Christof Steinkellner, BSc

Empirical Analysis of Social Networks of Computer Scientists

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Software Development and Business Management

submitted to

Graz University of Technology

Supervisor

Ass.-Prof.Dipl-Ing.Dr.techn. Elisabeth Lex

Knowledge Technologies Institute
Head: Univ.Prof.Dipl-Ing.Dr.techn. Stefanie Lindstaedt

Faculty of Computer Science

Graz, December 2015

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources. The text document uploaded to TUGRAZonline is identical to the present master's thesis dissertation.

Graz, ______________________________ ________________________________

Date Signature

EIDESSTATTLICHE ERKLÄRUNG[1]

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

Graz, am ______________________________ ________________________________

Datum Unterschrift

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Abstract

Twitter is a popular online social network among researchers. Researchers use Twitter to exchange views and opinions on various topics as well as discussing and promoting new ideas and publications.
The experiments conducted throughout this thesis are based on a Twitter dataset created by computer scientists whose research area is known. The thesis can roughly be divided into four parts. The first part explains the acquisition of the Twitter dataset. The second part presents various statistics of the dataset. It is shown that most Tweets are created during working hours and the users' activity differs greatly. Furthermore, the network created by the users' follower relationships is shown to have small world features and the research area affiliations are visible therein.
The thesis' third part is dedicated to the investigation of hashtag usage within the dataset. An analysis of the hashtag usage showed that most hashtags are very seldom used. The popularity of hashtag usage seems to be constant, but there are changes on short-terms. By using the mapping of users to research areas, the hashtags were assigned to research areas as well. Thus research area specific and general hashtags could be identified.
The forth part is focused on the distribution of hashtags through the observed Twitter network. Information flow trees were proposed to represent the distribution of each hashtag. These information flow trees show, among other things, each user's information spreading efficiency and how often a user is at the start or end of an information cascade. These findings were tested to correlate with user attributes like the Tweet and Retweet count and the follower relationships. These results showed that information flow trees are influenced by user attributes, but the correlation is only strong in some few cases.

Keywords. Online social network analysis, Information diffusion, Science 2.0, Information cascades

Kurzfassung

Unter Wissenschaftlern ist Twitter ein sehr beliebtes soziales Netzwerk. Dort diskutieren sie verschiedenste Themen und werben für neue Ideen oder präsentieren Ergebnisse ihrer aktuellen Forschungsarbeit.
Die in dieser Arbeit durchgeführten Experimente beruhen auf einem Twitter-Datensatz welcher aus den Tweets von Informatikern, deren Forschungsbereiche bekannt sind, besteht. Die vorliegende Diplomarbeit kann grob in vier Teile unterteilt werden: Zunächst wird beschrieben, wie der Twitter-Datensatz erstellt wurde. Danach werden diverse Statistiken zu diesem Datensatz präsentiert. Beispielsweise wurden die meisten Tweets während der Arbeitszeit erstellt und die Nutzer sind unterschiedlich stark aktiv. Aus den Follower-Beziehungen der Nutzer wurde ein Netzwerk erstellt, welches nachweislich small world Eigenschaften hat. Darüber hinaus sind in diesem Netzwerk auch die verschiedenen Forschungsbereiche sichtbar.
Der dritte Teil dieser Arbeit ist der Untersuchung der Hashtagbenutzung gewidmet. Dabei zeigte sich, dass die meisten Hashtags nur selten benutzt werden. Über den gesamten Beobachtungszeitraum betrachtet ändert sich die Verwendung von Hashtags kaum, jedoch gibt es viele kurzfristige Schwankungen. Da die Forschungsbereiche der Nutzer bekannt sind, können auch die Bereiche der Hashtags bestimmt werden. Dadurch können die Hashtags dann in fachspezifische und generelle Hashtags unterteilt werden.
Die Analyse der Weitergabe von Hashtags über das Twitter-Netzwerk wird im vierten Teil mittels sogenannter Informationsflussbäume betrachtet. Aufgrund dieser Informationsflussbäume kann gemessen werden wie gut ein Nutzer Informationen verbreitet und erzeugt. Dabei wurde auch die Hypothese bestätigt, dass diese Eigenschaften von der Anzahl der Tweets und Retweets und der Stellung im sozialen Netzwerk abhängen. Jedoch ist dieser Zusammenhang nur in Einzelfällen stark ausgeprägt.
Schlüsselwörter. Soziale online Netzwerke, Informationsverteilung, Science 2.0, Informationskaskaden

Contents

List of Figures

List of Tables

1 Introduction

Using online social media platforms has become more and more popular. Social media platforms are used by a wide range of people, they are not restricted to people affiliated with a certain age group, ethnic group, profession or any other group like this. Usually, online social media platforms are used to discuss and share current topics, news and activities or to promote new ideas, products and publications. Social media platforms can be used to converse and collaborate with people living far away from each other. Thus it comes without surprise that social media platforms are used by scientists, because scientists from different countries often work together on joint projects.

In scientific surroundings social media platforms are used above all to discuss and to share new ideas or intermediate results and to promote recently published papers. Twitter has become one of the most important social media platforms and is already heavily used during scientific conferences (Wen, Lin, et al., 2014). On that account it can be assumed that Twitter is a favored online social media platform among scientists. Hence, several scientific papers are already focused on the usage of Twitter in scientific surroundings.

This thesis investigates how hashtags are distributed on Twitter through a group of computer scientists. In particular, a Twitter dataset, consisting of all Twitter activities by specific users during a two and a half year period, was created for this thesis (see section 4.2). The specific users are Twitter accounts, which are known to be owned by computer scientists (Hadgu and Jäschke, 2014) and whose fields of research are known (Pujari et al., 2015).

In this thesis, an information flow tree describes the sequence of users, who have adopted a specific hashtag. A link is introduced between two users if one user has adopted a specific hashtag the other user has used or introduced before in time. Only these users where included in our information flow trees if they were influenced by another user and if they reused the hashtag within a defined time frame (see section 3.2.2). It is assumed that hashtags are distributed over to the Twitter follower network. Hence, in an information flow tree, a user A can only influence another user B if user B follows user A. The time frame is chosen in a way that the hashtag's distribution through the Twitter network is more likely than the distribution through other channels. Note that for each hashtag usage, the timestamp is available in our dataset. Therefore, our information flow trees can be used to investigate the distribution and reachability of a single hashtag within a certain time frame. In order to fit the process of creating information flow trees to the underlying data, the process can be modified by several parameters such as the time frame or the allowance to use multiple timestamps per user.

The tree form was chosen for the information flow trees, because previous work suggests that trees represent cascades well. Gomez Rodriguez, Leskovec, and Krause (2010) stated that an information cascade's influence structure is given as a directed tree. As described by Leskovec, Singh, and Kleinberg (2006), information cascades can be regarded to be trees or near-trees. Sadikov et al. (2011) stated that influence cascades have to resemble trees since each user can only appear once in an action sequence. Following this statement, each user within an information cascade, but the first user, was influenced by at least one other user within the cascade.

Users, who create and spread hashtags frequently, were identified by analyzing the information flow trees of several hashtags. Based on the hypothesis that more active and better connected users are probably better information spreaders than others, the measurements taken from the information flow trees (see section 3.3) and the user's activity and connectivity were compared (see section 4.6.4). This experiment should show if information flow trees and thereby the distribution of hashtags are influenced by the involved users' activity and social status. Thus, a well-connected user might spread hashtags more efficiently than a user with hardly any friends or followers. Likewise, a very active user might be more often at the beginning of an information flow tree than a user who tweets seldom.

The social relationships on Twitter play an important role at the creation of information flow trees, because they only consider information passed to neighbors in Twitter's social network. Hence, if a user reads a hashtag in another user's timeline and is thus influenced to use the hashtag, this connection is only represented in an information flow tree if there is a social relationship between these two users. The social relationships within the used dataset are discussed in section 4.3. Further, the social relationships of users interacting with each other and the user activities are investigated as well. This chapter also contains statistics showing how many Tweets were created at which time and the distribution of Tweets over users.

Hashtags play an important role in this thesis. Therefore, their usage within the Twitter dataset is analyzed in section 4.5. There it is displayed how often and when which hashtags are used. For some selected hashtags, a visualization showing when the hashtags were used, is analyzed. As each user's research area affiliation is known, the hashtag usage per research area can be calculated as well. With this data, an assignment of hashtags to research areas is possible. This assignment and its results are discussed in this section, too. The different subsets of hashtags which are used for several experiments, like the creation of information flow trees, are also explained in this section.

1.1 Research Questions

In order to clarify the problems addressed in this thesis, three research questions were stated. These questions summarize the main problems addressed by this thesis and are explained in detail in the following:

RQ 1: Research area specific hashtags

> *Are there hashtags, which are only used by users affiliated with the same area of research?*
> Hashtags are usually associated with certain topics. Since topics can be specific for certain research fields and the research area is known for each user within the dataset, identification of area specific hashtags should be possible. Furthermore, some hashtags might be used only

by users of a few research areas and other hashtags might have no noticeable affiliation with any specific research area.

RQ 2: Representing information flow

How to represent the diffusion of hashtags on Twitter?
Twitter is used to propagate various kinds of information. This information diffusion should be captured in order to see which users are responsible for creating and spreading pieces of information. This representation of the information flow should be close to reality and adjustable for different types of information and different diffusion speeds.

RQ 3: Identifying preeminent user attributes

Which user attributes correlate with the user's role within an information cascade on Twitter?
For each Twitter user the number of activities and social ties can be calculated. Twitter activities are Tweets and Retweets, social ties are given by a user's friends and followers. These four features might be indicative for the user's role within an information cascade. These roles describe the user's effectiveness as information spreader and if they are information sources or sinks.

2 Related Work

Due to the fact that Twitter is a very well-known microblogging service, many scientific studies use Twitter data. These studies are not restricted to any field of research. For instance, Twitter datasets can be used in political science: Twitter data was used to predict the German federal election in 2009 and to analyze how Twitter is used during an election (Tumasjan et al., 2010). Larsson and Moe (2011) analyzed the Twitter users during the Swedish election campaign in 2010. So these papers show that Twitter data can be used to summarize the sentiment during an election campaign and predict the election's outcome. Furthermore, the Twitter usage by the members of the U.S. congress was analyzed (Golbeck, Grimes, and A. Rogers, 2010). Another example of Twitter data usage is presented in Abel et al., 2012. This paper introduces a system to detect and summarize incidents like fires or earthquakes. The Tweets are used to create a profile of the incident to enable real-time analytics. Mendoza, Poblete, and Castillo (2010) verified if Tweets are reliable if they are created during a natural disaster and how these Tweets got propagated through the Twitter network. While Twitter can be used for merchandizing, rapid product feedback, quick news distribution and other commercial uses, these two papers show that Twitter can be also used to enhance civil life.

This thesis contributes to the existing research, by adding analysis of the usage of Twitter by computer scientists. Hence, this chapter is based on papers which are focused on the usage of social networks like Twitter in scientific surroundings.

2.1 Empirical analysis of social network usage

In a first step, scientific studies analyzing the usage of social networks by scientists are introduced.

2.1.1 Scientific usage of Twitter

The papers listed in this section investigate why Twitter is used by scientists and the relation between Twitter and the academic world. The study described by Zhao and Rosson (2009) investigates the usage of Twitter in working environments. The study's results show that Twitter is quite popular in working environments as an informal communication medium used for collaborative work. Honey and Herring (2009) tested if Twitter is a good platform for conversation and collaboration. They found that Twitter is already used widely for collaboration and the usage will eventually improve with user interface adaptions. These two papers show that Twitter is used and might simplify collaborative work. As collaborative work is often required for scientific projects, this can be a reason why scientists use Twitter.

First, one has to know why scientists use Twitter. Priem and Costello (2010) attended to this question as well as to how scholars cite on Twitter. They interviewed a few scientists and analyzed their Tweets. The authors found that Twitter citations of publications are often indirect citations and thereby differ from traditional citations in papers. However, these Twitter citations are perceived faster and are also regarded as measurement for the scientific impact of a publication. The connection between Tweets, scientific impact and future citations is studied by Eysenbach (2011). By analyzing Tweets linking to a publication, he discovered that future highly cited papers can be predicted mere three days after their publication. While these two papers already investigated the usage of Twitter by scientists, they are more focused on the correlation of Twitter popularity of scientists and actual scientific impact than conversational patterns.

2.1.2 Twitter usage at scientific conferences

Scientists tend to use Twitter during scientific conferences to exchange a broad range of information. For instance, Twitter is used to announce speeches or to comment on current presentations. The papers presented in this section investigate Twitter datasets collected at scientific conferences. Thus, they are similar to this thesis by examining the Twitter activities of researchers.

García et al. (2015) analyzed the Twitter data of 26 Computer Science conferences. They classified the Twitter users into language groups in order to see which languages were spoken at the conferences and what interaction occurs between these language groups. They verified their assumption that English is the most popular language, but a significant proportion of conference attendees use other languages as well. Further they discovered that users speaking only English and English-Japanese bilinguals mostly interact with their own language community while users using other languages interact with users of different language groups more frequently. Different languages are not considered by this thesis, but for future work it would be interesting to solemnly use English Tweets.

Wen, Lin, et al. (2014) analyzed the usage and communication patterns on Twitter of 16 Computer Science conferences over 5 years. They found that over the analyzed years information sharing became more important. They showed that Twitter is considered as an information platform during scientific conferences. The Retweet ratio and the usage of URLs in Tweets also increases steadily, while the usage of replies and mentions stays more or less the same. This study is partly similar to the analysis presented in 4.3.3.

Another study focused on the Tweets created at four academic conferences was done by Wen, Parra, and Trattner (2014). Twitter users were manually classified into five groups. Thereafter, the communication across the groups was monitored in order to see which groups got most attention and who are the most important users in a group. While the groups are quite different from the research areas used in this thesis, the investigation of the group's behavior is slightly similar.

The usage of Twitter during conferences was also analyzed by Reinhardt et al. (2009). The authors discovered that Twitter is often used as an additional way to discuss current topics and to share complementary information. The information exchange is usually done in plain text or web links. The publication notes that an advantage of microblogging is that it is not limited to a common location, so virtually anyone can participate in discussions. Moreover, with the help of Twitter, conference attendees can also follow parallel sessions.

Tweets of scientific conferences were also explored by Weller, Dröge, and Puschmann (2011). In this paper, the authors focused on the analysis of Twitter citations. They discovered that the observed scientists cite by quoting statements in Retweets and by linking to external sources with URLs. The author's research showed that classical citations and Twitter citations have slightly different purposes. However, they say that the frequency of an URL could still be used as measurement for the referenced resource's impact.

The questions if Twitter use during a conference is beneficial for the attendees and the overall research process is attended to by Ross et al. (2011). The authors of this paper also describe the observed usage of Twitter during a conference. One of their findings was that Tweets with the same hashtag are not one distributed conversation, but multiple monologues and loosely joined dialogues. This behavior explains why there are often parallel information flow trees.

The listed publications show why and how Twitter is used during scientific conferences and discuss various characteristics of the used Twitter dataset. It is shown that Twitter usage during conferences is an accepted way to communicate, but the Twitter usage has changed over the years as well as the users and user groups. These studies are all similar in some way to this thesis, but the used Twitter dataset is quite different. While these papers use datasets created at relatively short, but repeated occurrences, this thesis uses a continuous dataset over a longer period.

2.2 Communities and users in social networks

The reasons for community creation in social networks are various. The topical and geographical properties of Twitter's social network were analyzed by Java et al. (2007). This study shows that users build communities by connecting with users with similar intentions. This section also presents publications, which investigated if Twitter users can be classified and how to measure a user's influence. The hypothesis that people tend to connect to people with similar interest was strengthened by Wu et al. (2011). The authors of this paper discovered that celebrities tend to follow celebrities, bloggers tend to follow bloggers and so on. Hence, researchers might prefer communicating with other researchers of the same field. Another finding of this paper was that the majority of contents is created by a minority of users.

2.2.1 Classify users

Rao et al. (2010) present an approach to classify user attributes like gender, age, political orientation and regional origin from Twitter data. The authors use a stacked-SVM-based classification algorithm to extract the user attributes. While the approach is quite different from this thesis, the authors wanted to correlate Twitter data with user attributes or behavior as well.

2.2.2 User influence

When analyzing social networks, preeminent users, who have considerable more connections to other users, are found frequently. These outstanding users often influence other users by acting as a role model or idol. Influence in social networks was already studied long before online social networks were invented. Examples of those earlier studies are E. M. Rogers (1962) and Gold et al. (1956).

However, social influence is not easily measured. Anagnostopoulos, Kumar, and Mahdian (2008) present a model which tries to create a correlation between user actions and social ties. With this model the authors want to describe social influence. B. Sun and Ng (2012) identified influential users by using only the user's postings. The users were classified into starters and connectors. Leavitt et al. (2009) split Twitter users into two groups identified by their Tweets which are either conversation-based or content-based. Celebrities tend to use Tweets for conversation purposes, whereas news media rather spread contents. The papers address a similar research question as chapter 4.6.4, where the user attributes are correlated with the user's effectiveness of information spreading.

Cha et al. (2010) argue that the influential Twitter users usually are characterized by a well above average number of followers, plenty of Retweets and a lot of mentions in various Tweets. However, the study presented in this paper shows that the sets of high ranking users per dimension (followers, Retweets, mentions) have little overlap. The authors think that this happens because the dimensions mean different things. The number of followers shows a user's popularity, a Tweet is usually often retweeted if its content is of great value and the number of mentions is a measurement for a user's name value. The paper also analyzed other measures for influence like the total number of Tweets or the number of friends, but these measures were set aside, because they identified spammers as most influential users.

The research presented by Weng et al. (2010) also focuses on the influence of users in the Twitter network. The authors of this paper invented a measurement for user influence called TwitterRank. This ranking method is an extension to the PageRank algorithm Page et al. (1999) and uses the topical similarity between users and the link structure to indicate a user's influence. Again, this paper might explain results gained from the information flow trees.

2.3 Information diffusion

Twitter can often be described as monologue, if used to express one's opinion on various matters or to share resources and news with the followers (Ross et al., 2011). However, Twitter also allows its users to interact with each other. The question how information is distributed over Twitter inspired many scientific papers. Some of these papers are discussed in this section.

The modeling of information diffusion is often crucial for analyzing information diffusion in real data. These models are usually activation sequences or spreading cascades (Guille, Hacid, et al., 2013). Gomez Rodriguez, Leskovec, and Krause (2010) created a model showing how users got influenced by their neighbors. In this model, the influence structure is also represented as a tree. Time plays an important role in this model as well, because if there is more time between two usages of the same piece of information, the probability that one user influenced the other sinks. However, this model is different from the information flow trees used in this thesis, because its focus is on the most probable way a piece of information took, while information flow trees represent all ways a piece of information took equally.

Sadikov et al. (2011) created information propagation cascades from the Twitter network and the blogosphere. These cascades are trees just like the information flow trees. The authors argue that each user can only appear once in an action sequence and all users, but the user at the cascade's root, were influenced by other users. Thus, the influence cascades are trees. According to Leskovec, Singh, and Kleinberg (2006) information cascades are created if users were influenced by other users within a social network to adopt a new idea or action. As described in this paper, these information cascades can be considered as trees or near-trees.

2.3.1 Topic diffusion

Twitter can be used to share news and opinions on current matters. Due to this usage, Tweets are often connected to current events. Consequently, researchers have already worked on this subject.

Zubiaga et al. (2011) present an approach to detect trending topics and to classify these topics into news, current events, memes or commemoratives. This classification relies on 15 features, which are language independent and characterized by the trending topic's social diffusion. Further on, only Tweets sent before the topic became 'hot' and no external sources are needed for the classification. Hence, the classification can be done as soon as the topic becomes viral. Instead of whole topics Leskovec, Backstrom, and Kleinberg (2009) tracked so-called memes across the web. These memes are distinctive phrases. In their paper the authors presented a framework to track these memes on the web.

Romero, Meeder, and Kleinberg (2011) analyzed the spreading of hashtags through the Twitter network. The evaluated hashtags are linked to quite different topics. Their paper's goal is to describe the diffusion of different kinds of information in a shared environment, while this thesis is focused on the diffusion of hashtags and the resulting user roles. The authors of the paper claim that Twitter is an ideal platform to observe the spreading of a broad range of topics. The spread of a hashtag does not just depend on the associated topic, but as well as some more subtle features, which are described in their paper.

