income Distribution,* The Economic Journal **63**, pp. 318 (1953).

[168] J. C. Cordoba, *A Generalized Gibrat's Law,* International Economic Review **49**, 1463 (2008).

[169] H. A. Simon, *On a class of skew distribution functions,* Biometrika **42**, 425 (1955).

[170] J. Eeckhout, *Gibrat's Law for (All) Cities,* American Economic Review **94**, 1429 (2004).

[171] J. C. Kapteyn, *Skew Frequency curves in Biology and Statistics,* Molecular and General Genetics MGG **19**, 205 (1918).

[172] Y. Malevergne, V. Pisarenko, and D. Sornette, *Gibrat's law for cities: uniformly most powerful unbiased test of the Pareto against the lognormal*, Swiss Finance Institute Research Paper Series 09-40 (Swiss Finance Institute, 2009).

[173] H. Kesten, *Random Difference Equations and Renewal theory for product of Random Matrices,* Acta Mathematica **CXXXI**, 207 (1973).

[174] F. Black and M. S. Scholes, *The Pricing of Options and Corporate Liabilities,* Journal of Political Economy **81**, 637 (1973).

[175] G. Stringhini, M. Egele, C. Kruegel, and G. Vigna, *Poultry markets: On the underground economy of twitter followers,* in *Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks*, WOSN '12 (ACM, 2012) pp. 1–6.

[176] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, *Four degrees of separation,* CoRR **abs/1111.4570** (2011).

[177] A. Signorini, A. M. Segre, and P. M. Polgreen, *The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic,* PLoS ONE **6**, e19467 (2011).

[178] R. Chunara, J. R. Andrews, and J. S. Brownstein, *Social and news media enable estimation of epidemiological patternsearly in the 2010 Haitian cholera outbreak.* American Journal of Tropical Medicine and Hygiene **86**, 39 (2012).

[179] K. Lerman and R. Ghosh, *Information contagion: An empirical study of the spread of news on Digg and Twitter social networks,* in *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)* (2010).

[180] R. M. Anderson and R. M. May, *Infectious diseases of humans: dynamics and control* (Oxford University Press, 1991).

[181] M. E. J. Newman, *The spread of epidemic disease on networks,* Physical Review E **66** (2002).

[182] M. J. Keeling and K. T. D. Eames, *Networks and epidemic models, Journal of The Royal Society Interface*, Journal of The Royal Society Interface **2**, 295 (2005).

[183] R. Pastor-Satorras and A. Vespignani, *Epidemic spreading in scale-free networks,* (2001).

[184] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, *Epidemic Spreading in Complex Networks with˜Degree Correlations,* in *Statistical Mechanics of Complex Networks*, Lecture Notes in Physics, Vol. 625, edited by R. Pastor-Satorras, M. Rubi, and A. Diaz-Guilera (Springer Berlin / Heidelberg, 2003) pp. 127–147.

[185] M. E. J. Newman, *Exact Solutions of Epidemic Models on Networks*, Working Papers (Santa Fe Institute, 2001).

[186] P. Grassberger, *On the critical behavior of the general epidemic process and dynamical percolation,* Mathematical Biosciences **63**, 157 (1983).

[187] A. Vazquez, B. Rácz, A. Lukács, and A. L. Barabási, *Impact of Non-Poissonian Activity Patterns on Spreading Processes,* Physical Review Letters **98**, 158702+ (2007).

[188] R. Bellman and T. E. Harris, *On the Theory of Age-Dependent Stochastic Branching Processes.* Proceedings of the National Academy of Sciences of the United States of America **34**, 601 (1948).

[189] J. Iribarren and E. Moro, *Branching dynamics of viral information spreading.* Phys Rev E Stat Nonlin Soft Matter Phys **84** (2011).

[190] P. Van Mieghem, J. Omic, and R. Kooij, *Virus spread in networks,* IEEE/ACM Trans. Netw. **17**, 1 (2009).