2.3.2 Prediction of information diffusion

While analyzing the flow of information in a given dataset, the prospective information diffusion is the focus of several papers. Guille and Hacid (2012) present a model to predict the information spread of an topic on social networks. Another research dealing with information diffusion was done by Ma, A. Sun, and Cong (2012), who tried to predict the future popularity of a hashtag. Cheng et al. (2014) analyzed cascades on Facebook by tracking photos which got reshared and they were able to predict how cascades will continue to grow. The prediction of information diffusion is not part of thesis, but a prediction of future information cascades is basis for future work.

2.3.3 Information persistence

Often Twitter and other social media platforms are used to share information and resources. However, several shared resources become inaccessible after some time, because hyperlinks get invalid or resources were removed. In order to provide some figures on this loss of shared resources, SalahEldeen and Nelson (2012) explored the data loss on some social media datasets.

They found that after one year about 11% of all shared resources are lost and on each additional day 0.02% will vanish. The problem of disappearing online resources is not limited to social network. For instance, Sanderson, Phillips, and Van de Sompel (2011) study the number of valid links in papers and Ainsworth et al. (2011) investigate the number of archived webpages. Those studies are relevant for this thesis' crawling part, since some information might got lost over time.

2.3.4 Analyzing information cascades

Information cascades can be used in several ways. One possible usage is to check whether the social relationships are reflected in the information cascades. An analysis like this is also done in this thesis (see section 4.6). Kitsak et al. (2010) showed that the best connected users in social networks are not necessarily the best information spreaders, but users located at the network's so-called core. This analysis is quite different from the analysis of information flow trees conducted in this thesis, because the authors did not create any structures similar to information flow trees for their studies.

The distribution of advertisements in Sina-Microblogs (a chinese microblog) was analyzed by Yin et al. (2012). The authors of this paper discovered that news usually reach a wide audience efficiently, though celebrities do not influence advertisements as much as expected. The information cascade representation used in this paper consists of the reactions to a user's postings. This representation differs to information flow trees by allowing bidirectional edges and a different information distribution tracking approach. Arnaboldi et al. (2014) argue that the strength of the relationship between users affects the information diffusion. They used an information diffusion model which is based on the probability that a user influenced another user. Their paper suggests using the strength of social relationships when predicting which users are likely to start large information cascades.

Another way to examine information cascades is to look at the patterns of these cascades. This is done by Rattanaritnont, Toyoda, and Kitsuregawa (2012), who analyzed the cascade patterns for several topics on Twitter. Their results show that cascades created by hashtags from different topics have distinguishable patterns, but that observation also holds true for some hashtags on the same topic. The authors also noticed that the time of Tweets and the influence of a user on her or his followers are the best measures to distinguish cascade patterns. The cascades created for this study are quite similar to the information flow trees used by this thesis, but these cascades allow several starting points and the time between the usages is less important there.

Instead of explaining cascades by social relationships of the involved users, cascades can be used together with a graph of social relationships to detect communities of within the users. This hypothesis was investigated by Barbieri, Bonchi, and Manco (2013). They argued that a message cascade contains information about the user's communities as community structure is an important factor forming the cascades. In this publication, information cascades are not used directly, but within a more complex model.

3 Methodology

This chapter explains the experimental setting used for this thesis. At the beginning, the social media platform Twitter and associated terms are illustrated. The next section is about information flow trees. There, their purpose, properties and creation methods are explained in detail. The last section of this chapter covers the analysis of the information flow trees. There, the used approach to analyze information flow trees is explained. In this process, measurements for individual users and hashtags as well as measurements analyzing information flow trees collectively are computed.

3.1 About Twitter

Twitter[1] is a social media platform where users interact through so-called microblogs. Since this thesis is based on a Twitter-dataset, it is necessary to explain the most often used Twitter-terms and Twitter itself. Twitter's functionalities and features are described in this section, as well as the social relationships on Twitter.

3.1.1 Twitter Features

Twitter provides several features, which are discussed in combination with Twitter characteristics in this section.

[1] https://twitter.com/

Tweets Twitter is mainly a broadcast medium for microblogs. The microblogs used on Twitter are also called Tweets or statuses. Throughout this thesis, they will be called Tweets. Tweets are short textual status messages limited to a maximum of 140 Unicode characters. Attaching photos and videos to Tweets is also possible. If a Tweet got created, the Tweet appears on the timeline of all followers of the Tweet's creator. A Tweet does not consist solemnly of the text message, but there is also some meta-data for each Tweet available. This meta-data includes, among other things, the ID of the Tweet's creator, a unique Tweet-ID, the creation date, information whether the Tweet is a Retweet or reply, the mentioned users and the used hashtags.

Retweets If users want to share a Tweet of another user with their followers, the users create a so-called Retweet. A Retweet is a Tweet linked to another Tweet (the Tweet which got retweeted). Like all other Tweets, this Tweet can now be seen by all the user's followers.

Replies Replies are like Retweets special kinds of Tweets. They are responses to other Tweets. Therefore, they are linked to the Tweets, which they replied to.

Hashtags Hashtags are strings starting with #. No spaces or punctuation marks are allowed in hashtags. They are used to identify associated Tweets. There is no automatic tagging, but many events name a hashtag and users tend to include this hashtag, if they are tweeting about this event. The same holds true for a conversation about a certain topic where most related Tweets are tagged with the same hashtag.

Screen-names On Twitter users are usually identified by their so-called screen-name. This name is used if the user's Twitter account is addressed in a Tweet and always starts with a @ symbol. The Twitter name is the name of the Twitter account, this name is often the name or its abbreviation of the account's owner or the company running the account.

3.1.2 Social relationships on Twitter

Social relationships on Twitter are based on the follower-concept and are unidirectional relationships. A social relationship is created if some user A decides to follow user B. User A is then considered as B's follower and B as A's friend. So friends and followers on Twitter are just two views of the same data. Since the social relationships are unidirectional, user B does not have to follow user A in return. Usually user B takes no part in the creation of the relationship, because each user is allowed to follow almost every other user. However, some Twitter accounts are restricted so if B would be an access restricted user account, B would have to allow A to follow her or him. Users usually follow each other because if a follower relationship has been created, the following user gets all new Tweets of B delivered to her or his Twitter-timeline.

3.2 Information Flow Trees

In this thesis, information flow trees are created to model how information is distributed in a directed social network. The following section explains what information flow trees are and their purpose in detail. Thereafter two different ways to generate information flow trees are introduced. At the end of this section it is explained how information flow trees can be used to analyze user behavior, to compare users and to model spreading of information.

3.2.1 About Information Flow Trees

Information flow trees show how a piece of information gets distributed in a social network. Various pieces of information can be examined with information flow trees. For instance, the distribution of hashtags, topics, URLs or other things like that, which are used on social media platforms can be analyzed. In this thesis information flow trees are just used for hashtags on Twitter (**RQ 2**). Hence, in this thesis information flow trees always model how a hashtag travels from one user to another in a Twitter follower network.

In this setup, an edge of an information flow tree is created if some user uses a hashtag, which is afterwards used by one of the user's follower on Twitter. Since Twitter is usually not the sole information source of Twitter users, the link is only created if both users' Tweets were created within a certain time frame. If there is too much time between two Tweets, it is assumed that the information got distributed outside the Twitter network. Trees are chosen instead of directed networks, because there should be no circles in the information distribution. If a hashtag is used once, the hashtag is not forgotten within the chosen time frame. So if a user uses a certain hashtag several times within the time frame, it is assumed that only the first usage was influenced by another user. However, other users might get influenced by all hashtag usages of a certain user.

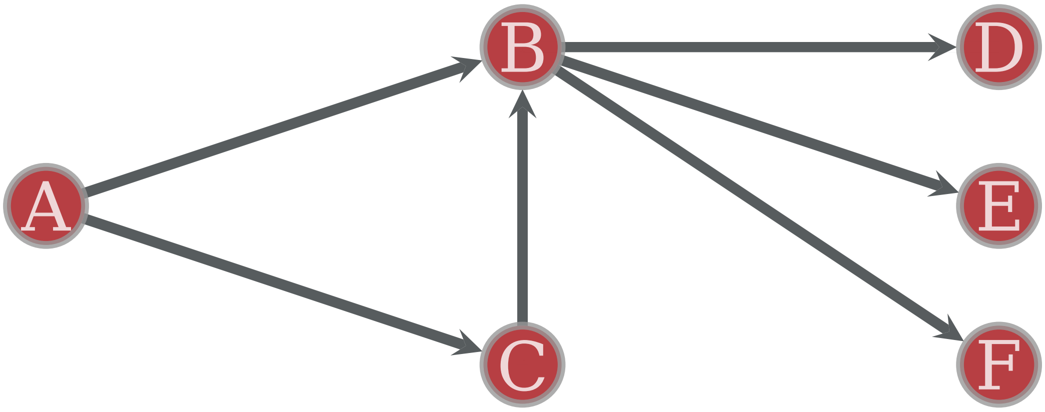


Figure 3.1: **Example of an information flow tree** - User A created the surveyed hashtag. Then A's followers C and B used the hashtag. Afterward B's followers D, E and F used the hashtag.

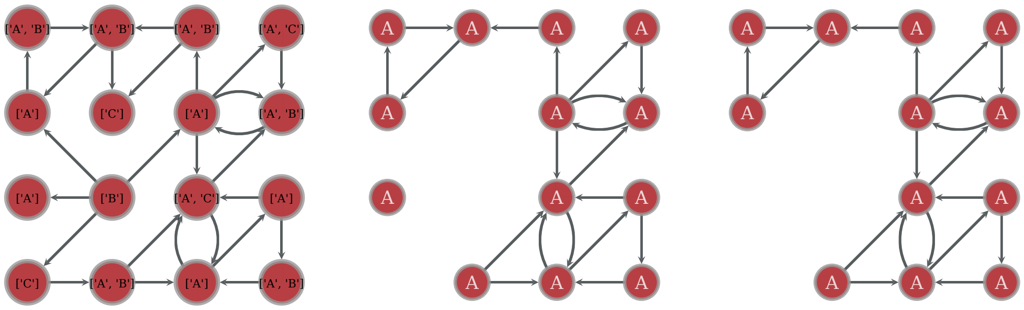
An exemplary information flow tree can be seen in figure 3.1. This information flow tree is created by the users A, B, C, D, E and F, who all used the same hashtag and are related to each other by several follower relationships. In order to create exactly this information flow tree, the users have to use the hashtag in an exact sequence. If the users had the same follower relationships, but used the hashtag in another sequence, most likely another information flow tree would have been created. Since the sequence of the hashtag usage is crucial, the time sequence has to be as follows: User A created a new hashtag or used a hashtag that was not used some time ago. Then C and thereafter B used the hashtag. After user C, the users D, E and F used the hashtag in some arbitrary sequence. In order to get the information flow tree as shown in the figure, there must exist several follower relationships between the six users. At least B and C have to follow A, B has to follow C and D, E and F have to follow B in order to create the shown information flow tree. However, there might be other follower relationships as well, which were not included in the information flow tree, due to the time sequence of hashtag usage. For instance, if A follows all other users, there will be no changes in the resulting information flow tree, but if F would follow A, there would be an edge from A to F.

3.2.2 Creation of Information Flow Trees

In the following subsection the creation of information flow trees is explained in detail. The creational process consists of two parts: The filtering of the underlying follower network and the creation of information flow trees based on the filtered network.

Filtering of Follower networks

To create information flow trees a follower network and data indicating when and which hashtag is used is needed. In this thesis, a follower network is used where each user-vertex has the used hashtags and their timestamps attached. The first step to create information flow trees for a certain hashtag is to filter the whole follower network, so that only the users, who used the hashtag, remain in the network. This process is illustrated in figure 3.2. First, all users, who did not use the hashtag *A* are excluded from the follower network. After this step there might be several users who are not connected to other users. These users can be removed from the network as well, because they cannot be a part of any information flow tree, since a follower connection is therefore required.



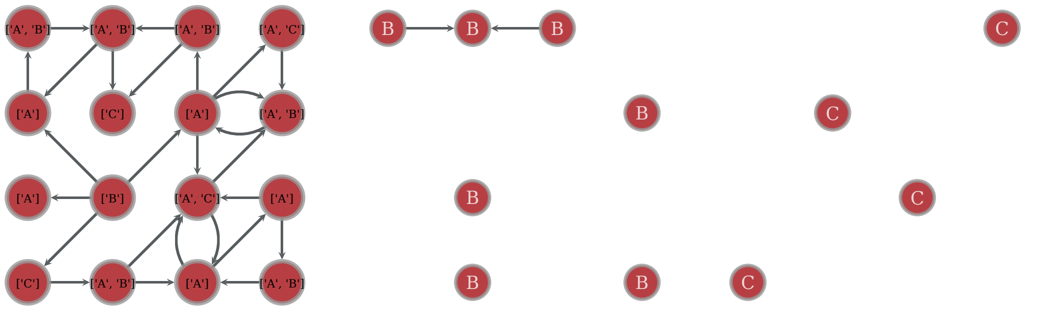
(a) Example follower network with used hashtags

(b) Filtered network so that only users who used hashtag *A* are shown

(c) Removed the nodes without any edges from the network shown in (b)

Figure 3.2: **Filtering a follower network by the used hashtags** - This figure shows the first step at creating information flow trees. Figure (a) shows an exemplary follower network. The node's labels show which tags were used by each user. In figure (b) the network is filtered in a way that only the users who used the tag *A* remain in the network. Since users without any connection cannot be a part of any information flow tree, these users can be removed as well. The resulting filtered follower network, which is used to extract information flow trees, is shown in figure (c).

The filtering of the network is mandatory for the following steps, but it also shows an unsolvable problem of the creation process: Even if there are several users, who used the hashtag, there might be few or even none information flow trees if the users do not follow each other. In figure 3.3 this problem is illustrated. There, the same network as in the previous example is filtered for the hashtags *B* and *C*. The resulting network for *B* is very small and for *C* there exists no network at all. This problem will probably arise when creating information flow trees for general hashtags with few usages.



(a) Example follower network with used hashtags

(b) Filtered network so that only users who used hashtag *B* are shown

(c) Filtered network so that only users who used hashtag *C* are shown

Figure 3.3: **Nodes get removed from follower network when filtered for specific hashtags** - This figure shows a problem of the creation of information flow trees. If the network shown in figure (a) is filtered for hashtag *B*, the remaining network is very small (see figure (b)). If the network is filtered for hashtag *C*, all connections are severed, as can be seen in figure (c)

Creation of Information Flow Trees from filtered follower networks

This section explains how information flow trees are extracted from filtered follower networks. In order to explain the process in detail, an exemplary filtered follower network is used. This network can be seen in figure 3.4. The users are distinguished by their positions in the network plot. Through all following examples in this section the position of each user remains the same. The node's labels show when each user used the hashtag. When working with real data, the time information is stored by UNIX-timestamps and the maximum time difference between two usages is given in seconds. Here, the timestamps just indicate the succession of the hashtag usage. The time difference between each successive usage is always the same. The example consists of a follower network with a single component, but in real data, the follower network could be split in several components as well.

Two slightly different methods to create information flow trees were used. Which of these methods fits the ongoing analysis better, is unknown at the beginning. Therefore, both are explained in the following.

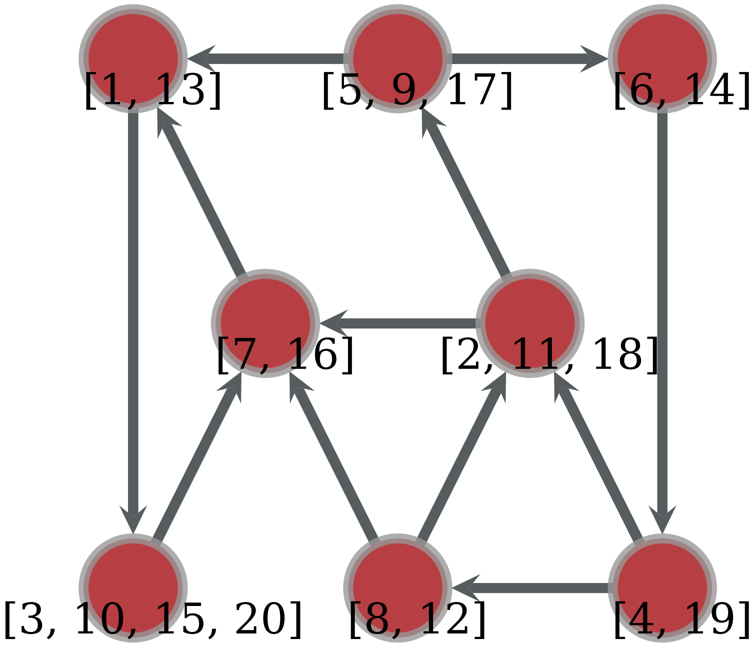


Figure 3.4: **Example of filtered follower network with timestamps** - Each node in this figure represents a user. The node's labels indicate when a user used the hashtag. The edges indicate the follower relationships. For instance the user positioned on the top-left follows the user positioned at the bottom-left. The displayed network is not based on any real network.

Creation of information flow trees for a hashtag with one allowed timestamp per user Information flow trees represent the information distribution in a network. Each edge represents a passing of information from one user to another. On Twitter, this information passing is for instance done if a user uses a certain hashtag and one of the user's followers reads this hashtag and uses the hashtag, too. However, if there is too much time between these two usages, the second user might have been influenced another way to use the hashtag. Hence the time between two hashtag usages is crucial for creating information flow trees.

Summing up, to create an information flow tree the following data is needed:

- A follower network only containing users who used the hashtag
- Data showing when a certain user used the hashtag
- A time frame. This time frame indicates how much time can pass between two usages of the hashtag, so that it is highly likely that the information was distributed through the Twitter network and thus can be added to the resulting information flow tree.

The first step at the creation of information flow trees is to determine which users can serve as starting points in the resulting trees. Therefore, for each hashtag usage, a timestamp-user-pair is created. So for each hashtag usage the user and the usage time is known. A timestamp-user-pair, which is at the beginning of an information flow tree, has to fulfill the following conditions:

- The user must not have used the hashtag before within the given time frame
- The user must not have a friend who used the hashtag before her or him within the time frame

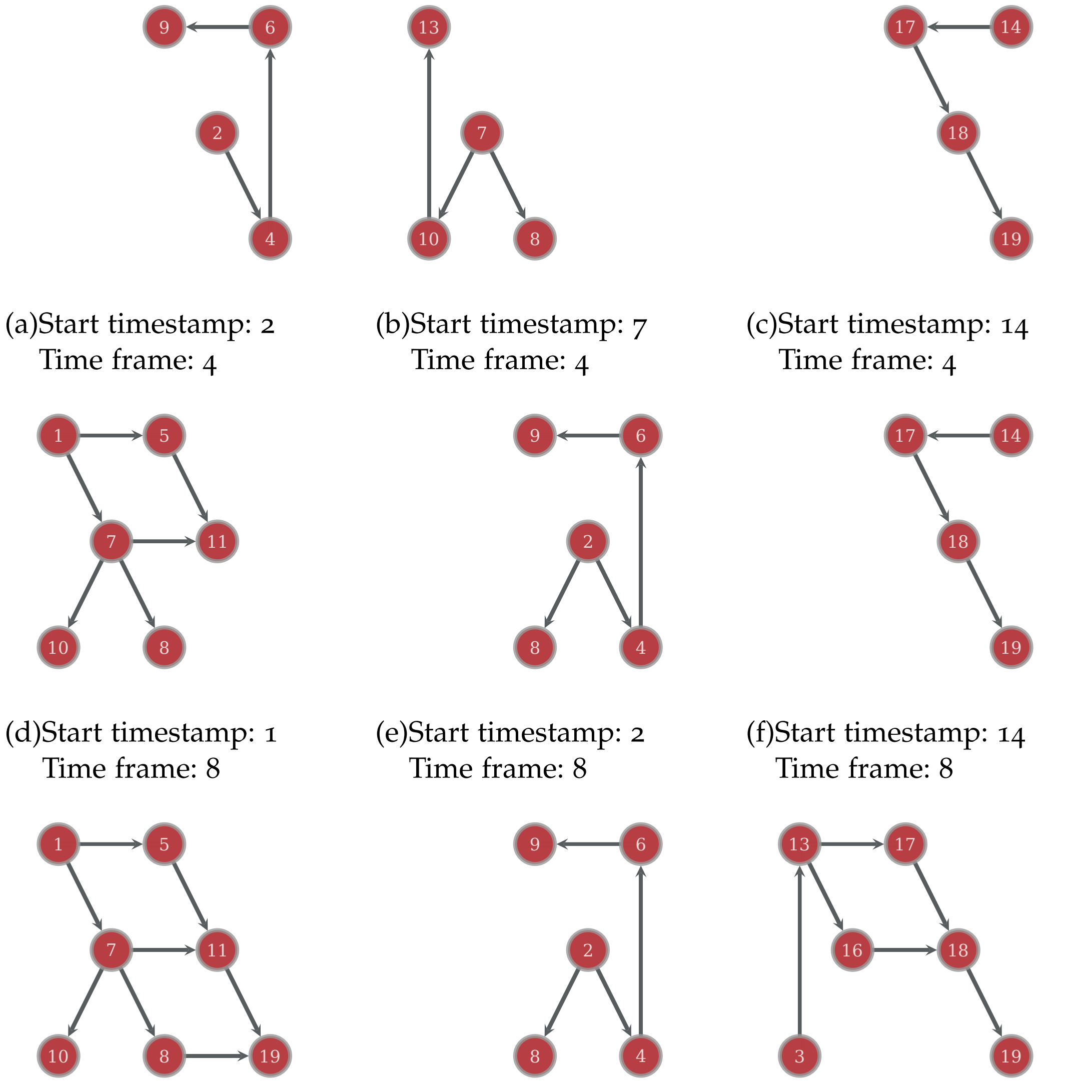
For each timestamp-user-pair that fulfills these conditions, an information flow tree gets created. The tree is created by adding new timestamp-user-pairs to its leafs. Beginning at the tree's leaf with the timestamp-user-pair with the lowest timestamp, all unused timestamp-user-pairs are tested if they could be added to the information flow tree. New timestamp-user pairs are only allowed to be added to the tree's leafs if they fulfill the following conditions:

- The new pair's user must not be in the tree already. There can be just one node and thus only one timestamp for each user in each tree
- The new pair's user has to follow the leaf's user
- The timestamp of the new pair has to be greater than the leaf's timestamp
- The time difference between the new pair's timestamp and the leaf's timestamp must not be greater than the given time frame

If there are no pairs which can be added, the creation process proceeds to the leaf with second lowest timestamp. This process is repeated until there are no tree ends, where a new timestamp-user end can be added. Thereafter the information flow tree is finished. However, if the tree consists of a single node, the tree is disregarded.

Figure 3.5 shows the information flow trees for the network displayed in figure 3.4. For this example, the time frames four, eight and twelve were used. A time frame of four means that, for instance, after the 1st timestamp, only the 2nd, 3rd and 4th timestamps are within this time frame. In the sub-figures 3.5a to 3.5c all possible information flow trees for a time frame of four successive hashtag usages are shown. Three resulting trees seem quite few, but they are the only allowed trees for the rules defined above. For instance, the information flow trees starting at the 1st, 3rd, 5th, 15th, 19th and 20th timestamps consist of just a single node and are thus disregarded. The users with the 4th, 6th and 8th timestamps are already parts of other trees with these timestamps and therefore not allowed to start their own trees. The same holds true for the 9th, 10th and 13th timestamp-user-pairs. However, these are special cases, because if a information flow tree would start there, the tree would continue in different directions than in the already existing trees. For instance, the tree starting at 9th hashtag usage would create this tree: $(9 \rightarrow 11 \rightarrow 12)$, but this would be a part of a circle, because the hashtag was already distributed over this way (see figure 3.5a). The same problem would occur, if a tree would start at the 10th or 13th hashtag usage.

The information flow trees created with an allowed time frame of eight successive timestamps are not so different from those where only four successive timestamps were allowed, the trees just get bigger and sometimes have different starting points. For instance the tree shown in 3.5b is a subtree of 3.5d and 3.5a is a subtree of 3.5e. The trees created with a maximum of twelve allowed successive timestamps are also not so different from those created with a time frame of eight allowed successive timestamps. The information flow trees shown in 3.5g and 3.5h are just bigger and the tree shown in 3.5i changed somewhat more compared to 3.5f.



(a)Start timestamp: 2
Time frame: 4

(b)Start timestamp: 7
Time frame: 4

(c)Start timestamp: 14
Time frame: 4

(d)Start timestamp: 1
Time frame: 8

(e)Start timestamp: 2
Time frame: 8

(f)Start timestamp: 14
Time frame: 8

(g)Start timestamp: 1
Time frame: 12

(h)Start timestamp: 2
Time frame: 12

(i)Start timestamp: 3
Time frame: 12

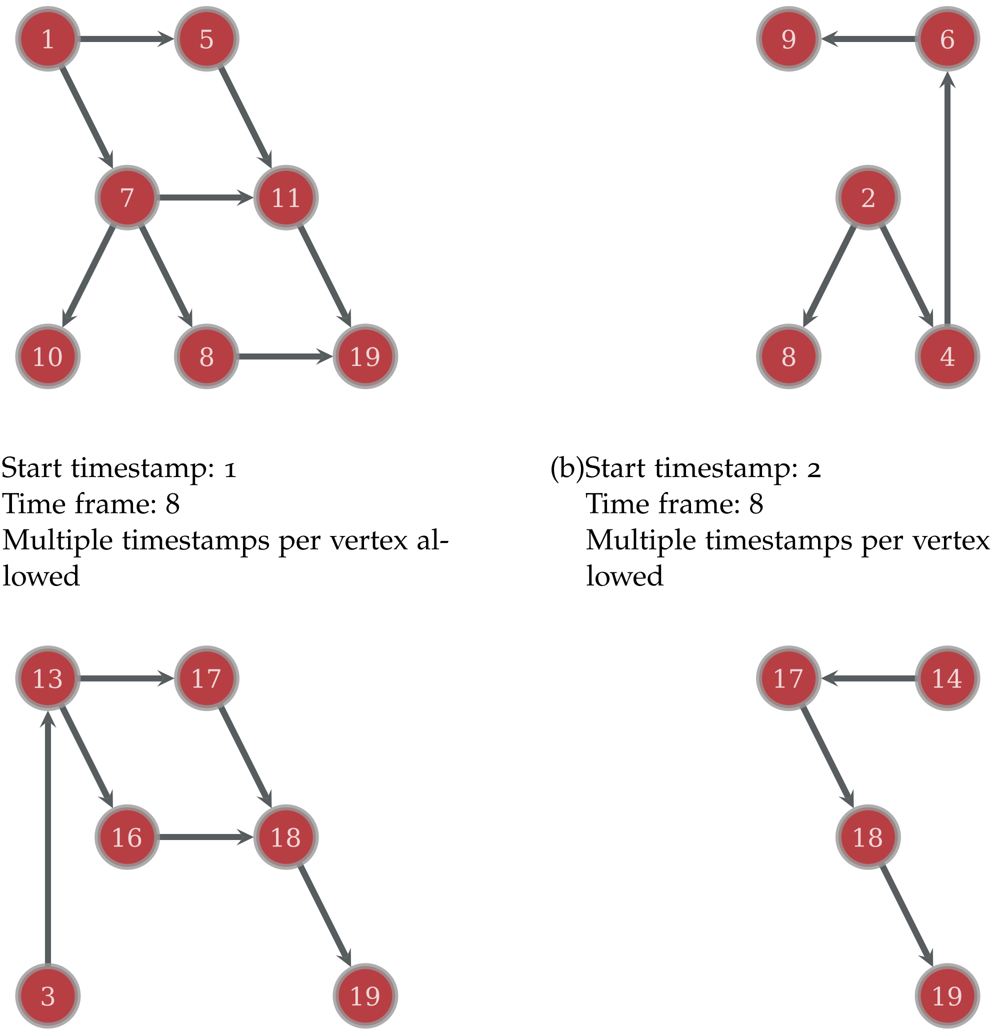
Figure 3.5: **Information flow trees extracted from the example network** - The figure shows all possible information flow trees for the example network with time frames of 4, 8 and 12 allowed successive timestamps. The maximum allowed successive timestamps is 4 for figures a to c, 8 for figures d to f and 12 for figures g to i. Other information flow trees are not possible for the given time frames, since either the trees consists of just a single node or the resulting tree would be a sub-tree of another tree

Creation of information flow trees for a hashtag with multiple allowed timestamps per user The creation of information flow trees for a hashtag with multiple allowed timestamps differs slightly from the creation were only one timestamp per user is allowed. The only difference between these two methods is that the tree's leafs, where new timestamp-user pairs are allowed to be added can have multiple timestamps. This is only possible if the user has used the hashtag once more within in the time frame. This change should disregard less Tweets of a user. For instance, in the network displayed in figure 3.4 the user at the top center has used the hashtag at the 5th, 9th and 17th occasion. If this user would be included into an information flow tree with the 5th hashtag usage, the user could never influence the user at the right center if the allowed time frame was four, because the 11th and 18th hashtag usage are not within the allowed time frame. However, this user also used the hashtag at the 9th occasion, which is within a time frame of four. This 9th hashtag usage could influence the user at the right center to using the hashtag at the 11th occasion. Hence, with multiple allowed timestamps, the user with the 5th hashtag usage could influence the user with the 11th hashtag usage within an allowed time frame of four.

To explain the difference of these two methods, an example of this creation method is shown in figure 3.6. In this figure the resulting information flow trees for the network from figure 3.4 with an allowed time frame of eight and multiple allowed timestamps per user are shown. The information flow tress shown in 3.6b and 3.6d are identical to those shown in 3.5e and 3.5f.

Compared to figure 3.5d, in figure 3.6a only the bottom-right vertex is added, although the time difference to its neighbors is greater than eight. With the new addition of several allowed timestamps per user, this is allowed. The user with 8th hashtag usage also used the hashtag at the 12th occasion, which is within the allowed time frame. Consequently, the user probably just reused the hashtag without further extrinsic influence. The 12th and 19th timestamps are also within this allowed time frame, so one user probably influenced the other user for using the hashtag. The same holds true for the user, who used the hashtag as 11th and again as 18th, which allows him to connect to the user with the 19th timestamp.

The information flow network shown in 3.6c is an entire new tree. Without the permit to use multiple timestamps per user, the user at the bottom-left



(a)Start timestamp: 1
Time frame: 8
Multiple timestamps per vertex allowed

(b)Start timestamp: 2
Time frame: 8
Multiple timestamps per vertex allowed

(c)Start timestamp: 3
Time frame: 8
Multiple timestamps per vertex allowed

(d)Start timestamp: 14
Time frame: 8
Multiple timestamps per vertex allowed

Figure 3.6: **Information flow trees extracted from the example network with multiple allowed timestamps per vertex** - The difference of the information flow trees shown here to those in figure 3.5 is that here multiple timestamps per vertex are allowed. Differences can be seen in the subfigures (a) and (c). In these figures there are edges between timestamps whose difference is bigger than the time frame of eight allowed successive timestamps

would have never qualified as a starter, because the user was influenced by the user, who used the hashtag at the 7th and 16th position. However, using this method the user can create a information flow tree starting at the 3rd timestamp, which is only possible with the other method if the time frame is greater than ten allowed successive timestamps.

Hence, the permit to use multiple timestamps per user should provide more and maybe longer information flow trees. Probably this method also models the information flow a little better than the other method, because with this method the first three users using the hashtag start their own information flow trees.

Excluding weekends from the time frame An analysis of the number of Tweets created per weekday (see figure 4.4a) shows that during weekends considerably less Tweets were created than during workdays. Hence, weekends could be excluded when calculating the time difference between two hashtag usages. For instance, if one usage was on Thursday and another on Monday, with a time frame of three days, the time difference would be too great. However, if the weekend is excluded from this calculation, the time difference would be within the time frame of three days. The weekends can be excluded in all described information flow tree creation methods.

The approach to disregard the weekends when calculating the time difference seems to be reasonable as there is less activity on the weekends, but there are drawbacks as well: For short time frames, the excluding of weekends might distort the results. Besides some hashtags might have more or equal activity on weekends while other hashtags might have none activity during weekends, so a global excluding of weekends might distort the results even more.

3.3 Analysis of Information Flow Trees

If many information flow trees got created for a great number of hashtags and Tweets, analyzing all those trees manually in an efficient way is very unlikely. Therefore, several measurements to aid the information flow tree analysis are used. They are explained in the following. The measurements are sorted into measurements linking to users or hashtags.

3.3.1 Information Flow Tree measurements linked to users

This section discusses the information flow tree measurements linked to users. All these measurements are calculated together and saved into the user dataset (see section 4.1.2). These measurements could be calculated on their own as well, but this approach saves time and gives a better overview on the resulting data.

Probably, the measurements discussed in the following correlate with other properties of the user dataset (see section 4.1.2). For instance, a user who got retweeted very often, might be a good information spreader. In order to calculate the correlation between two user properties like this, the spearman rank correlation is used (Spearman, 1904). This correlation is used because it works for data with arbitrary distribution. The spearman correlation is calculated by the following formula (d is the difference of the ranks for one observation):

$$r_{Sp} = 1 - \frac{6 * \sum_{i=1}^{n} d_i^2}{n * (n^2 - 1)}$$

Users frequently appearing in information flow trees A user who appears in many information flow trees is probably an important user no matter at which position.

Users frequently starting information flow trees A user starts an information flow tree if the user introduces a new hashtag or uses a hashtag which was not used for a longer time period and at least one of the user's followers has to use the same hashtag within the given time-range afterwards. Hence, users starting a significant number of information flow trees are most likely important users in the observed community. These users are also referred to as information sources.

Users frequently at the end of information flow trees Users at the end of an information flow tree used a hashtag within the given time-range after at least one of their friends, but none of these users' followers used the hashtag within the given time-range. Within the given time frame, these users are the last users who reliably noticed the hashtag. Being at the end of an information flow tree can have various reasons, so a user at the end does not have to be unimportant. However, if a user is only found at the end of information flow trees, the user can be considered as an information sink.

Users acting as information spreaders An information spreader is a user who spreads information. So every user in an information flow tree, except those at the end, can be considered as an information spreader. However, the goal of this measurement is to identify how efficient a user spreads information. Two ways to put a user's information spreading efficiency to numbers are used:

One way is to look at both the in- and out-links of each vertex of the information flow tree. To get a comparable score the out-links are divided by the in-links, but this also means that all sources have to be excluded from this measurement, because there would be a division by zero. If this measurement is used, users with higher scores are considered better information spreaders.

However, one might argue that the number in-links have no meaning when measuring the information spreading efficiency. So a user with 100 in- and 50 out-links is probably a better information spreader than a user with 2 in- and 5 out-links. The previously explained measurement above would prefer the second user with a score of 2.5 over the first user, who just scored 0.5.

Hence, another way to measure a user's information spreading efficiency is to look at the users' out-links exclusively. If measured this way, only the number of out-links are counted and the information sources and even the information sinks can be considered as well.

In order to compare a user's information spreading importance, the average score over all information flow trees is used. This way the scores are easily exchangeable. If the second measurement is used, the scores of users at the start or end of information flow trees can be added to the mean value as well.

Information flow tree length If a well-known user is within an information flow tree, it might become bigger than usual. In order to put this variation to numbers, the sizes of all information flow trees which can be connected to a user are compared. An information flow tree is connected to a user, if the user is within the tree. To measure the tree's size, the number of edges, vertices and the pseudo diameters[2] can be used. Since a user's impact on the information flow tree size might be associated with the user's position in the information flow tree, the comparison can be done for users at different positions separately. So the information flow tree sizes can be calculated with certain users at the begin, at the end, neither at the begin nor at the end and at an arbitrary position on the information flow tree. In order to get comparable values for all users, for each user the mean tree length of all trees, where the user is present, is used.

3.3.2 Information Flow Tree measurements linked to hashtags

This section explains the information flow tree measurements linked to hashtags. All measurements explained here are calculated together and saved into the hashtag dataset (see section 4.1.3). All these measurements could be calculated individually, but this approach gives a better overview

[2] https://graph-tool.skewed.de/static/doc/topology.html#graph_tool.topology.pseudo_diameter

from the resulting data, saves time and supports to create visualizations afterwards without recalculating the measurements.

Information flow tree length This measurement is similar to the measuring of the user's information flow tree length. Again either the number of edges, vertices or the pseudo diameter can be used for comparison. This measurement should show if some hashtags tend to create bigger information flow trees or multiple smaller ones. In order to compare the hashtags, the mean length of all trees belonging to a hashtag can be used.

Information flow tree duration The duration of each information flow tree belonging to a hashtag is analyzed by this measurement. The duration is the elapsed time between the first usage of the hashtag covered by the tree (always the source) and the last usage (done by one of the sinks). This measurement should show how long a hashtag stays active. The use of this measurement is probably only reasonable if the information flow trees were not created by partitioning the active time.

4 Experiments and Results

This chapter explains the results to the experiments discussed in the previous chapter (see chapter 3). The chapter's first three sections are about the used dataset. At first, the dataset's structure and contained properties are explained. Thereafter, the dataset acquisition is discussed covering the crawling of the Twitter-API and the post-processing of the crawler's results. The third section is about the dataset characteristics. There, social relationships within the dataset, the users' activity and their interactions are discussed. Thereafter follows a short section describing the research area affiliation of the users and which share of Tweets was created by users affiliated with certain research areas. The next part is dedicated to the analysis of hashtags. Hence, this part discusses the usage of hashtags and the questions if the hashtag usage changed over time and if hashtags can be assigned to certain research areas. The chapter's last section is focused on the analysis of information flow trees. The optimal information flow tree creation parameters are discussed, as well as the resulting information flow trees for some selected hashtags. At the section's end, the third research question (Whether user attributes can influence information flow tree measurements) is attended to.

4.1 Dataset

The dataset used for this thesis consists of the attributes of Tweets and attributes of Twitter-users. A crawler was used to create a dataset of Tweets and users, which is used throughout this thesis. This crawler is described in detail in section 4.2.2. First, a description of the Tweet and user attributes is given. Then, the dataset of hashtags is explained. This dataset is used to save results referring to individual hashtags.

4.1.1 Dataset of Tweets

Each Tweet-record in the dataset consists of the attributes listed in the following. All these attributes are obtained by the Twitter crawler and the dataset of Tweets is neither changed nor amended at any other point.

- A unique Tweet-ID
- The Tweet's text
- Date and time of the Tweet's creation
- User-ID linking to the Tweet-creator
- Retweet-Information:
  - How many times the Tweet was retweeted in total
  - The IDs of the Retweets contained in the dataset
  - If the Tweet is a Retweet contained in the dataset, the ID of the retweeted Tweet
- If the Tweet is a reply to another Tweet, the ID of that Tweet
- A list of hashtags used in the Tweet (can be empty)
- A list of users, represented by their ids, mentioned in the Tweet (can be empty)

4.1.2 Dataset of users

The dataset of user attributes is not exclusively created by the Twitter crawler, because analysis results linked to individual users are also added to the dataset. The adding of the results is not absolutely necessary, but it saves time when creating views and visualizations of the already calculated results and it also provides an overview. The following listing shows the attributes saved in the user dataset and where they were created:

- Attributes obtained with the Twitter crawler:
  - A unique user-ID
  - The name and the screen-name of the user
  - The location string provided by Twitter
  - Follower and friend data:
    - The total number of friends and followers

        - A list of friends, which are in the dataset
        - A list of followers, which are in the dataset

- The user's research areas (see explanation in the following)
- Attributes calculated from the dataset of Tweets:
    - The count of Tweets associated with the user
    - The number of mentions by other users
    - The count of users who mentioned this user
    - The number of times the user mentioned other users (if there is more than one mention in a Tweet, each mention is counted separately)
    - The count of users mentioned by this user
    - The number of retweets created by this user
    - The count of users retweeted by this user
    - The mean count of retweets by users from the dataset each of the user's Tweets got
    - The mean count of retweets each of the user's Tweets got (all over Twitter)
    - The number of users who retweeted this user
    - The count of replies
    - The number of users the user replied to
    - The total number of hashtag usages
    - The count of individual used hashtags
- Attributes calculated by the information flow analysis (see section 3.3):
    - The number of information flow trees containing the user
    - The number of information flow trees where the user was found at the source
    - The number of information flow trees where the user was found at the sink/end
    - The information spreader measurements for each information flow tree containing the user
    - The information flow tree length measurements for each tree subdivided by the user's position in the tree.

The users' fields of research were obtained from Pujari et al. (2015). This paper provides a way to map computer scientists to their fields of research (all possible research areas and their acronyms can be seen in table 4.1).

Acronym	Computer Science research area
AI	Artificial Intelligence
ATH	Algorithms & Theory
CA	Computer Architecture
CB	Computational Biology
CDP	Concurrent, Distributed & Parallel Computing
CG	Computer Graphics
CN	Computer Networking and Networked Systems
DM	Data Management
ED	Education
HCI	Human-Computer Interaction
OS	Operating Systems
PL	Programming Languages
SE	Software Engineering
SNP	Security & Privacy

Table 4.1: **Research areas and their acronyms** - taken from (Pujari et al., 2015)

The users' affiliations to their research areas were calculated in two steps. First, the researchers were mapped to conferences by analyzing their publications and then the conferences were assigned to research areas. Thereof, the research area affiliations of the scientists were calculated. Thus, each researcher is assigned to at least one research area. The affiliation to an area is given in percent and the sum of all affiliation always sums up to 100 %.