[191] P. Van Mieghem, *Markovian SIR and SIS epidemics on networks,* (1013), submitted.

[192] H. Kwak, C. Lee, H. Park, and S. Moon, *What is twitter, a social network or a news media?* in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10 (ACM, New York, NY, USA, 2010) pp. 591–600.

[193] P. Domingos and M. Richardson, *Mining the network value of customers,* in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01 (ACM, 2001) pp. 57–66.

[194] H. W. Watson and F. Galton, *On the probability of the extinction of families,* Journal of the Anthropological Institute **4** (1875).

[195] K. Rothman, S. Greenland, and T. Lash, *Modern Epidemiology* (Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008).

[196] P. Erdős and A. Rényi, *On the evolution of random graphs,* Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).

[197] Twitter, *http://twitter.com,* (2013).

[198] L. L. Yu, S. Asur, and B. A. Huberman, *Artificial Inflation: The Real Story of Trends and Trend-setters in Sina Weibo,* in *SocialCom* (2012).

[199] B. Bollobas, O. Riordan, J. Spencer, and G. Tusnady, *The Degree Sequence of a Scale-Free Random Graph Process,* Random Structures & Algorithms **18**, 279 (2001).

[200] E. L. Crow and K. Shimizu, *Lognormal Distribution, Theory and Application* (Marcel Dekker Inc., 1988).

[201] N. A. Marlow, *A normal limit theorem for power sums of normal random variables,* The Bell System Technical Journal **46**, 2081 (1967).

[202] M. E. J. Newman, *Power laws, Pareto distributions and Zipf's law,* Contemporary Physics (2005).

[203] W. Shockley, *On the Statisticsof Individual Variations of Productivity in Research Laboratories,* Proceedings of the IRE (1957).

[204] Foursquare in numbers, *https://foursquare.com/about,* (2014).

[205] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz, *Redrawing the map of great britain from a network of human interactions,* PLoS ONE **5**, e14248 (2010).

[206] B. Pang, L. Lee, and S. Vaithyanathan, *Thumbs up? Sentiment Classification using Machine Learning Techniques,* in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2002).

[207] P. D. Turney, *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,* in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (2002).

[208] G. Mishne and N. Glance, *Predicting movie sales from blogger sentiment,* in *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)* (2006).

[209] H. Yu and V. Hatzivassiloglou, *Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences,* in *Proceedings of the 2003 conference on Empirical methods in natural language processing* (2003) pp. 129–136.

[210] L. Barbosa and J. Feng, *Robust sentiment detection on twitter from biased and noisy data,* in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (2010) pp. 36–44.

[211] R. Cilibrasi and P. Vitanyi, *Automatic meaning discovery using Google,* Manuscript, CWI (2004).

[212] E. Riloff and J. Wiebe, *Learning extraction patterns for subjective expressions,* in *Proceedings of the 2003 conference on Empirical methods in natural language processing* (Association for Computational Linguistics, 2003) pp. 105–112.

[213] J. Wiebe and E. Riloff, *Creating subjective and objective sentence classifiers from unannotated texts,* Computational Linguistics and Intelligent Text Processing , 486 (2005).

[214] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, *Target-dependent twitter sentiment classification,* Proc. 49th ACL: HLT **1**, 151 (2011).

[215] E. Cambria, A. Hussain, C. Havasi, and C. Eckl, *Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems,* Development of Multimodal Interfaces: Active Listening and Synchrony , 148 (2010).

[216] A. Go, R. Bhayani, and L. Huang, *Twitter Sentiment Classification using Distant Supervision*, Tech. Rep. (Stanford University, 2009).

[217] Tweet Sentiments, *Tweet Sentiments, http://tweetsentiments.com,* (2013).

[218] Lingpipe, *http://alias-i.com/lingpipe,* (2013).