4.1.3 Hashtags

A dataset of hashtags was also created to provide an overview and quick access to the measurements linked to individual hashtags. The properties stored for each hashtag and how the properties were obtained are explained in the following:

- Properties obtained from the Tweet and user dataset:
  - The name of the hashtag (always stored without #)
  - The number of usages for each user
  - The timestamps indicating when the hashtag was used

  - A mapping indicating how often a hashtag was used by a user belonging to a certain research area (if a user belongs to 60 % to area A and to 40 % to area B and uses the hashtag, 0.6 is added to the area A count and 0.4 is added to the area B count)
  - A mapping which shows the number of times the hashtag was used together with a certain other hashtag in the same Tweet
- Properties calculated by the information flow analysis (see section 3.3):
  - The number of information flow trees created for the hashtag
  - The information flow tree length measurements for each of the hashtag's information flow trees
  - A list of the duration of each of the hashtag's information flow trees

4.2 Dataset acquisition

The dataset used for this thesis contains almost all Tweets in a certain time period of 5 578 Twitter-accounts. These accounts were selected because they are in a list of Twitter-accounts[1], which were verified to be owned by computer scientists. This list is the result of the research done by Hadgu and Jäschke (2014). They identified computer scientists by a machine learning approach. Thereby, a set of Twitter-accounts probably owned by scientists were classified into researchers and non-researchers.

Some of these Twitter-accounts are inaccessible, so they are omitted from the list of Twitter-accounts which are crawled. The list of Twitter-accounts was further reduced by allowing only Twitter-accounts of scientists whose research area was known. The mapping of researchers to their fields of research was done with the results from Pujari et al. (2015) (see section 4.1.2 for a more detailed explanation of the research areas).

The time period for the Tweets starts on 1st January 2013 and ends on 10th May 2015. I tried to get all Tweets of the named Twitter-accounts in this time period, but still some Tweets are missing in the dataset due to the

[1] https://github.com/L3S/twitter-researcher/blob/master/data/candidates_matched.tsv

limitations of the Twitter-API. This section explains how the dataset was acquired from the Twitter-API. First, the limitations of the Twitter-API are described, then the actual crawling part is explained and at the end, the cleaning processes of the dataset are shown.

4.2.1 Twitter-API Limitations

The Twitter-API[2] is limited in several ways to prevent misuse and free commercial use of Twitter data. These limitations have various impacts on the crawling task and the resulting datset and are explained in the following.

The most obvious API limitation is the so-called rate limit. All API requests are limited to a certain number of requests per fifteen-minute-window. There is an own limit for each request, so if a request is blocked due to the rate limit, the other request limits are not affected. For an overview of the different rate limits please visit `https://dev.twitter.com/rest/public/rate-limits`. The rate limit is not a big problem for the dataset acquisition, because only the computing time increases, but the resulting dataset is not affected directly. However, if some Twitter data is changed during the execution of the crawler, these changes are not recognized, because Twitter only provides the current state, but no historic data. So if for instance a user profile got changed, there is no straightforward and easy way to determine the profile's previous state and when the profile got changed. This behavior can lead to a dataset inconsistency, if a user stops following another user during the crawling process, the relation might appear for one user, but not the other user if their social relationships are crawled at different times. Hence, the resulting dataset might include some users who are not followed by their friends. This error can be fixed by restoring these missing relations.

Another problem at crawling Twitter data is that some Tweets and Twitter-user-accounts are not accessible. There are two reasons for inaccessible data on the Twitter API. Often the requested data was deleted and therefore got inaccessible, because the Twitter API provides no historical data. Since there

[2] `https://dev.twitter.com/rest/public`

is no simple way to circumvent this problem, all deleted data cannot be included in the crawled dataset. The other reason for inaccessible data is that some Tweets or Twitter-accounts are not publicly available. Again there is no straightforward way to get this data, so this data is omitted from the dataset, too.

The Twitter API limits several requests to the most recent data. For instance, the search request does not provide results older than a week (Twitter, 2015). However, of all these limitations, the limit to get only the 3 200 latest Tweets of a user is the most important. While most users created less than 3 200 Tweets in the search period (see section 4.3.3), the older Tweets of these most active users in the dataset are missing. Currently there is no way to get these missing Tweets with the Twitter API without payment.

4.2.2 Twitter crawler

The Twitter crawler is used to create a dataset of Tweets and Twitter-user accounts which are used in this thesis. The section's first part describes the crawler used for this thesis and its parts. Thereafter, the crawling strategy is described.

Implementation details of the Twitter crawler

Each of the following paragraphs explains one part of the Twitter crawler and its purpose in detail.

Connection to the Twitter-API The connection to the Twitter-API implements all used Twitter-API-requests. The main task of this part is to query the Twitter-API with specific requests and parse the responses. In order to reduce the impact of the Twitter-API's rate limit as much as possible, this connection allows using several credentials at the same time. The Twitter-API differs between user and application authentication which count the rate-limit separately. Some API-requests also have different rate limits for the two authentication methods. Hence, the Twitter connection allows to use both user and application authentication at the same time. The implemented Twitter-API requests are listed here:

- **GET statuseslookup**
  Takes a list of Tweet-IDs for input and returns all available Tweet data for these IDs
- **GET statusesshow:id**
  Takes a single Tweet-ID for input and returns the available data for this Tweet
- **GET userslookup**
  Takes a list of User-IDs or Twitter-screen-names for input and returns all available information of the user accounts referenced by these IDs or identified by these screen-names (For crawling the dataset only the search for Twitter-screen-names is used)
- **GET usersshow**
  Takes a single Twitter-screen-name for input and returns the user-account with this screen-name. A user-ID can also be used for input, but this functionality is not required for crawling the dataset for this thesis
- **GET statusesretweets:id**
  Looks up all available Retweets for a provided Tweet (not used for crawling this dataset)
- **GET friendsids and GET followersids**
  Looks up the IDs of all friends/followers of a given user
- **GET friendslist and GET followerslist**
  Looks up the full user accounts of all friends/followers of a given user
- **GET statusesuser_timeline**
  Gets all available (latest 3200) Tweets of a given user account. Additionally this implementation allows to limit the returned Tweets by a given time period (so no Tweets older than a specific date are returned)

Threaded features All implemented API-requests are called by individual threads, so the requests and responses can be processed individually. Each thread uses only one Twitter authentication via the Twitter connection. Hence, if multiple authentications are used, at least as many threads per request have to be created. The usual threaded feature's work routine is the following: First of all, the thread gets the next item which has to be crawled from the data management part and queries the Twitter-API through its assigned Twitter connection. The response is then saved by the data management part. If the thread runs into a rate-limit-timeout, the thread sleeps until a query is allowed again. If the thread has no more work to do, the thread is terminated.

Data management The crawler's data management part manages all crawled data and uncrawled data. So all threaded features and the crawler controller hold an data management instance to store and access data. The data management is also responsible to save the crawled data to the database. This action is triggered every minute by the crawler's controller. If there is already some crawled data in the database, this data is loaded before any other request. So the crawler can be stopped and continued at every time with a minimum of data-loss.

Crawler controller The crawler's controller is the central controlling instance. The controller defines which data has to be crawled and which threaded features are used. If a threaded feature terminates too early, the feature is restarted by the controller. This situation usually arises if one threaded feature has to wait for the result of another threaded feature. For instance without crawled user-accounts, the friend and follower connection of theses user-accounts cannot be crawled beforehand. The crawler controller also has a built-in ability to stop the crawling process in a save way so that all running processes are stopped gracefully without any data-loss. The controller also serves as a wrapper for all crawler functions, because the other parts are all created and called by the controller.

Crawling strategy

The crawling strategy used to obtain this thesis' dataset can be described by several successive steps. Please note that most steps run asynchronously, so these steps do not have to be completed before the start of the next step. Furthermore, the dataset created by the crawler is only a snapshot. Therefore, all Twitter-data created or edited after the crawler's execution cannot be included.

- The list of scientists' Twitter-account-names are added to to the crawler
- The Twitter-accounts are looked up by using the Twitter-API. Both the GET userslookup request and the GET usersshow request are used to get all Twitter-accounts as fast as possible
- For each retrieved user-account the Tweets created by this account are looked up with the Twitter-API (GET statusesuser_timeline). The search for those Tweets is limited to Tweets created in or after 2013
- For each user-account the friend and follower relationships are obtained with the GET friendsids, GET followersids, GET friendslist and GET followerslist Twitter-API requests. This step usually takes most of the time, because even if both the app and user authentication are used simultaneously, these requests can only get a maximum of 159 000 friends or followers per 15-minute window.

Due to the long execution time of the crawler, the data-snapshot obtained from the Twitter-API might not be an exact snapshot. The researcher did his best to get the best snapshot considering all limitations.

4.2.3 Clean the dataset and prepare it for use

As a matter of principle, the dataset could be used directly after the crawling has been finished. However, there are some problems in the dataset which need to be addressed before actually using the dataset in order to avoid problems later on. These problems, their implications and reasons are explained in the following:

Friend and Follower connections outside the dataset During the crawling process, the ID's of each user's friends and followers were crawled. Naturally, many of these user-IDs point to users who are not in the dataset. For this thesis only the social relationships between users within the dataset are relevant. Hence all these unnecessary follower and friend connections can be deleted without further implications. After this cleaning process less disk space should be occupied and the computing time of several tasks using the dataset should be reduced (e.g. iterating over friends or followers).

Linking missing Retweet chains Usually a Retweet is linked to the retweeted Tweet and a Tweet is linked to all its Retweets. However, due to the crawler's inexactness some of these links might be missing. To solve this problem, the missing links are created. So if a Tweet links to a Retweet, but the Retweet does not link to the Tweet, this link is created. The same holds true for the other way around. Naturally, only the Retweet connections within the dataset are restored.

Equalize Follower and Friend lists After the friend and follower connection outside the dataset were removed, there should be as many friend as follower relationships by the reason that they represent just two views of the same data. However, after the crawling process this is not the case. In order to correct the data, for each missing follower relationship per friend relationship this relationship is created and vice versa.

4.3 Dataset characteristics

This section describes the dataset used for this thesis in more detail. First of all, the follower and friend relationships within the dataset and the network created by the follower relationships are examined. Then follows an examination of the time when Tweets were created. Thereafter follows a short description of the distribution of the users' activity. At the end of this section, the user interactions are discussed.

4.3.1 Followers and Friends

The social relationships play an important role when studying the interactions between users on Twitter. The network built by these relationships is the foundation for analyzing the information distribution through the observed Twitter network. Hence, this section examines the social relationships between all users within the dataset in more detail. Properties of follower network used for the creation of information flow trees (see section 3.2.2) are also described in this section.

Follower and Friend Statistics

The used Twitter dataset (see section 4.1.2) consists of 5 569 users. Together, these users have more than 13.5 million follower relationships and about 3 million friend relationships. Many of these relationships are connections to users not contained in the dataset. For this thesis, these relationships are considered irrelevant and are discarded to save memory and computing time when working with these relationships. Hence, the statistics presented in this section use only follower and friend relationships between users represented within the dataset.

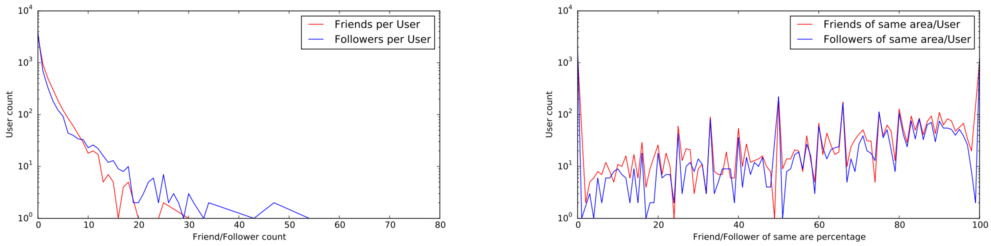
The exclusion of connection to users outside the dataset reduces the amount of follower and friend relationships to 88 070. In other words, the dataset just includes 0,7 % of all follower and 2,9 % of all friend relationships. The number of follower and friend relationships is the same after the filtering, because a follower relationship is just the reverse of a friend relationship. However, at first this was not the case due to some crawling incorrectness, but this error was corrected (see section 4.2.3).

An overview of the observed user's social relationships is shown in table 4.2. Like stated above, this data only applies to the users inside the used dataset. For instance, the observed users will probably have more than 15.8 followers on average. They have on average 15.8 followers within dataset, but they probably have more followers who are not included in the used dataset.

Since only social relationships within the dataset are considered, each user has averagely 15.8 friends and followers. Figure 4.1a illustrates the distribution of friends and followers per user. The plot clearly shows that followers and friends are not evenly distributed, so there are users who have clearly more friends than followers and vice versa. This fact is fortified

Users	5 569
Followers per user	15.8
Followers per user from the same research area	51 %
Friends per user	15.8
Friends per user from the same research area	57 %
Friends per follower	2.5
Users following their followers	49 %
Users following their followers of the same research area	35 %
Users followed by their friends	35 %
Users followed by their friends of the same research area	26 %
Users without friends	910 (16 %)
Users without friends and Tweets	423 (8 %)
Users without followers	1 493 (27 %)
Users without followers and Tweets	576 (10 %)
Users without friends and followers	780 (14 %)
Users without friends and followers and Tweets	370 (7 %)
Users without Tweets	995 (18 %)

Table 4.2: **Follower and friend statistics** - This table provides an overview of the follower and friend relationships' statistics. All values, which are not representing user count values, are mean values over all users.



(a) Friends and followers per user

(b) Percentage of friends and followers from the same research area

Figure 4.1: **Distribution of friends and followers per user and the shared research areas** - In figure (a) the distribution of friends and followers per user is shown. The x-axis shows the count of followers or friends, while the y-axis shows how much users have this amount of followers or friends. Figure (b) shows how many users' friends or followers are affiliated to the same research area.

by the mean proportion of friends per follower of 2.5. This value shows that most users follow more users than users are following themselves, while some few users have more followers than friends. Due to the fact that the fields of research are known to each user, one might think that users tend to create more relationships with users within their own research area. In order to verify this hypothesis, two users are considered belonging to the same research area, if they are affiliated to the same research area regardless their affiliation percentage. So even a user belonging to 100 % and another user belonging to 1 % to the same area are considered belonging to the same research area.

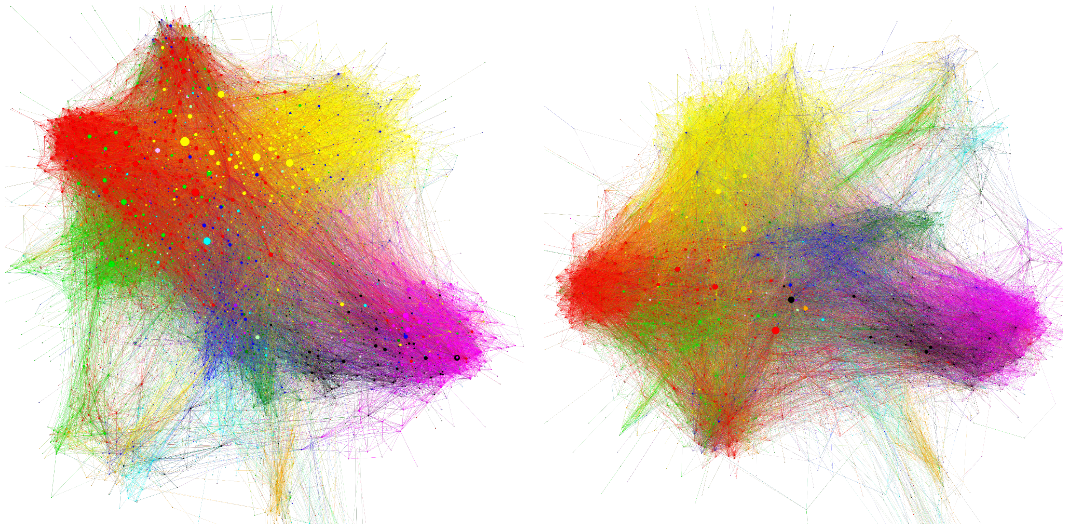
However, only 51 % of the average user's followers and 57 % of the average user's friends share the same fields of research. In figure 4.1b the distribution of the share of users affiliated to the same research area as the befriended or followed user. This distribution is not skewed as the distribution of friends and followers per user, but almost evenly distributed. There seems to be a slight tendency to prefer users of the same profession over other users. However, one must keep in mind that the affiliation percentage was not considered to calculate these results. Another method to verify the hypothesis that users from the own research area are preferred when creating follower relationships is to look at a the networks formed by the follower and friend relationships (see section 4.3.1).

Another property of the used dataset shown in table 4.2 is how many of the average user's followers are also friends (called *users following their followers*). Almost half of all of the average user's followers are followed by the average user. This figure gets even more significant for followers from the same research area, although the percentage is only 35 %. But this percentage is calculated for all followers not only those sharing the same research area. If only the followers from the same research area are considered, the average user follows almost 70 % of his own followers. However, the percentage of friends who follow oneself is only 35 % and 26 % for users of the same research area (or 46 % if only these users are considered). This characteristic is probably caused by the skewed friend to follower proportion.

An examination of the used dataset also shows that some users have no followers, friends and/or Tweets (see table 4.2). This examination shows that 1493 users (27 %) have no followers and 910 users (16 %) have no friends within the dataset. Of these users there are 780 users (14 %) who have neither friends nor followers within the dataset, which indicates that if a user has no friends, the user has very likely no followers as well, but the contrary relation is not so clear. Of the users without friends and followers, approximately the half of them can be considered as totally inactive, because they did not create a single Tweet in the observed time-period. However, probably all users without followers and friends and those without Tweets are irrelevant for this thesis, because the social relationships between the users and their Tweets are an important properties for almost all analyzes. Interestingly almost 18 % of all observed users never wrote a Tweet in the observed time period. Of these 995 users, 629 have either no friends or followers or both.

Networks formed by follower and friend relationships

By using all the social relationships of the Twitter dataset, directed networks showing the follower and friend relationships can be created. These two networks are similar, but all edges are reversed, because followers and friends are just two views of the same data. In this thesis, the follower network is more important, since the information flow trees are based thereof.



(a) Follower relationships (b) Friend relationships

Figure 4.2: **Visualization of follower and friend relationships** - The visualization of the follower and friend relationships was done with Gephi (Gephi.org, 2015). The coloring of the nodes is based on the user's research area with the highest percentage (in both plots the same color is used for each research area). The node's size is proportional to the node's in-degree. So in subfigure (a) the users who follow many other users have bigger nodes and subfigure (b) shows the users followed by many others accentuated.

Figure 4.2 shows the follower and friend networks. There the user's nodes are colored according to their research area with the highest affiliation percentage. So if a user is affiliated 80 % to AI and 20 % to HCI, the user's node is colored in the color belonging to AI. The size of the user's nodes is proportional to their in-degree. Hence, in the plot of follower relationships, users, who follow many other users, are highlighted and in the plot of friend relationships, users with many followers are highlighted. The position of each node was calculated by Gephi's ForceAtlas2 layout algorithm (Jacomy et al., 2011). This algorithm is not deterministic, so at each run the result might look differently. Therefore, and because the edge direction is considered as well, the research areas are not located at the same locations in 4.2a and 4.2b.

These plots show that users affiliated to the same research area are often surrounded by other users of the same research area. However, if the figures are examined more closely, the research areas are not separated in a clear way. For instance, the red area (data management) appears in two accumulation points and only the 4-5 research areas with most users are clearly visible. Hence, the plots show that there are no sharp borders between the users affiliated to different research areas.

Small world properties of the follower network

As proposed by Milgram (1967), many real-world networks have so-called small world properties. All these networks have in common that their diameter is small, so all vertices are connected with each other through only a few other vertices. Another property of small world networks is that if two vertices are both connected to another vertex, the two vertices are very likely connected to each other, too.

In order to check if a network has small world properties, the network is compared with a random network of the same size. If the network has small world properties it has to fulfill two requirements: First, the network's clustering coefficient has to be considerably larger than the random network's clustering coefficient. Second, the average path length has to bigger than or at least equal to the random network's average path length (Watts and Strogatz, 1998).

Concluding, the average path length (L) and the clustering coefficient (C) of

	Average path length (L)	Clustering coefficient (C)
Follower network	3.8363	0.2086
Random network (mean)	3.6370	0.0104

Table 4.3: **Small world properties of the follower network** - The values for the random network are the mean values over ten random networks with the same amount of vertices as the tested network. The shown values indicate that the follower network has indeed small world properties

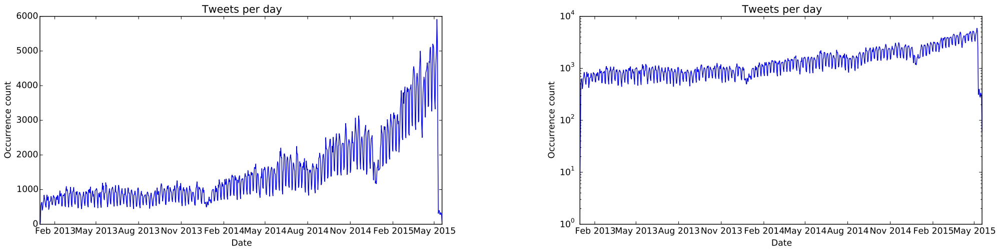
small world networks have to fulfill the following properties:

$$L > L_{random}$$
$$C \gg C_{random}$$

Table 4.3 shows the comparison of the follower network with random networks to check the network's small world properties. The random networks have as many vertices as the follower network. The results show that the follower network has indeed small world properties. The follower network's average path length of 3.8363 is slightly greater than the average path length of the random networks. This is also reflected in the follower network's diameter which is 8 instead of 6 or 7 like most often in the random networks. The follower network's clustering coefficient is significantly greater than the average clustering coefficients of the random networks, too.

4.3.2 Creation date of Tweets

The dataset used for this thesis consists totally of 1 538 661 Tweets. This section will explain in detail when these Tweets were created and if there are certain patterns in the distribution of Tweets over time. Therefore, the Tweets per day, week, and hour of the day are analyzed and discussed.

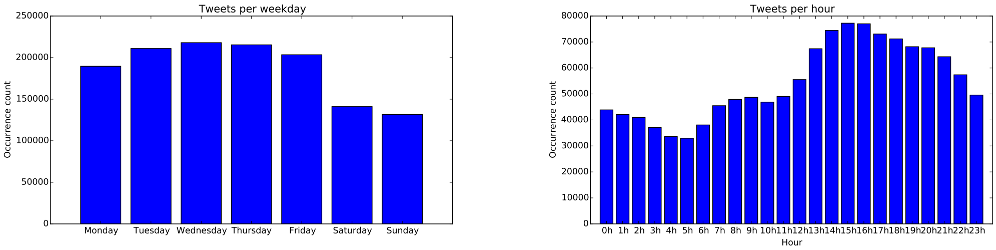


(a) linear scale

(b) logarithmic scale

Figure 4.3: **Tweets per day** - This figure shows the number of Tweets for each day in the observed period. The subfigures (a) and (b) show the same data, only the y-axis-scale is different. The number of Tweets per day grows steadily. In 2015, the growth is approximately exponential as figure (b) shows. The perpetual short increases and decreases are actually weeks (see figure 4.4a).

Figure 4.3 shows the number of Tweets for each day in the observed period. Overall, the number of Tweets per day grows steadily. There are no days where an extraordinary amount of Tweets can be observed. So there are no events, like scientific conferences or other events which might be of importance for a significant part of the observed users, which were able to push the number of Tweets while they took place. However, in 2013 and 2014 between Christmas Eve and New Year, the number of Tweets per day dropped. So the only event having a considerable impact on the number of tweets per day is not connected to computer science at all, but is simply Christmas holiday. In 2015, the number of Tweets seems to be growing exponentially, because in the logarithmic scale the ascent is linear. The growth is probably caused, on the one hand, by crawling errors. One crawling fault is that older Tweets are not always shown or might got deleted. Online discussions about the Twitter-API suspect that older Tweets are sometimes not shown due to infrastructure limitations, but there is no proof thereof. On the other hand Twitter's popularity might have increased among computer scientists. However, this growth might have other reasons as well, but these reasons will not be investigated by this thesis.



(a) Tweets per day of the week

(b) Tweets per hour of the day

Figure 4.4: **Tweets per weekday and hour of the day** - Plot (a) shows the tweets per weekday, which explains the ups and downs in figure 4.3). Most Tweets are created during working days, while the number of Tweets per day drops on weekends. Figure (b) shows the Tweets per hour of the day. It can be seen that during the night hours (23-7h) less Tweets were created than during the day.

The number of Tweets per day seems to be increasing and decreasing periodically. These ups and downs are in fact always occurring within exactly one week. So the Tweets per weekday were analyzed, too. The result can be seen in 4.4a and shows the reason for the periodic increase and decrease of Tweets per day, because on every weekend, the number of Tweets per day drops. Since more Tweets were created during the workdays than during the weekdays, computer scientists apparently are tweeting more during work than in their spare time. Due to the fact that researchers might prefer using Twitter on work days than actually at work, the Tweets per hour of the day were also analyzed. The result of this analysis can be seen in figure 4.4b. This figure shows that indeed more Tweets were created during the day than during the night. However, this analysis might not be accurate because the different time zones were not considered, because they are not included in the Tweet meta-data. So a computer scientist located in Japan might tweet during the day, but in this statistic her or his Tweets are counted as Tweets created in the night. Consequently, the question if computer scientists mainly use Twitter at work cannot be answered definitely, but still there is evidence to assume this hypothesis to be true.

4.3.3 Activity of users

The dataset used in this thesis consists of 1 538 661 Tweets and 5 569 users whereof only 4 574 users created at least one Tweet. Naturally, the Tweets are not equally distributed over the users. Hence there are users who created only a few Tweets and users who created a lot of Tweets.

Figure 4.5 shows eleven different user groups and the share of each group on the total number of Tweets and users. The users were grouped by their number of Tweets. The number of Tweets needed for each group can be seen on the x-axis. The blue bars indicate the group's share on the total number of users and the red bars show which percentage of all Tweets were created by each group. This plot shows that more than 70 % of all active users (users who created at least one Tweet) created less or equal than 300 Tweets, but they are responsible for only about 21 % of all Tweets. The 121 users with more than 1 500 Tweets, created almost as much Tweets as this big group. Altogether 50 % of all Tweets were created by just 513 users. The plot also shows that most users rarely write Tweets and that most of the observed activity on Twitter was done by a relatively small group of users. In summary, the distribution of Tweets per user is highly skewed.

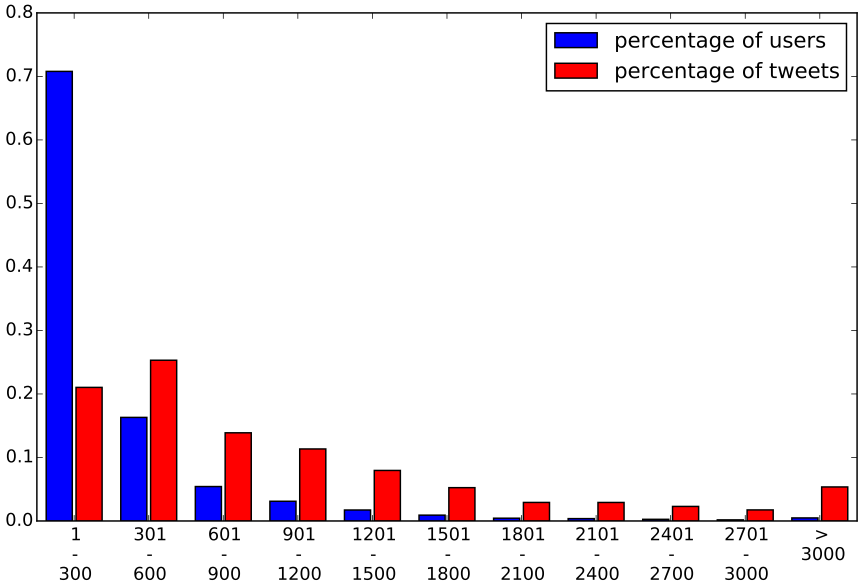
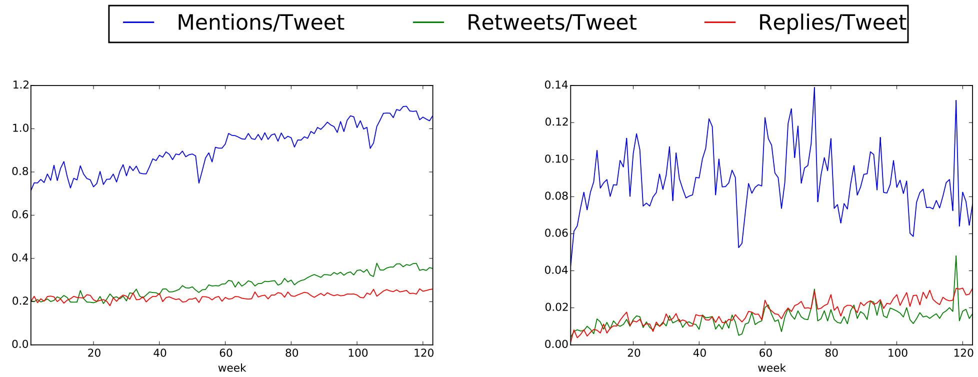


Figure 4.5: **Tweets per user** - In this figure the users are grouped by their number of Tweets. The required number of Tweets for each group is given on the x-axis. Each group features two bars: The blue bar indicates what percentage of the dataset's users are in a group and the red bar shows the percentage of Tweets created by the group's users

4.3.4 User interactions

The user interactions on Twitter discussed in this section include Retweets, user-mentions and replies. Following each other on Twitter can also be considered as user interaction, but social relationships in this dataset are already discussed in section 4.3.1. The first part of this section attends to the number of user interactions and if there is an observable growth or decay in the usage of Retweets, user-mentions or replies. The second part investigates the relationships of users retweeting, mentioning or replying each other. Please keep in mind that the discussed user interactions differ from each other. While a Tweet can be either a Retweet or a reply or none of these, there can be multiple user-mentions in a single Tweet.



(a) Total user interactions

(b) User interactions of users within the dataset

Figure 4.6: **Mentions, Retweets and Replies per Tweet over time** - The three lines show the number of mentions, Retweets and replies in each observed week. The plot shown in (a) is the plot of all observed user interactions in the Twitter dataset. Hence, user interactions between users who are not observed will be included too (for instance mentioning a user not within the dataset). The data shown in (b) consists only of user interactions which took place between users captured by the dataset.

Figure 4.6 shows the number of user interaction per Tweet for each week. The number of interactions per Tweet was chosen, because otherwise the increasing number of Tweets per day would influence the result. I decided to use weeks instead of days for these plots to smooth the results. Otherwise it might be harder to observe trends if there were too much outliers. The first subplot (4.6a) shows all user interaction which could be observed in the dataset of Tweets. For instance, mentions of users not within the dataset are included as well. The second subplot (4.6b) shows only the user interactions between users included in the dataset. The first plot shows that overall the share of replies stays more or the less the same over the observed time period. The number of user-mentions per Tweet and the percentage of retweets grows approximately in the same way, but while the percentage of retweets gets nearly doubled, the mentions per Tweet grow only by about a half.

The user interaction statistics between users inside the dataset are quite different from the observed interactions without this limitation. The most obvious difference is that there are roughly only a tenth of all interactions and the interactions per week oscillate more. The growth patterns of figure 4.6b are also different from those seen in figure 4.6a: The number of user-mentions stays more or less the same and the growth of the Retweet share is less evident. However, the reply share, which hardly grew in the other plot, grows plainly here.

Figure 4.7 shows the distribution of user mentions over time based on the users' social relationships. Only mentions between users inside the dataset are used. The user mentions are subdivided into four classes, which are defined by the involved users' relationships. The relationships used for this classification are the follower and friend relationships and the affiliation of users to the same research areas. The plot shows that users unconnected by any social relationship or membership in the same research area are rarely mentioned. Only about 5 % of all mentions are done between users like this. Users tend to mention users belonging to the same research area. About 80 to 90 % of all mentions are done between users affiliated to the same research area. Please note that the actual affiliation percentages are not considered here.

Users connected to any follower or friend relationship also mention each other frequently. However, most times these users also share an affiliation to the same research area. Only about 10 % of all mentions done between users with a friend or follower relationship but no common research area. Interestingly, the share of mentions done by users connected by following

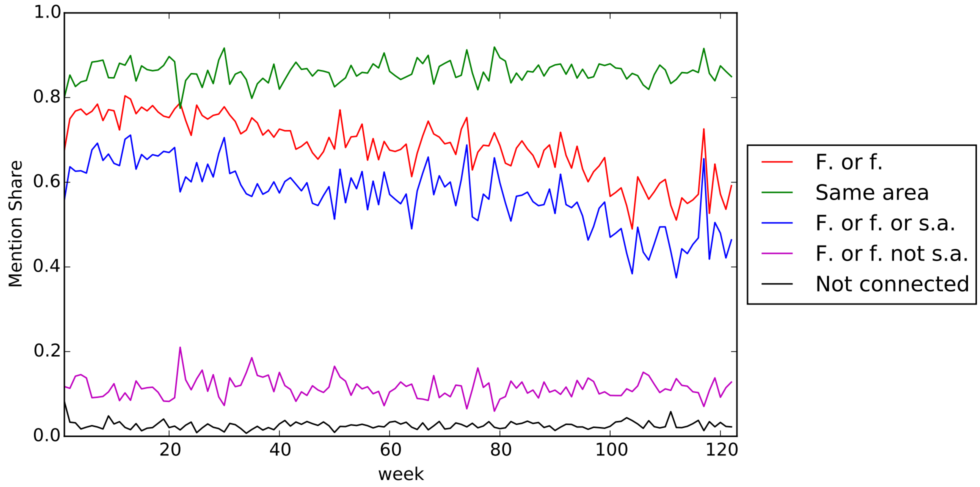


Figure 4.7: **Mentions subdivided by user relationships** - This plot compares the user mentions by the relationships between the involved users. The relationships are either follower or friend relationships or affiliations to the same research area. So there are five classes: users connected by a friend or/and follower connection (*F. or f.*), users affiliated to the same research area (*Same area*), users connected by both (*F. or f. or s.a.*), users connected by a friend or/and follower connection but not sharing a research area (*F. or f. not s.a.*) and users not connected to each other (*Not connected*)

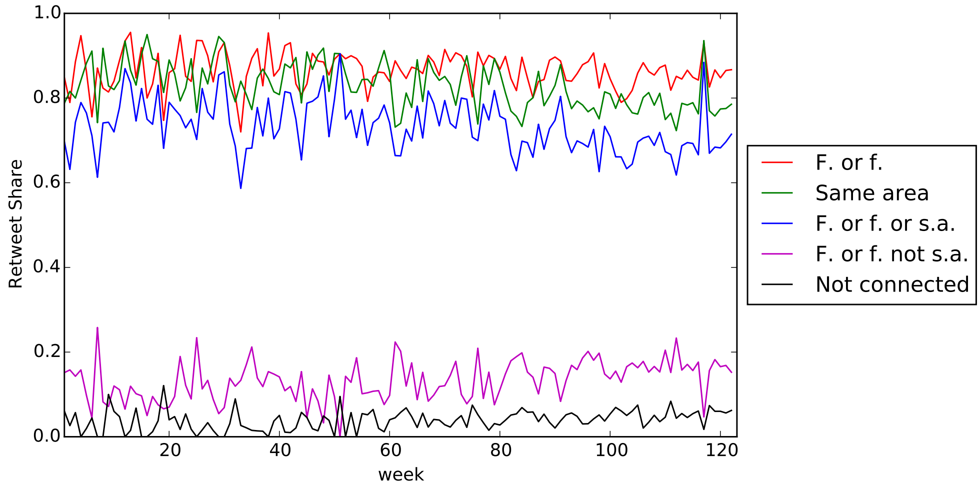


Figure 4.8: **Retweets subdivided by user relationships** - The data represented in this plot shows the share of Retweets by five different user relationship classes. The relationship classes are determined by the involved users' follower and friend relationships and their common research area affiliations. The classes are the same as in figure 4.7

each other decreases over the observed period by about ten percentage points.

The line plot displayed in figure 4.8 represents the share of Retweets partitioned into five classes. These five classes are identical to those explained in the previous paragraph. Retweeting users who are neither friends, followers nor members of the same research area are very rare. Most often users who share a common research area or are friends or followers retweet each other. However, over the whole observed time retweeting users affiliated with the same research area seems to get a bit less popular, while a few more unconnected users got retweeted. If users retweet a follower or friend, the users usually share a research area as well.

Figure 4.9 shows the share of replies by users grouped by their relationships. The relationships are likewise subdivided as in figure 4.7 and figure 4.8. The data represented by this plot clearly shows that replying each other is mostly done between users of the same research areas. Replies to users related to each other by follower and friend relationships is quite popular at

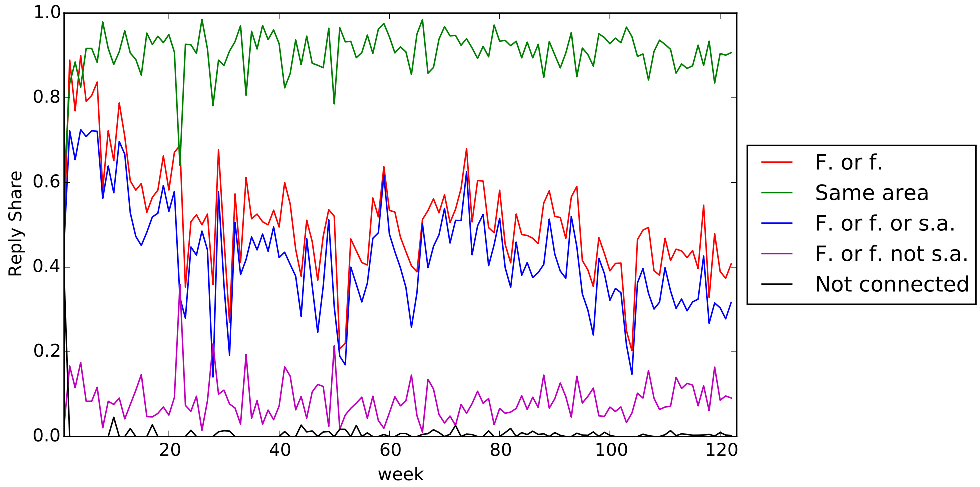


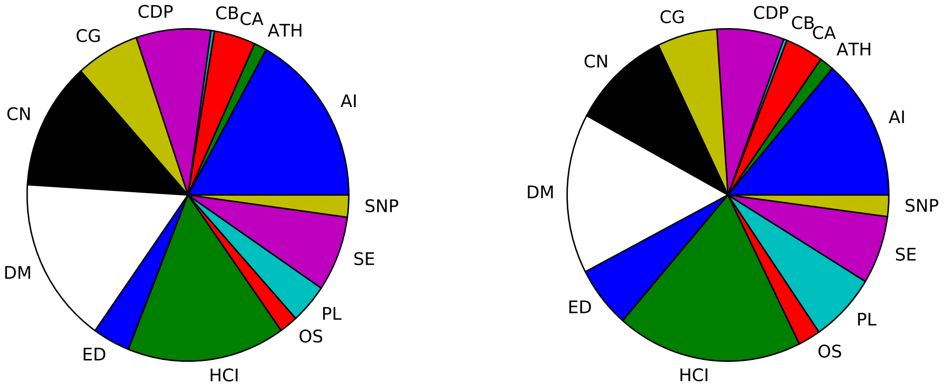
Figure 4.9: **Replies subdivided by user relationships** - This plot shows the share of users replying to Tweets of their friends and followers or computer scientists of the same field of research. Again, the user relationships are divided into five different classes.

the beginning of the observed period, but gets less popular over time. The reply share of these groups is reduced to about a half.

To sum up, users most often interact with users of their own research areas. Interactions between users who neither share a research area nor a social relationship on Twitter occur rarely. Many interactions also take place between users who are connected by follower relationships. However, most of these interactions can also be explained by involved users' common research interests. The interaction share of users solemnly connected by follower relationships usually rank between 10 and 20 %. On the other side, the interaction share of users with the same research area is usually above 80 %. These results provide further evidence that users often communicate with users sharing the same research area affiliation. Hence, area specific hashtags are quite likely.

4.4 Research area affiliations

Due to the fact that the research areas of each user are known (see section 4.1.2), the number of users belonging to a certain research area can be calculated. This is done by summing up the affiliation percentages of all users affiliated with a certain research area. Additionally, the number of Tweets created by a certain research area's users can be calculated. This calculation is straightforward if a user belongs to a single research area. Then the Tweet count for the user's area is just increased by one. However, users can belong to several research areas. So if for instance a user is affiliated to research area A1 to 90 % and with research area A2 to 10 % and creates a Tweet, the Tweet count of area A1 is increased by 0.9 and the Tweet usage count of area A2 increased by 0.1.