[219] N. Shuyo, *Language Detection Library*, http://code.google.com/p/language-detection/ (2010).

[220] D. Klein and C. D. Manning, *Accurate Unlexicalized Parsing,* in *Proceedings of the 41st Meeting of the Association for Computational Linguistics* (2003).

[221] WordNet, *A lexical database for English. http://wordnet.princeton.edu/,* (2013).

[222] D. K. C. M. Kristina Toutanova and Y. Singer, *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network,* in *Proceedings of HLT-NAACL* (2003).

[223] the University of Pennsylvania, *(Penn) Treebank Tag-set,* (2013), http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html.

[224] G. R. Boynton, A. Bates, E. Bettis, M. Bopes, R. Brandt, D. Fohrman, J. Hahn, T. Hart, C. Headley, J. Kopish, R. Maharry, J. Matson, K. Mohoff, R. Mraz, M. Palmer, L. Pena, B. Phillips, A. Rhodes, H. Rosman, C. Sievers, D. Tate, S. Tyrrell, J. Villarreal, P. Wiese, and A. Wignall, *The Reach of Politics via Twitter - Can That Be Real?* Open Journal of Political Science (2013).

[225] DeviantArt, *http://www.deviantart.com/,* (2014).

[226] Y. Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, *Analysis of topological characteristics of huge online social networking services,* in *Proceedings of the 16th international conference on World Wide Web* (ACM, 2007) p. 844.

[227] B. Klimt and Y. Yang, *Introducing the Enron corpus,* in *First Conference on Email and Anti-Spam (CEAS)* (2004).

ACKNOWLEDGMENTS

I thank the royal KPN N.V., TNO and the Technical University of Delft for funding me through the trans-sector research academy for complex networks and services (TRANS). In particular I am very grateful to my promotor Piet Van Mieghem for the opportunity to work together with him in a very interesting group at the TU-Delft and for interesting conversations. His enduring passion of science and his ability to spot the "beauty of research" had quite an impact on me.
Special thanks also to my copromotor Christian Doerr who has been a mentor in the last four years, always being able to patiently discuss any kind of problems I faced and pointing towards the right direction. It was a great pleasure to work with you.
I want to thank the members of my Ph.D. committee for their work and time spent on this thesis. I am honored by your attendance at this important event in my life.

Of cause this thesis would not have been possible without the endless support of my parents and my brother, I think there is no way to express the magnitude of my gratefulness in words.
I also need to thank Susanne Melde, who convinced me (at least partially) to pursue a Ph.D. in Delft, for doing so and supporting me ever since.
During the time in Delft I had the pleasure to meet quite some special persons who, for whatever reason, I now call friends. This extension of my own social network is interestingly not completely explainable by the content provided in this thesis but I am really, honestly, glad to have met you: Wynand Winterbach, Ruud van de Bovenkamp, Niels van Adrichem, Stojan Trajanowski, Jil Meier, Javier Martín Hernández, Daniela Vaman, Yann Dufournet, Igor Stepanov, Dajie Liu, Fernando Kuipers, Wendy Murtinu, Huijuan Wang, Nico Baken, Rob Kooij, Cong Li, Anteneh Beshir, António Madureira, Siyu Tang, Edgar van Boven, Evangelos Pournaras, Ebisa Negeri, Martijn van Egdom, Song Yang, Farabi Iqbal, Rogier Noldus, Marcus Märtens, Xiangrong Wang, Yakup Koc, Annalisa Socievole and Bo Qu.

Thank you all for making my time in Delft the best of all times.

Norbert Blenn
Delft, March 2014

CURRICULUM VITÆ

Norbert BLENN

26-03-1981 Born in Burgstädt, Germany.