(a)Area affiliations of Users (b)Area affiliations of Tweets

Figure 4.10: **Research area affiliations** - The summarized research area affiliations of all users are shown in (a). Figure (b) shows the summarized research area affiliations of all Tweets. The abbreviations used in these plots are explained in table 4.1

Figure 4.10 shows the result of this calculation. In 4.10a the research areas' share of the users and in 4.10b the research areas' share of the dataset of Tweets is shown. The comparison of these two figures shows that the users of some research areas are more active on Twitter than the users of other research areas. For instance, the researchers affiliated to programming languages (PL) are more active than the researchers affiliate to artificial intelligence (AI). The plots also show that the users and Tweets are not equally distributed over all research areas. Some research areas clearly have a greater share of the whole dataset.

4.5 Hashtag Analysis

This section is focused on the hashtag usage within the used dataset. At the beginning, the overall hashtag usage is discussed. Thereafter follows an explanation which subsets of hashtags are used throughout this thesis. Then the hashtag usage over time is discussed: Firstly the overall usage over time is discussed and then the usage over time of selected hashtags is explained in detail. At the section's end, the assignment of hashtags to research areas is described.

4.5.1 Overall hashtag usage

Each Tweet can contain several hashtags, but there are also Tweets without hashtags. This section discusses the hashtag usage within the dataset of Tweets. Unsurprisingly, there are hashtags which are more popular than others. In total, 114 418 individual hashtags were found, but the majority of these hashtags vanished shortly after their first usage. By way of example, only 747 or 0.65 % of these hashtags were used at least 100 times.

Figure 4.11 illustrates the correlation of hashtag occurrence and hashtag quantity. The hashtags are grouped by their occurrence count and for each of these groups the number of members is shown on the y-axis. By the way of example, there are a little more than 100 hashtags which were used 321 to 640 times. The plot shows that the distribution of occurrences per hashtag is highly skewed. For instance there are only 19 hashtags which were used more than 1 000 times, but there 69.169 hashtags which were used only once.

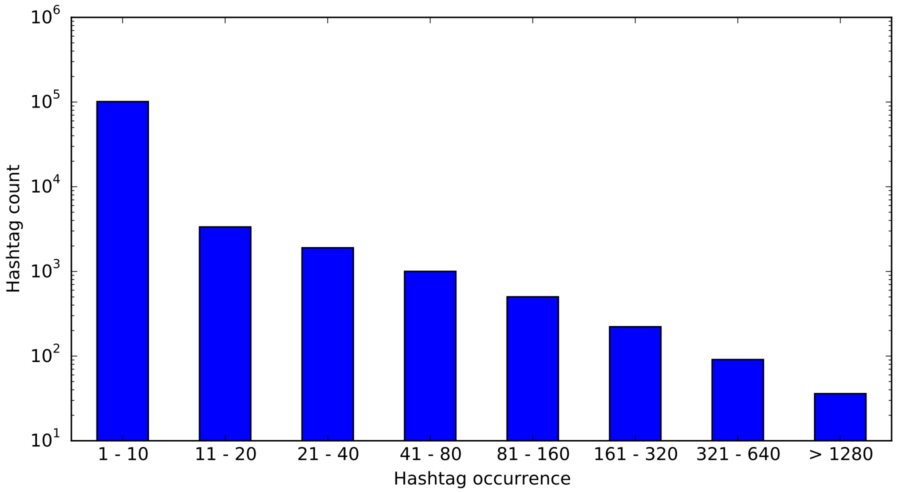


Figure 4.11: **Hashtag occurrence** - In this plot the hashtags were grouped by their occurrence-count. For instance, this plot shows that there are about 1 000 different hashtags, which were used 41 to 80 times.

4.5.2 Selected Hashtags for further analysis

The data presented in the previous section (4.5.1) shows that the majority of all hashtags are rarely used. These seldom used hashtags are problematic for several analyzes done in this thesis. For instance, information flow trees cannot be created for hashtags with only one usage and the creation for hashtags with few usages is often fruitless, too.

Hence, several subsets of all hashtags are used throughout this thesis. One often used subset consists of all hashtags with at least 100 usages by at least 10 different users. Using more than one individual user is especially important for the information flow tree creation, because a hashtag used 500 times by a single user was apparently not distributed through the observed Twitter network. The restriction to a minimum of 100 usages ensures that the hashtag is used several times. This set consists of 530 individual hashtags.

Sometimes another subset is used as well. This subset consists of all hashtags with at least 250 usages by at least 10 different users. There are 172 hashtags fulfilling these limitations and they are used if a smaller set of hashtags is needed.

Most results presented in this thesis could be calculated for both hashtag subsets explained above, but the results would be too long to include in this thesis. An example is the analysis presented in figure 4.14. If this plot would be calculated for 172 or even 530 hashtags, the result would fill several pages. Hence, 22 hashtags were randomly chosen out of all hashtags with at least 250 usages by at least 10 different users. These 22 hashtags are used for all computations where more hashtags would produce too long results. In the following they are often called *selected hashtags*. These 22 hashtags are:

- 3dprinting
- agile
- apachecon
- bigdata
- bitcoin
- charliehebdo
- chiplay
- cometlanding
- cybersecurity
- dvcon
- eclipse
- ibm
- iswc2014
- leadership
- netneutrality
- oscars
- rdaplenary
- sdn
- tech
- ted
- worldcup
- www2014

4.5.3 Hashtag usage over time

Given the changing nature of online social networks, the usage of hashtags might change over the observed time period. Hence, using hashtags in Tweets might become more popular or more new hashtags get generated at some point.

In order to see if there are changes or patterns in the usage of hashtag over the observed time period, figure 4.12 was created. Since there are so many seldom used hashtags, only hashtags with at least 100 usages from at least 10 different users were considered for this plot. The line indicating the usages of hashtags per Tweet (each Tweet can contain multiple hashtags) shows that the average overall popularity of hashtag usage does not change much. However, the weekly number of hashtags per Tweet oscillates heavily, so the hashtag usage is not constant. The blue line, which indicates the total number of hashtags usages per week, increases clearly over time. Though

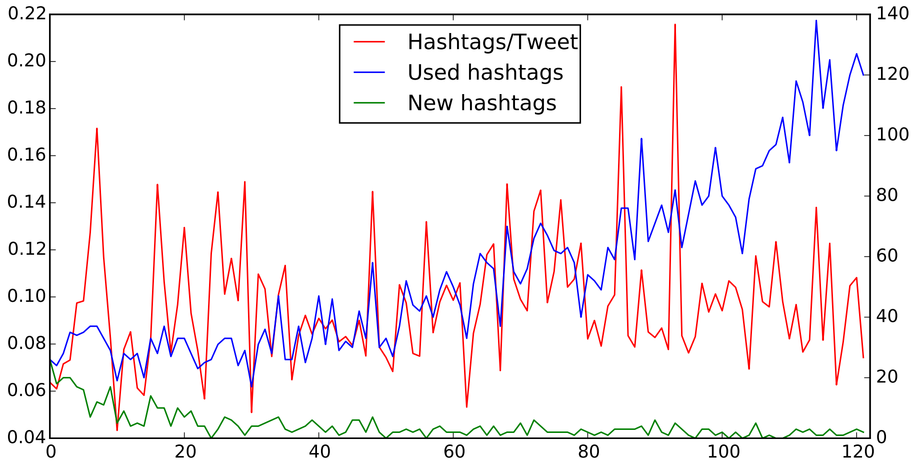


Figure 4.12: **Hashtag usage** - The plot shows the number of hashtags per Tweet, the number of used individual hashtags and the number of newly introduced hashtags for each observed week. The left-handed y-scale belongs to the hashtags per Tweet data, the right-handed y-scale belongs to the two other data representations

this increase is not remarkable because the daily number of Tweets is also increasing by approximately the same amount (see figure 4.3). The green line in figure 4.12 shows how many hashtags were used for the first time (regarding the used dataset) each week. This data representation is almost constant after a decrease in the first 30 weeks. This decrease is probably caused by the more popular hashtags, which are often used regularly over the whole observed period.

4.5.4 Detailed hashtag usage over time of selected hashtags

Hashtags are often linked to events and popular topics, thus some hashtags might be used only within a limited time period or reoccur at certain intervals. Hence, the usage of specific hashtags over the observed period is discussed in this section.

Figure 4.13 shows the usage of the selected hashtags over the observed period. This data representation shows that only the hashtags *cometlanding*

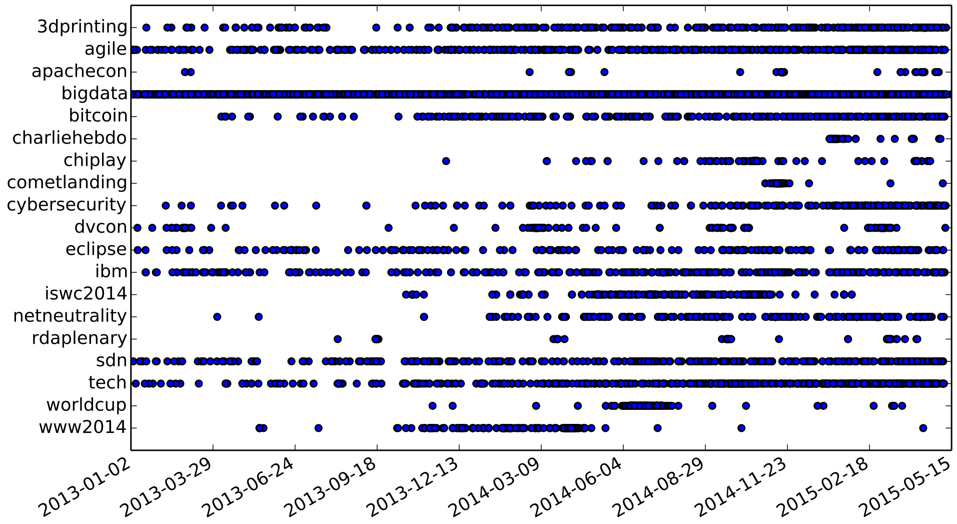


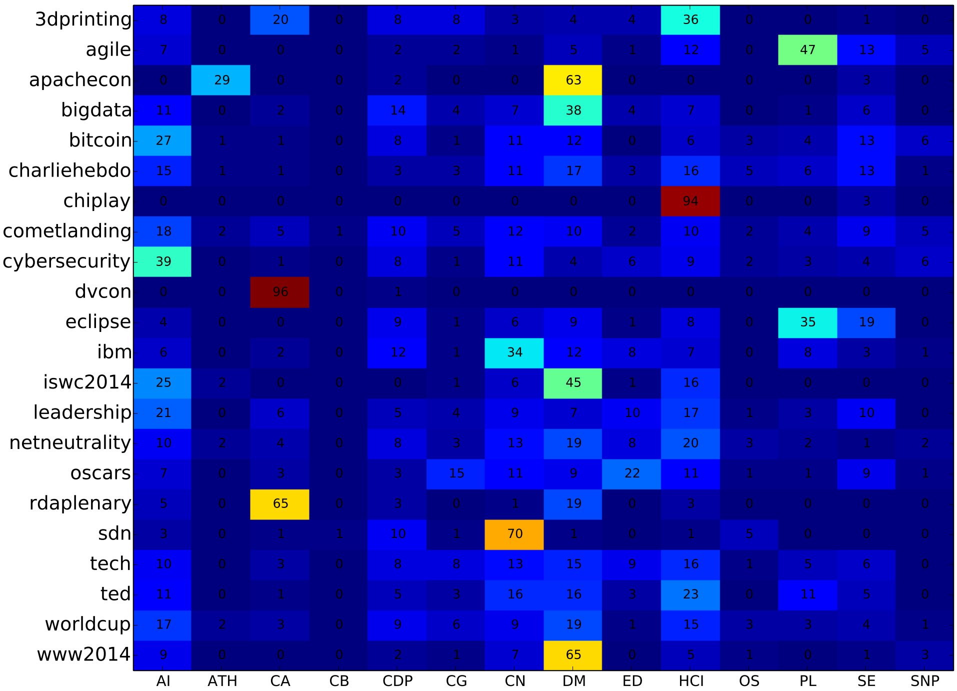
Figure 4.13: **Hashtag usage over time** - This plot shows the usage over time for each hashtag. Each dot represents a usage of a specific hashtag at a specific time. Each dot's hashtag is named on the y-axis and the time can be seen on the x-axis. This way, patterns in the hashtag usage are observable

and *charliehebdo* are active just once in a short time period. Both hashtags are not directly related to computer science which could explain the short usage time. Most hashtags displayed in this figure have no real usage peaks, they are almost constantly used throughout the observed time period. The only reoccurring hashtags are *dvcon*, *oscars* and *rdaplenary*, which are all hashtags linked to reoccurring events. Interestingly, conference hashtags like *iswc2014*, *chiplay* and *www2014* are much longer present in the Twitter dataset than the actual conferences took place. Contrariwise, the hashtag for the football world cup (hashtag *worldcup*), which is not linked to computer science, was mainly used during the world cup.

4.5.5 Research area specific and general hashtags

Due to the fact that the research areas of each user are known (see section 4.1.2), the affiliation of hashtags to certain research areas can be calculated. By the use of this result, research area specific and general hashtags can be identified (**RQ 1**). The subdivision of hashtags to the fourteen research areas is done in a similar way as the affiliation of Tweets to research areas (see section 4.4). The only difference between these two calculations is that this calculation is done per hashtags instead of Tweets. So if for instance the hashtags H1 and H2 were used in a Tweet and the Tweet's creator is affiliated to research area A1 to 90 % and to research area A2 to 10 %, the usage count of area A1 for H1 and H2 are increased by 0.9 and the usage count of area A2 for H1 and H2 are increased by 0.1.

Figure 4.14 shows the result of this assignment of hashtags to research areas for selected hashtags. There, the affiliation of a hashtag to a research area is given in percent (the percentages are rounded, so the sum is not always 100). The affiliation percentages are color coded to distinguish the different values. Hence, the visual differentiation of research area specific and general hashtags is possible. Research area specific hashtags are hashtags, which are mainly used by users affiliated to only a few different research areas. General hashtags are hashtags used by users of several different research areas and ideally the hashtag usage is distributed evenly over all research areas.



	AI	ATH	CA	CB	CDP	CG	CN	DM	ED	HCI	OS	PL	SE	SNP
3dprinting	8	0	20	0	8	8	3	4	4	36	0	0	1	0
agile	7	0	0	0	2	2	1	5	1	12	0	47	13	5
apachecon	0	29	0	0	2	0	0	63	0	0	0	0	3	0
bigdata	11	0	2	0	14	4	7	38	4	7	0	1	6	0
bitcoin	27	1	1	0	8	1	11	12	0	6	3	4	13	6
charliehebdo	15	1	1	0	3	3	11	17	3	16	5	6	13	1
chiplay	0	0	0	0	0	0	0	0	0	94	0	0	3	0
cometlanding	18	2	5	1	10	5	12	10	2	10	2	4	9	5
cybersecurity	39	0	1	0	8	1	11	4	6	9	2	3	4	6
dvcon	0	0	96	0	1	0	0	0	0	0	0	0	0	0
eclipse	4	0	0	0	9	1	6	9	1	8	0	35	19	0
ibm	6	0	2	0	12	1	34	12	8	7	0	8	3	1
iswc2014	25	2	0	0	0	1	6	45	1	16	0	0	0	0
leadership	21	0	6	0	5	4	9	7	10	17	1	3	10	0
netneutrality	10	2	4	0	8	3	13	19	8	20	3	2	1	2
oscars	7	0	3	0	3	15	11	9	22	11	1	1	9	1
rdaplenary	5	0	65	0	3	0	1	19	0	3	0	0	0	0
sdn	3	0	1	1	10	1	70	1	0	1	5	0	0	0
tech	10	0	3	0	8	8	13	15	9	16	1	5	6	0
ted	11	0	1	0	5	3	16	16	3	23	0	11	5	0
worldcup	17	2	3	0	9	6	9	19	1	15	3	3	4	1
www2014	9	0	0	0	2	1	7	65	0	5	1	0	1	3

Figure 4.14: **Research area affiliation of hashtags** - This plot shows which share of hashtags was used by users affiliated to a certain research area. The abbreviations used in the x-axis represent research areas (see table 4.1). The percentages shown in the plot are color coded for easier differentiation of outstanding values.

For instance, the hashtags *chiplay* and *dvcon* were almost solemnly used by users associated with human computer interaction (HCI) and computer architecture (CA) respectively. The figure also shows obvious examples for general hashtags like *cometlanding* and *tech*. However, there are also hashtags which are not easily visually classified into general or research area specific hashtags. In order to classify all hashtags in equal measure, a simple assignment of hashtags to research areas is used. This hashtag-area-assignment divides the hashtags into general and area specific hashtags. The area specific hashtags are further divided by the number of involved research areas. Theoretically there is no limit for research areas for area specific hashtags, but as it happens the maximum number of areas for an area specific hashtag is three for the used data.

The hashtag-area-assignment works as follows: For each hashtag, the mean value and standard deviation of the research area affiliation percentages are calculated. The sum of these two values forms a threshold. Each percentage, which is greater than this threshold, is selected. If the sum of all differences between these selected percentages and the threshold is also greater than the threshold, the hashtag is assumed to be area specific. The number of selected percentages indicates the number of involved research areas per area specific hashtag. All hashtags which do not fulfill these requirements are assumed to be general hashtags.

Hashtags	General	1 area	2 areas	3 areas
22 selected	bitcoin, charliehebdo, cometlanding, leadership, netneutrality, oscars, tech, ted, worldcup	agile, bigdata, chiplay, cybersecurity, dvcon, ibm, rdaplenary, sdn, www2014	3dprinting, apachecon, eclipse, iswc2014	-
100 usages	118 (22%)	195 (37%)	168 (32%)	49 (9%)
250 usages	31 (18%)	74 (43%)	53 (31%)	14 (8%)

Table 4.4: **Research area assignment of hashtags results** - The assignment was done for three subsets of all hashtags within the dataset (see section 4.5.2). For the 22 selected hashtags the concrete results for each hashtag are given, for the other subsets only the sum over the hashtags of each class is given.

The result of this assignment can be seen in table 4.4. This result shows that most hashtags are considered as research area specific hashtags, at least with the used assignment. Around 40 % of all hashtags are specific to one area and about 30 % are specific to two research areas. Only less than 10 % of all hashtags are considered as specific to three different research areas. Of the 22 selected hashtags, 9 hashtags are general, 9 are specific to one area and 4 are specific to two areas. Considering hashtags with less than 50 % affiliation to one research area might not seem right. However, for instance *agile* is only affiliated to 47 % to programming languages (PL), but this area is clearly the most important one for this hashtag. The result of the automatic assignment also fits for the visual classification of the 22 selected hashtags. Hence, the answer to **RQ 1** is that research area specific hashtags exist and it is possible to classify them automatically.

Figure 4.14 also shows that users of some research areas like *CB, OS* and *SNP* hardly used any of the analyzed hashtags. This observation matches with the number of users and Tweets affiliated with these areas. Figure 4.10 shows that these research areas have only few users and Tweets. Another observation is that the users of some areas tend to use more general than area specific hashtags. For instance, the users affiliated to computer architecture (CA) use mainly research area specific hashtags, while the users of artificial intelligence (AI) often use general hashtags. However, this observation might only be true for the analyzed hashtags.

4.6 Information Flow Tree Analysis

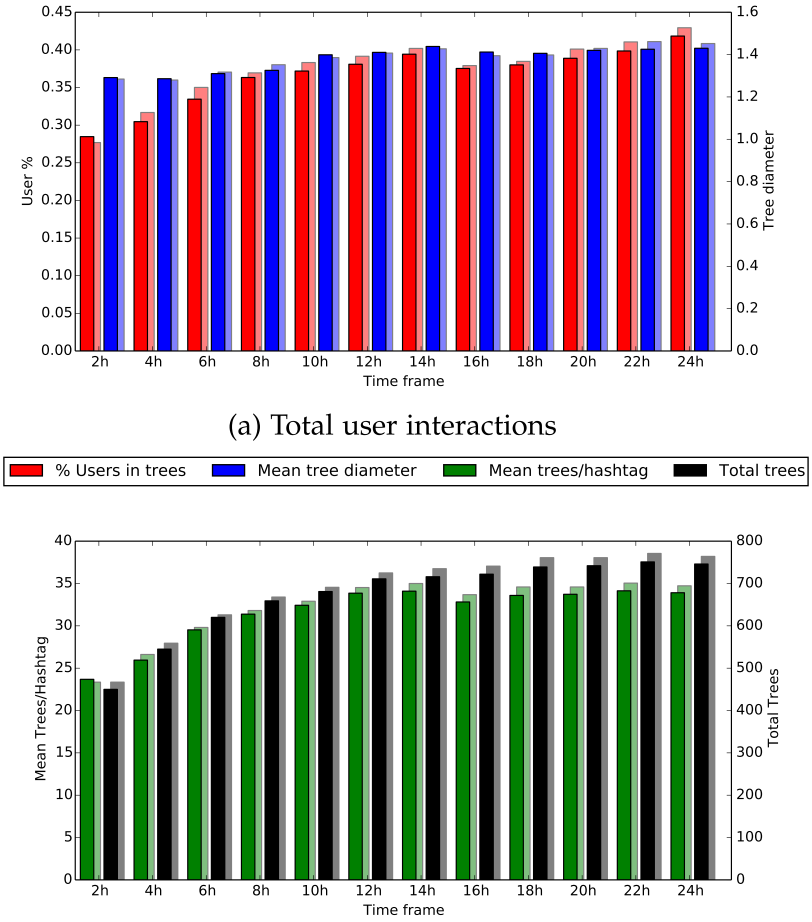
The following section is focused on the analysis of information flow trees. First of all, the optimal parameters for the information flow tree creation for the used dataset are discussed. Thereby the optimal time frame is explained as well. Thus, the second research question (**RQ 2**) should get answered appropriately. Thereafter follows an explanation of the information flow trees created for some selected hashtags. The section's last part is dedicated to the question if attributes calculated from the users' activity influence the measurements taken from the resulting information flow trees.

4.6.1 Information Flow Tree creation variants

In section 3.2.2 the creation of information flow trees was discussed. There, two slightly different methods to create information flow trees were explained. Additionally, both methods support excluding weekends from the maximum allowed time between two hashtag usages. This exclusion of weekends is not reasonable, since the optimal time frame for the used dataset calculated in this section is well below one day. Excluding weekends would prolong the time frame from less than one day to several days in some cases. However, whether allowing multiple timestamps per user yields better information flow trees has to be evaluated. It is assumed that the better information flow tree creation method yields more and longer information flow trees than the other.

Figure 4.15 presents the statistics of the information flow trees created for 22 selected hashtags. The first plot is showing the mean tree diameters of all produced information flow trees (blue bars) as well as the percentage of users represented in the trees (red bars). This percentage was calculated by counting all individual users inside an information flow tree created for a specific hashtag and dividing this value by the count of all individual users using this hashtag. For the visual representation the mean value over all used hashtag measurements was used. The second plot is showing the average number of information flow trees created for each hashtag (green bars) and the number of totally created information flow trees (black bars). Each data row consists of a solid and a half transparent bar. The solid bar shows statistics for the information flow tree creation method which allows only one timestamp per user and the half transparent bar shows the statistics for the other creation method.

Usually, the information flow tree creation method which allows multiple timestamps per user creates slightly more and bigger trees with more users than the other method. While there is only a slight difference, still this method seems to be the better choice for the used data. Summing up the displayed results, the number of generated information flow trees and their size grows almost steadily with the time frame. Starting around the time frame of eight hours the growth lessens. Between the time frames of fourteen and eighteen hours the mean number of information flow trees per hashtag



(a) Total user interactions

(b) User interactions of users within the dataset

Figure 4.15: **Statistics of information flow tree creation for selected hashtags with different parameters** - These plots show statistics for the information flow trees created for 22 selected hashtags with different parameters. The half transparent bars always show the information flow tree creation variant with multiple allowed timestamps, while the solid bars represent the creation variant with only one allowed timestamp per vertex. Each bar has its own y-axis in order to compare the values more easily. All y-axis are linearly scaled and they always belong to the bars closer to them.

and the number of included users drops slightly, but this could be just an artifact caused by the used hashtags.

In summary the second information flow tree creation method fits better and the time frame should be probably greater than eight hours. However, the choice of the optimal time frame for creating information flow trees should not depend on the results quality. Instead the time frame should be chosen in a way so that actual human behavior is best represented by the resulting information flow trees.

4.6.2 Optimal time frame

For creating information flow trees, the chosen time frame is crucial. A well-chosen time frame ensures that users are included in an information flow tree only if they were likely influenced by another user to use a certain hashtag. In order to get influenced, the hashtag should travel through the Twitter network. If the time frame is too big, the probability that the hashtag was distributed in a different way than the Twitter network increases. The best way to get a good time frame would be using the results of an existing study concerning this relation. However, there seems to be no study like this.

Hence, another way to get the optimal time frame for the creation of information flow trees is needed. Retweets are direct reactions to Tweets of other users and followed users get retweeted most often. So the time between the publication of a Tweet and its Retweets might be a good indicator for the optimal time frame.

Figure 4.16 shows how much time elapses until Retweets are created in the used dataset. Most Retweets are created relatively early after the initial Tweet's publication and within 24 hours all Retweets are written. Consequently, any time frame greater than 24 hours is probably wrong for the creation of information flow trees for this dataset. Only one and a half hour after the creation of the initial Tweets, 50 % of all Retweets are written. On average, a Retweet gets written 4.6 hours after the initial Tweet. The standard deviation for all time differences between Retweet and initial Tweets is 6.1 hours. In a time frame of 10.7 hours (sum of mean and standard

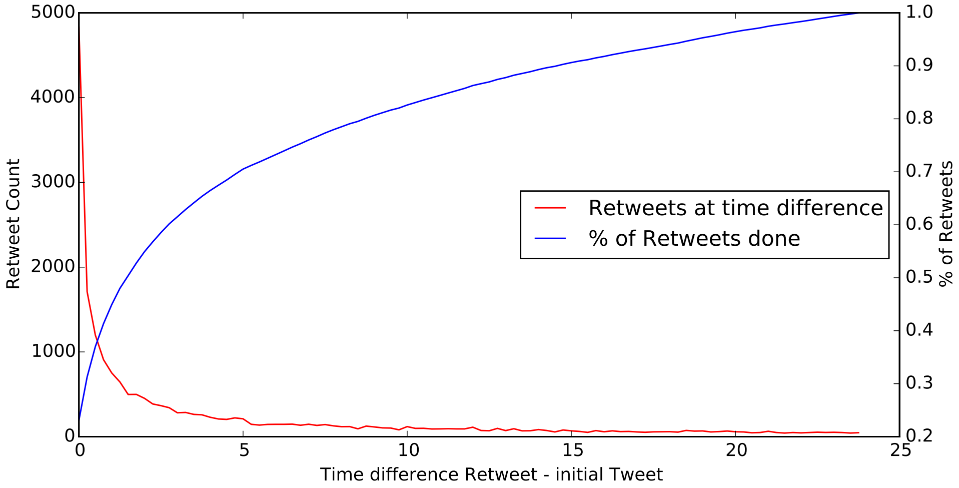


Figure 4.16: **Retweet distribution** - The distribution of the time difference between Retweet and retweeted Tweet is shown in this figure. The x-axis shows the number of hours between Retweet and initial Tweet. The red line indicates how many Retweets were written within a certain time period (the time was measured in quarter hours to avoid too much clutter). The left y-axis shows this Retweet count. The blue line is representing the percentage of written Retweets until a certain time period (the right y-axis is affiliated with the blue line).

deviation), 84 % of all Retweets are written. However, an information flow tree should represent the information distribution as good as possible and if 16% possibly not included distribution ways seems to be a little too much. 90 % of all Retweets are written within a time frame of 14.5 hours, 95 % within 18.75 hours, 97 % within 20.75 hours and 99 % within 22.75 hours.

Hence, the optimal time frame for the creation of information flow trees with this data is probably between 15 and 22 hours. Regarding the results from 4.6.1 a time frame of 19 hours is chosen for all following experiments. This time frame ensures that if Retweets were tracked, 95.3 % would be included and there seems to be no problem creating information flow trees with this time frame with the thesis' dataset.

4.6.3 Information flow trees created for selected hashtags

In order to show some statistics on the resulting the information flow trees, all information flow trees for the 22 selected hashtags were created. Instead of looking at the statistics for all hashtags at once, the information flow trees for each hashtag were analyzed separately. Hence, the total number of information flow trees, their length and their duration was analyzed for each hashtag.

Figure 4.17 shows the results for those three attributes. These results show that the information flow trees created for the hashtags differ greatly. However, with more information flow trees generated, the probability to get many trees with many users and covering a longer time period rises. All hashtags have in common, that their information flow trees seldom cover more than two days and the trees' diameter is mostly just two or three. The hashtag *agile* seems to benefit most from the allowance of multiple timestamps per user, because while all its information flow trees have a diameter of just one, there are trees covering more than nineteen hours. The hashtags creating the longest information flow trees are *charliehebdo* and *www2014*. The information flow trees of many hashtags have a diameter of just one. There seems to be no correlation between tree diameter and tree duration, so the response times for certain topics are longer than for other topics.

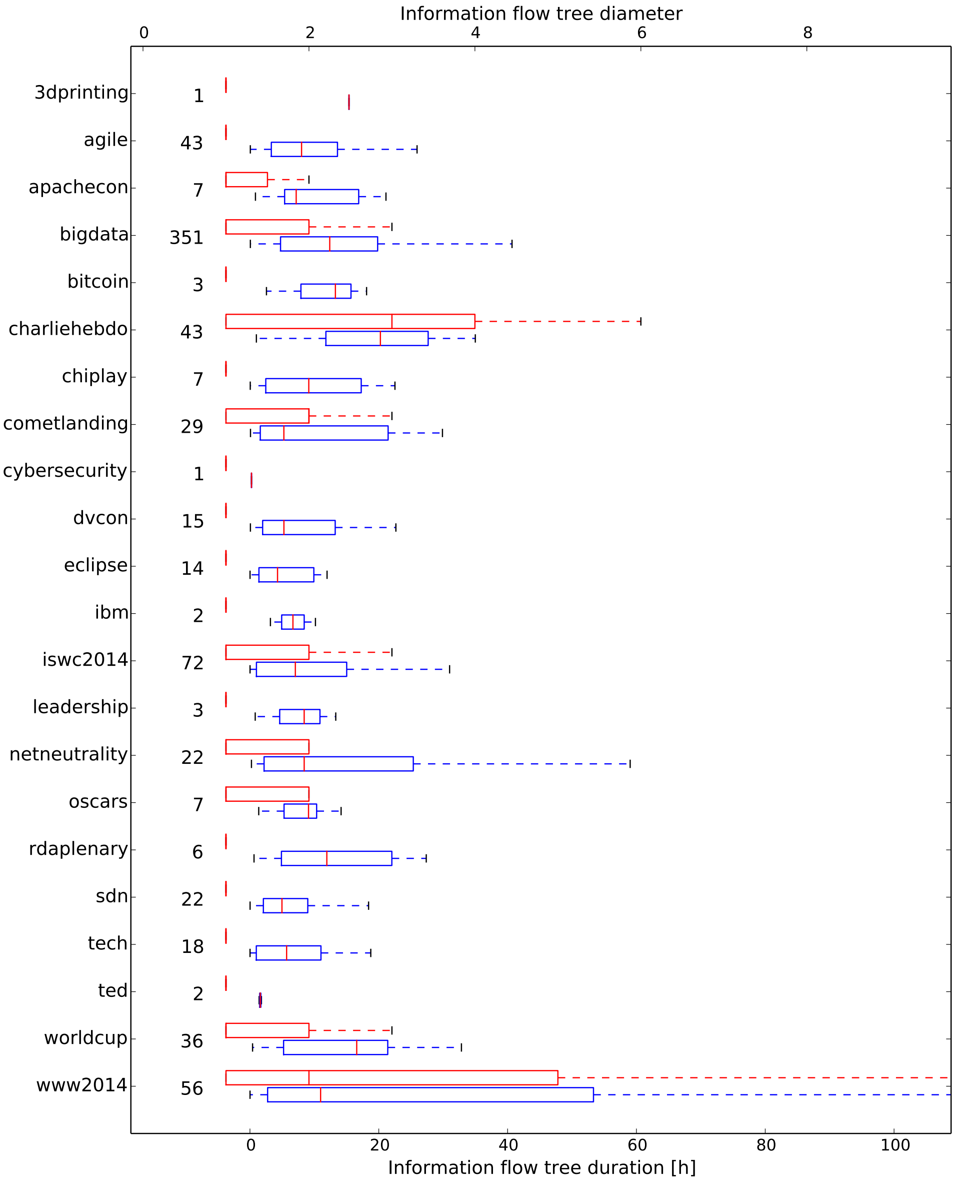


Figure 4.17: **Count, size and duration of information flow trees created for selected hashtags** - The numbers in the plot are indicating the total number of created information flow trees for the current hashtag. The red boxplots show the diameters of the information flow trees created for each hashtag. The blue boxplots show the duration of all created information flow trees for each hashtag. The duration is the maximum difference between the earliest and the latest timestamp in the information flow tree. The upper x-axis belongs to the red boxplots, while the lower y-axis is affiliated with the blue boxplots.

4.6.4 Information flow tree measurements correlating with other user attributes

One of this thesis' hypothesis is that users playing an important role at distributing a hashtag through the Twitter network have attributes correlating with the measurements calculated from the information flow trees (**RQ 3**). If this hypothesis is correct, good information spreaders and creators could get recognized without looking at the information flow trees. Hence there might be some user attributes which correlate with some attributes calculated from the information flow trees.

The social ties of a user were deemed important for the user's role in the information flow trees. So the user's prominence in the Twitter network is probably important, which is measured here by the user's count of followers and friends within the dataset. Besides the user's social relationships, the activities are considered to be important. Therefore, the number of Tweets and Retweets a user created within the observed time period are also used as user attributes. Chosen as relevant information flow tree measurements for users are the source and sink count, how often a user appears in individual trees and how long they get and how well a user spreads information. The user attributes and information flow tree measurements which are tested if they correlate with each other are listed and described in detail in the following:

The used user attributes are:

- **Followers** - The count of a user's followers within the dataset
- **Friends** - The count of a user's friends within the dataset
- **Retweets** - The number of Retweets a user wrote within the observed time period in reaction to another Tweet within the dataset
- **Tweets** - The total number of Tweets a user wrote within the observed time period

The used information flow tree measurements are:

- **Information flow trees** - Count of user appearances in individual information flow trees
- **Information spread** - The average number of a user's out-links in each information flow tree (the sinks are disregarded)

- **Sink** - The number of times a user is acting as information sink
- **Source** - The number of times a user is acting as information source
- **Tree length** - The average length of information flow trees where the user is present

Table 4.5 shows the results of the correlation between all user attributes and information flow tree measurements. These correlations were calculated for all users with at least one information flow tree. If all users were used, the correlation coefficients would be higher, but this is clearly the wrong way, because inactive users have no information flow trees.

While almost all user properties correlate with the information flow tree measurements, there is quite a wide range between strong and weak relationships. In summary, the Tweet count is least indicative for predicting any user role within the information flow trees. Only the source count and a little bit less the number of information flow trees are correlating with the number of Tweets. This behavior is expected, because if a user creates more Tweets, there is a higher probability that the user is in more information flow trees and thus also starts more information cascades.

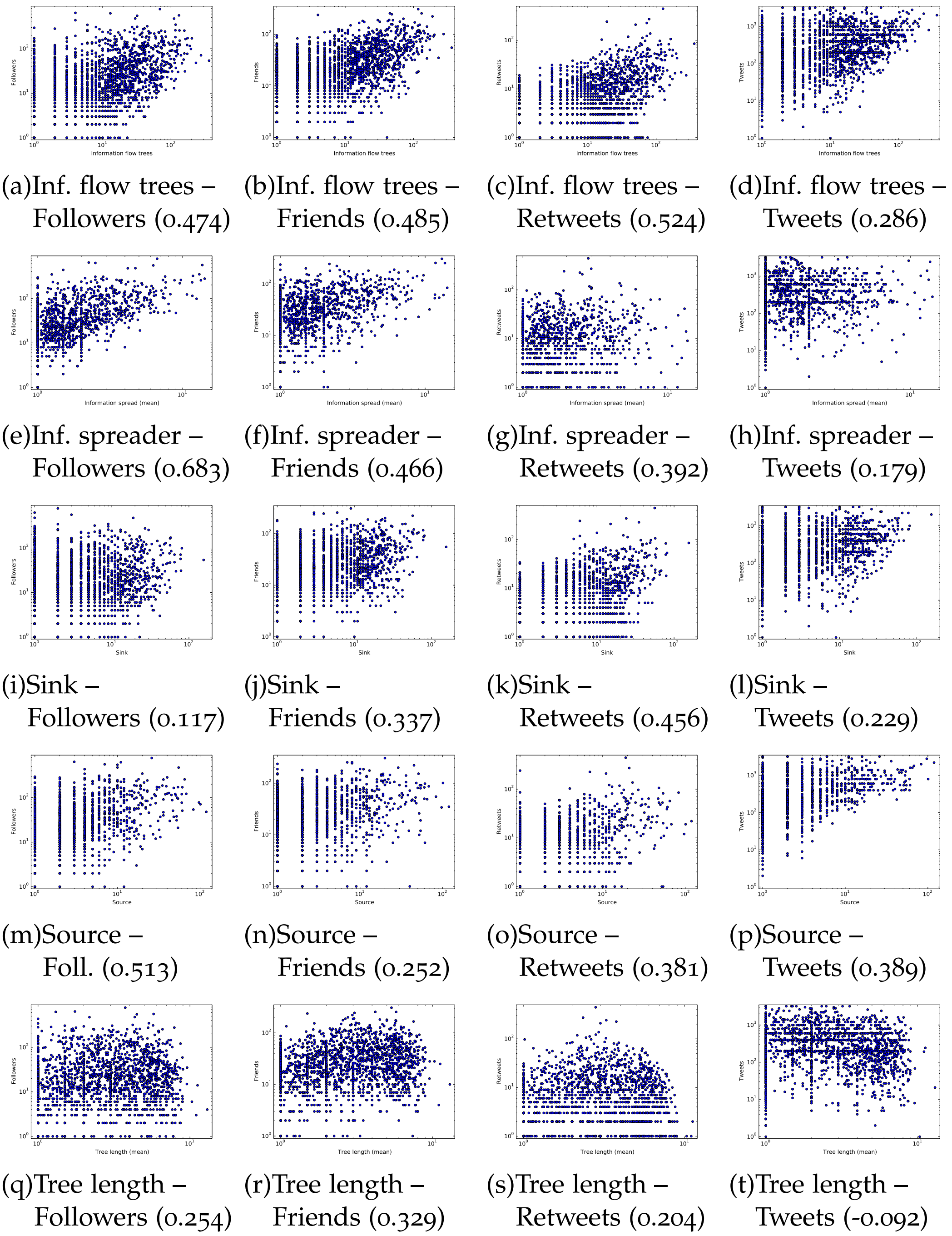
The follower and Retweet count are the best user attributes to predict information flow tree measurements. On average, they score a spearman correlation coefficient of about 0.4. The best correlation is between the users' follower count and information spread efficiency. This result fits quite nicely, because a user with many followers is more likely to spread a piece of information to more other users than a user with only a few followers.

The information flow tree count and the information spreading measurement are correlating best with all used user attributes. The source count seems more tightly connected with the user attributes than the sink count. Probably the source count is more difficult to analyze, because the reasons to stop an information cascade are manifold and might relay on the topic as well. The average information flow tree length is only weakly correlating with the user attributes.

In order to illustrate the correlations between the user attributes and information flow tree measurements, figure 4.18 was created. There, all twenty comparisons are displayed in scatterplots. Although the plots are very small, a comparison of the different correlations is possible, because all correlations

User property	Info. flow tree measurement	Spearman's rho
Followers	Information flow trees	0.474
	Information spread	0.683
	Sink	0.117
	Source	0.513
	Tree length	0.254
Friends	Information flow trees	0.485
	Information spread	0.466
	Sink	0.337
	Source	0.252
	Tree length	0.329
Retweets	Information flow trees	0.524
	Information spread	0.392
	Sink	0.456
	Source	0.381
	Tree length	0.204
Tweets	Information flow trees	0.286
	Information spread	0.179
	Sink	0.229
	Source	0.389
	Tree length	-0.092

Table 4.5: **Spearman correlation results of user attributes and information flow tree measurements** - The results of the spearman correlation between all user attributes and information flow tree measurements are given by the spearman rho coefficient. The spearman rho coefficient ranges from -1 to 1. All positive values indicate a relationship, but the higher the value gets, the better is the correlation. Values between 0 and 0.2 indicate a very weak, values between 0.2 and 0.4 indicate a weak, values between 0.4 and 0.6 indicate a moderate and values between 0.6 and 0.8 indicate a strong relationship.