EDUCATION

1992–1999 Gymnasium, Burgstädt

2001–2007 Technische Universität, Dresden
Study Thesis: Development and conception of Wikipedia as hyper-audio.
Supervisor: Dr. rer. nat. H. Donker

Publication: Gestaltung von Hyperlinks in einer Hyperaudio-Enzyklopädie
(design of hyper links in an hyperaudio-encyclopedia)
N. Blenn, H. Donker, Mensch und Computer 2007, Weimar
(M&C Forschungspreis für den besten Beitrag [Best Paper Award]).

Diploma Thesis: Entwicklung eines portablen Stereo Videoaufnahmesystems
für die Präsentation auf einer Stereoprojektionswand
(Development of a portable stereo video recording system
for presentation at an stereo projection screen)
Supervisor: Prof. dr. S. Gumhold

2007–2010 Research Assistant at Computergraphics and Visualization
Technische Universität, Dresden

Publication: A Tool For Automatic Preprocessing Stereoscopic-Video
N. Blenn, N. v. Festenberg, S. Gumhold, Electronic Imaging 2010, San Jose, USA

2010–2014 PhD candidate at Delft University of Technology
Network Architectures and Services Group

Thesis: Content Propagation in Online Social Networks
Doctoral Advisor: Prof. dr. ir. P. F. A. Van Mieghem
Supervisor: Dr. C. Doerr

LIST OF PUBLICATIONS

12. **C. Doerr, N. Blenn**, *Metric convergence in social network sampling*, 2013, Proceedings of the 5th ACM workshop on HotPlanet, Hong Kong, China.

11. **C. Doerr, N. Blenn, P. Van Mieghem**, *Lognormal Infection Times of Online Information Spread*, 2013, PLoS ONE 8(5): e64349. doi:10.1371/journal.pone.0064349.

10. **C. Doerr, N. Blenn, S. Tang, P. Van Mieghem**, *Are Friends Overrated? A Study for the Social Aggregator Digg.com*, Computer Communications, 35(7), pp. 796–809, 2012.

9. **N. Blenn, C. Doerr, N. Shadravan, P. Van Mieghem**, *How much do your friends know about you? Reconstructing private information from the friendship graph*, 2012, Eurosys 2012, 5th Workshop on Social Network Systems.

8. **D. Liu, N. Blenn, P. Van Mieghem**, *A Social Network Model Exhibiting Tunable Overlapping Community Structure*, 2012,1st International Workshop on Advances in Computational Social Science, June 4-6, Omaha, Nebraska, USA.

7. **D. Liu, N. Blenn, P. Van Mieghem**, *Characterizing the Structure of Affiliation Networks*, 2012, 12th International Conference on Computational Science (ICCS), June 4-6, Omaha, Nebraska, USA.

6. **N. Blenn, C. Doerr, K. Charalampidou, P. Van Mieghem**, *Context-Sensitive Sentiment Classification of Short Colloquial Text*, 2012, IFIP Networking 2012, May 21-25, Prague, Czech Republic.

5. **N. Blenn, C. Doerr, S. van Kester, P. Van Mieghem**, *Crawling and Detecting Community Structure in Online Social Networks using Local Information*, 2012, IFIP Networking 2012, May 21-25, Prague, Czech Republic.

4. **N. Blenn, C. Doerr, P. Van Mieghem**, *Content Propagation in Online Social Networks*, 2011, Poster, ICT.Open, November 14-15, Veldhoven, The Netherlands.

3. **P. Van Mieghem, N. Blenn, C. Doerr**, *Lognormal Distribution in the Digg Online Social Network*, 2011, The European Physical Journal B, Vol. 83, No. 2, pp. 251-261.

2. **S. Tang, N. Blenn, C. Doerr, P. Van Mieghem**, *Digging in the Digg Social News Website*, IEEE Transactions on Multimedia, Vol. 13, 2011, No. 5, October, pp. 1163-1175.

1. **C. Doerr, S. Tang, N. Blenn, P. Van Mieghem**, *Are Friends Overrated? A Study for the Social News Aggregator Digg.com*, Networking 2011, Valencia, Spain.