(a)Inf. flow trees – Followers (0.474)

(b)Inf. flow trees – Friends (0.485)

(c)Inf. flow trees – Retweets (0.524)

(d)Inf. flow trees – Tweets (0.286)

(e)Inf. spreader – Followers (0.683)

(f)Inf. spreader – Friends (0.466)

(g)Inf. spreader – Retweets (0.392)

(h)Inf. spreader – Tweets (0.179)

(i)Sink – Followers (0.117)

(j)Sink – Friends (0.337)

(k)Sink – Retweets (0.456)

(l)Sink – Tweets (0.229)

(m)Source – Foll. (0.513)

(n)Source – Friends (0.252)

(o)Source – Retweets (0.381)

(p)Source – Tweets (0.389)

(q)Tree length – Followers (0.254)

(r)Tree length – Friends (0.329)

(s)Tree length – Retweets (0.204)

(t)Tree length – Tweets (-0.092)

Figure 4.18: **Scatter plots of correlations between user attributes and information flow tree measurements** - Each subplot's label is showing the compared values for each user and in the brackets the spearman's rho coefficient of these values is shown. On the y-axis of each plot are the user attributes, the x-axis is showing the information flow tree measurements. All axes are scaled logarithmically.

are displayed together. Each row is dedicated to an information flow tree measurement and the columns show the different user attributes.

These plots show, like the results in table 4.5, that the number of Tweets created by a user does not affect the resulting information flow trees overly much. Even the source count correlates only weakly with the number of Tweets. However, the number of Retweets has much more influence on the resulting information flow trees than just the Tweet count. This is probably caused by the fact that Retweets usually carry the same hashtags as the initial Tweet and are done within the 19 hour time frame. Thus, if a user creates a Retweet to one of her or his friend's Tweets, it is very likely that those two users are represented within an information flow tree. In addition, Retweets are more conversation based than mere Tweets.

Only the three best correlations (information spreader - followers, information flow trees - Retweets and source - followers) have spearman correlation coefficients above 0.5. That these correlations are the best is not really surprising. Users with more followers can more easily spread information and are more likely to be repeated by any of their followers. Retweets might form information flow trees on their own as explained in the previous paragraph. Overall, the user attributes do not influence the information flow trees for hashtags overly much. Hence, the answer to the third research question cannot be verified completely positive. The user attributes do influence the information flow trees, but only some of the shown comparisons correlate strongly.

5 Conclusion

This thesis analyzed a Twitter dataset of computer scientists. The thesis is mainly focused on the usage of Twitter by computer scientists and the hashtag usage within this dataset was particularly investigated. The Twitter dataset was created by crawling the Twitter-API for the Tweets of specific Twitter accounts. The Twitter accounts were selected based on results of previous studies of other authors.

First of all, the used Twitter dataset was examined as a whole. In the course of this examination, the social relationships between the Twitter users were investigated. These follower and friend relationships are not equally distributed through the set of users. Hence there are users connected to many others, while some users have only weak or none connections at all. Interestingly there is no apparent rule or tendency on how many of a user's followers share the user's research interests.

The network created by the follower relationships between the users plays an important role at studying the information distribution through the Twitter network. Further experiments indicated that this network has small world properties. The research area affiliations of the users are visibly represented in this network, but there are no real borders between the users of different research areas.

Moreover, the Tweets themselves were examined. Apparently most Tweets were created during workdays and probably during work, but that is partly speculation. The Tweet per user distribution was found to be highly skewed, since most users are seldom active, while a small subset of users creates a major part of all Tweets.

User interactions on Twitter take place by mentioning, retweeting and replying to each other. The results on the examination of these interactions show that most of these interactions are done between users affiliated to the same research areas. Interacting with followers and friends is also popular, but the popularity recedes over time. Users not connected by follower relationships or common research areas hardly interact with each other.

A large part of this thesis is dedicated to the analysis of hashtag usage. It was shown that the hashtag usage is not equally distributed. So there are some heavily used hashtags, while the majority of hashtags is only used a few times. The usage patterns of the more often used hashtags change frequently, but overall the hashtag usage stays constant. One of this thesis' hypotheses was that some hashtags might be specific to some research areas (**RQ 1**). A visualization of the area affiliation of some selected hashtags indicated, that there are indeed hashtags, which are for the most part used by users affiliated to some particular research areas. Furthermore, an automatic approach was presented, which allows assigning the hashtags into area specific and general hashtags.

A substantial part of this thesis attends to the analysis of the diffusion of hashtags through the Twitter network. In order to represent this information flow (**RQ 2**), information flow trees were proposed. Information flow trees model the flow of pieces of information through a network. In this case, the follower network is used. This way information flow trees show how users are influenced by other users to use certain hashtags. Most important for the creation of information flow trees is a good choice for a time frame. This time frame determines the maximum allowed time between two hashtag usages to get included into an information flow tree. By investigating the time between the creation of a Tweet and its Retweets, a time frame of nineteen hours was chosen.

After the optimal parameters for the creation of information flow trees were identified, the information flow trees of some selected hashtags were analyzed. Thereby the created trees, their length and duration were compared. These information flow trees differ quite much and there is no evident correlation between the created trees and the usage patterns. Thereafter the information flow trees for a greater set of hashtags were calculated in order to find an answer to the third research question (**RQ 3**). This hypothesis states that the measurements gained from the information flow trees correlate with certain user attributes. The executed analysis showed that there are indeed features correlating with each other, but most correlations were found to be quite weak. However, there were still some notable correlations: Users with many followers are usually good information spreaders and often create new information and users, who create many Retweets, are usually found in many information flow trees.

5.1 Future work

This thesis is based on the usage of hashtags within a scientific community. The insights gained by this thesis lead to new aims and questions which might be approached in future works. These future works are explained shortly in the following:

The information flow trees described in this thesis could be used to track other pieces of information than hashtags. For instance, URLs or buzzwords could be tracked and the data source for these trees is not limited to Twitter data, so data from other online social networks could be used as well. The information flow trees themselves could also get enhanced. For instance, the links could get weighted by considering the time difference and possible repetitions. Hence, if two usages occurred closely after each other, the link would be tighter, than for two usages which are apart by almost the whole allowed time frame.

The analysis of the existing information flow trees is also not exhaustive. For instance, the way how information gets distributed on Twitter could get evaluated more closely. When looking at the information flow tree of a specific hashtag, most usages within the tree might originate from Retweets rather than ordinary Tweets. This way, a classification between information spreading and information creating activities might be possible. Furthermore, the user roles could be looked closer at. Currently only the source and sink count and the user's efficiency at spreading information is measured. However, there is no classification of users into information sources, sinks or spreaders.

In order to provide better meaning closer to reality information flow trees, only Tweets of the same language could be considered for a possible transfer of information. However, users using different language might rarely communicate and thus seldom follow each other. Yet, the links between users originating from different language groups are forming another research question. Another possible way to get better information flow trees is the grouping of hashtags which are standing for the same topics or events. Examples of this would be grouping the hashtags *fb* and *facebook* or *charliehebdo* and *jesuischarlie*. Eliminating possible typing errors in the hashtags is also a possible way to improve the information flow trees.

Information flow trees represent the observed flow of pieces of information through a social network. However, a prediction experiment is also conceivable. This way, the information flow of some piece of information could get predicted. The results of this thesis show that there are attributes which influence the information flow trees, but the correlation is not very strong. Maybe a combination of features or additional information like the user's research area and interests could help to get a good prediction.

5.2 Real world contributions

Mainly, this work is a scientific work, but its results could be used for other projects, which are not building thereof, as well. Based on the thesis' results and possible future work, commercial usage might be possible, too. However, to use this work commercially several adjustments would have to be made.

The Twitter crawler used in this thesis could be easily adjusted to crawl other Twitter datasets. Different crawling strategies should be possible, due to the crawler's modularity. While a commercial usage is unlikely, the crawler could be used for other scientific projects.

Another, but more hypothetical, usage is to use this thesis' results for marketing and promotion. Since the results allows to find the most effective information spreader and to detect information sinks, this knowledge could be used to help promoting new products more effectively.

Acknowledgments

First, I would like to thank my supervisors Elisabeth Lex and Subhash Pujari for supporting me during the process of creating this thesis.

I also want to thank my friends and fellow students for working together with me on numerous projects and assignments during studying. Special thanks go hereby to David, Markus and Philipp, who worked with me on most subjects and were always ready to share their ears and minds.

Thanks also go to my family and friends who always supported me and often helped me throughout the course of my studies to order my thoughts and free my mind. In particular I want to thank my parents who let me study without any pressure and my sister for correcting language errors not just in this thesis but also in several other written essays.

Bibliography

Abel, Fabian et al. (2012). "Twitcident: Fighting Fire with Information from Social Web Streams." In: *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*. New York, New York, USA: ACM Press, p. 305. URL: http://dl.acm.org/citation.cfm?id=2187980.2188035 (cit. on p. 5).

Ainsworth, Scott G. et al. (2011). "How much of the web is archived?" In: *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries - JCDL '11*. New York, New York, USA: ACM Press, p. 133. URL: http://dl.acm.org/citation.cfm?id=1998076.1998100 (cit. on p. 13).

Anagnostopoulos, Aris, Ravi Kumar, and Mohammad Mahdian (2008). "Influence and correlation in social networks." In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*. New York, New York, USA: ACM Press, p. 7. URL: http://dl.acm.org/citation.cfm?id=1401890.1401897 (cit. on p. 9).

Arnaboldi, Valerio et al. (2014). "Information diffusion in OSNs: the Impact of Nodes' Sociality." In: *Proceedings of the 29th Annual ACM Symposium on Applied Computing - SAC '14*. New York, New York, USA: ACM Press, pp. 616–621. ISBN: 9781450324694. DOI: 10.1145/2554850.2555000. URL: http://dl.acm.org/citation.cfm?id=2554850.2555000 (cit. on p. 13).

Barbieri, Nicola, Francesco Bonchi, and Giuseppe Manco (2013). "Cascade-based community detection." In: *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, p. 33. DOI: 10.1145/2433396.2433403. URL: http://dl.acm.org/citation.cfm?doid=2433396.2433403 (cit. on p. 14).

Cha, Meeyoung et al. (2010). "Measuring User Influence in Twitter : The Million Follower Fallacy." In: *International AAAI Conference on Weblogs and Social Media*, pp. 10–17. ISBN: 9781450304931. DOI: 10.1.1.167.192. URL: http://www.icwsm.org/2010/ (cit. on p. 10).

Cheng, Justin et al. (2014). "Can cascades be predicted?" In: *Proceedings of the 23rd international conference on World wide web*, pp. 925–936. ISBN: 9781450327442. DOI: 10.1145/2566486.2567997. arXiv: 1403.4608. URL: http://dl.acm.org/citation.cfm?id=2567997$%5Cbackslash$nhttp://dl.acm.org/citation.cfm?doid=2566486.2567997 (cit. on p. 12).

Eysenbach, Gunther (2011). "Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact." In: *Journal of medical Internet research* 13.4, e123. ISSN: 1438-8871. DOI: 10.2196/jmir.2012. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3278109%5C&tool=pmcentrez%5C&rendertype=abstract (cit. on p. 6).

García, Ruth et al. (2015). "Language, Twitter and Academic Conferences." In: p. 4. URL: http://arxiv.org/abs/1504.03374 (cit. on p. 7).

Gephi.org (2015). *Gephi - The Open Graph Viz Platform*. URL: http://gephi.github.io/ (visited on 10/20/2015) (cit. on p. 48).

Golbeck, Jennifer, Justin M. Grimes, and Anthony Rogers (2010). "Twitter use by the U.S. Congress." In: *Journal of the American Society for Information Science and Technology*, n/a–n/a. ISSN: 15322882. URL: http://doi.wiley.com/10.1002/asi.21344 (cit. on p. 5).

Gold, David et al. (1956). *Personal Influence: The Part Played by People in the Flow of Mass Communications*. DOI: 10.2307/2088435 (cit. on p. 9).

Gomez Rodriguez, Manuel, Jure Leskovec, and Andreas Krause (2010). "Inferring networks of diffusion and influence." In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*. New York, New York, USA: ACM Press, pp. 1019–1028. ISBN: 9781450300551. DOI: 10.1145/1835804.1835933. URL: http://dl.acm.org/citation.cfm?id=1835804.1835933 (cit. on pp. 2, 11).

Guille, Adrien and Hakim Hacid (2012). "A predictive model for the temporal dynamics of information diffusion in online social networks." In: *Proceedings of the 21st international conference ...* Pp. 16–20. URL: http://dl.acm.org/citation.cfm?id=2188254 (cit. on p. 12).

Guille, Adrien, Hakim Hacid, et al. (2013). "Information diffusion in online social networks: A survey." In: *ACM SIGMOD Record* 42.2, pp. 17–28 (cit. on p. 11).

Hadgu, Asmelash Teka and Robert Jäschke (2014). "Identifying and analyzing researchers on twitter." In: *Proceedings of the 2014 ACM conference on Web science - WebSci '14*. New York, New York, USA: ACM Press, pp. 23–32. ISBN: 9781450326223. DOI: 10.1145/2615569.2615676. URL: http://dl.acm.org/citation.cfm?id=2615569.2615676 (cit. on pp. 1, 37).

Honey, C. and S.C. Herring (2009). "Beyond Microblogging: Conversation and Collaboration via Twitter." English. In: *2009 42nd Hawaii International Conference on System Sciences*. IEEE, pp. 1–10. URL: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=4755499 (cit. on p. 6).

Jacomy, Mathieu et al. (2011). "Forceatlas2, a continuous graph layout algorithm for handy network visualization." In: *Medialab center of research* 560 (cit. on p. 48).

Java, Akshay et al. (2007). "Why we Twitter: Understanding microblogging usage and communities." In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis - WebKDD/SNA-KDD '07*. New York, New York, USA: ACM Press, pp. 56–65. URL: http://dl.acm.org/citation.cfm?id=1348549.1348556 (cit. on p. 9).

Kitsak, Maksim et al. (2010). "Identification of influential spreaders in complex networks." In: *Nature Physics* 6.11, pp. 888–893. ISSN: 1745-2473. DOI: 10.1038/nphys1746. URL: http://dx.doi.org/10.1038/nphys1746 (cit. on p. 13).

Larsson, A. O. and H. Moe (2011). "Studying political microblogging: Twitter users in the 2010 Swedish election campaign." In: *New Media & Society* 14.5, pp. 729–747. ISSN: 1461-4448. URL: http://nms.sagepub.com/content/14/5/729.short (cit. on p. 5).

Leavitt, Alex et al. (2009). "The Influentials: New Approaches for Analyzing Influence on Twitter." In: *Web Ecology* 04, p. 18. URL: http://www.webecologyproject.org/wp-content/uploads/2009/09/influence-report-final.pdf (cit. on p. 10).

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg (2009). "Meme-tracking and the dynamics of the news cycle." In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*. New York, New York, USA: ACM Press, p. 497. URL: http://dl.acm.org/citation.cfm?id=1557019.1557077 (cit. on p. 12).

Leskovec, Jure, Ajit Singh, and Jon Kleinberg (2006). "Patterns of Influence in a Recommendation Network." English. In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Wee-Keong Ng et al. Vol. 3918. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 380–389.

ISBN: 978-3-540-33206-0. DOI: 10.1007/11731139_44. URL: http://dx.doi.org/10.1007/11731139_44 (cit. on pp. 2, 11).

Ma, Zongyang, Aixin Sun, and Gao Cong (2012). "Will this #hashtag be popular tomorrow?" In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12*. New York, New York, USA: ACM Press, p. 1173. ISBN: 9781450314725. DOI: 10.1145/2348283.2348525. URL: http://dl.acm.org/citation.cfm?id=2348283.2348525 (cit. on p. 12).

Mendoza, Marcelo, Barbara Poblete, and Carlos Castillo (2010). "Twitter under crisis: Can we trust what we RT?" In: *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*. New York, New York, USA: ACM Press, pp. 71–79. URL: http://dl.acm.org/citation.cfm?id=1964858.1964869 (cit. on p. 5).

Milgram, Stanley (1967). "The small world problem." In: *Psychology today* 2.1, pp. 60–67 (cit. on p. 49).

Page, Lawrence et al. (1999). *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Previous number = SIDL-WP-1999-0120. Stanford InfoLab. URL: http://ilpubs.stanford.edu:8090/422/ (cit. on p. 10).

Priem, Jason and Kaitlin Light Costello (2010). "How and why scholars cite on Twitter." In: *Proceedings of the American Society for Information Science and Technology* 47.1, pp. 1–4. ISSN: 00447870. DOI: 10.1002/meet.14504701201. URL: http://doi.wiley.com/10.1002/meet.14504701201%20https://www.asis.org/asist2010/proceedings/proceedings/ASIST%5C_AM10/submissions/201%5C_Final%5C_Submission.pdf (cit. on p. 6).

Pujari, Subhash Chandra et al. (2015). "Social Activity versus Academic Activity: A Case Study of Computer Scientists on Twitter." In: *Proceedings of the 15th International Conference on Knowledge Technologies and Data-Driven Business (i-KNOW 2015)* (cit. on pp. 1, 35–37).

Rao, Delip et al. (2010). "Classifying latent user attributes in twitter." In: *Proceedings of the 2nd international workshop on Search and mining user-generated contents - SMUC '10*. New York, New York, USA: ACM Press, p. 37. ISBN: 9781450303866. DOI: 10.1145/1871985.1871993. URL: http://dl.acm.org/citation.cfm?id=1871985.1871993 (cit. on p. 9).

Rattanaritnont, Geerajit, Masashi Toyoda, and Masaru Kitsuregawa (2012). "Analyzing patterns of information cascades based on users' influence and posting behaviors." In: *Proceedings of the 2nd Temporal Web Analytics Workshop on - TempWeb '12*. New York, New York, USA: ACM Press, p. 1. ISBN: 9781450311885. DOI: 10.1145/2169095.2169097. URL: http://dl.acm.org/citation.cfm?id=2169095.2169097 (cit. on p. 13).

Reinhardt, Wolfgang et al. (2009). *How People are using Twitter during Conferences*. URL: http://lamp.tu-graz.ac.at/~i203/ebner/publication/09%5C_edumedia.pdf (visited on 06/12/2015) (cit. on p. 8).

Rogers, Everett M. (1962). "Diffusion of innovations." In: *The Free Press* (cit. on p. 9).

Romero, Daniel M., Brendan Meeder, and Jon Kleinberg (2011). "Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter." In: *Proceedings of the 20th international conference on World wide web - WWW '11*. New York, New York, USA: ACM Press, p. 695. ISBN: 9781450306324. DOI: 10.1145/1963405.1963503. URL: http://dl.acm.org/citation.cfm?id=1963405.1963503 (cit. on p. 12).

Ross, C. et al. (2011). "Enabled backchannel: conference Twitter use by digital humanists." en. In: *Journal of Documentation* 67.2, pp. 214–237. ISSN: 0022-0418. URL: http://www.emeraldinsight.com/doi/full/10.1108/00220411111109449 (cit. on pp. 8, 10).

Sadikov, Eldar et al. (2011). "Correcting for Missing Data in Information Cascades." In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM '11. Hong Kong, China: ACM, pp. 55–64. ISBN: 978-1-4503-0493-1. DOI: 10.1145/1935826.1935844. URL: http://doi.acm.org/10.1145/1935826.1935844 (cit. on pp. 2, 11).

SalahEldeen, Hany M. and Michael L. Nelson (2012). "Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?" In: p. 12. arXiv: 1209.3026. URL: http://arxiv.org/abs/1209.3026 (cit. on p. 12).

Sanderson, Robert, Mark Phillips, and Herbert Van de Sompel (2011). "Analyzing the Persistence of Referenced Web Resources with Memento." In: p. 4. URL: http://arxiv.org/abs/1105.3459 (cit. on p. 13).

Spearman, Charles (1904). "The proof and measurement of association between two things." In: *The American journal of psychology* 15.1, pp. 72–101 (cit. on p. 29).

Sun, Beiming and Vincent TY Ng (2012). "Identifying influential users by their postings in social networks." In: *Proceedings of the 3rd international* *workshop on Modeling social media - MSM '12*. New York, New York, USA: ACM Press, p. 1. ISBN: 9781450314022. DOI: 10.1145/2310057.2310059. URL: http://dl.acm.org/citation.cfm?id=2310057.2310059 (cit. on p. 10).

Tumasjan, Andranik et al. (2010). "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." In: *ICWSM*, pp. 178–185. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852 (cit. on p. 5).

Twitter, Inc (2015). *GET search/tweets — Twitter Developers*. URL: https://dev.twitter.com/rest/reference/get/search/tweets (visited on 10/13/2015) (cit. on p. 39).

Watts, D J and S H Strogatz (1998). "Collective dynamics of 'small-world' networks." In: *Nature* 393.6684, pp. 440–2. ISSN: 0028-0836. DOI: 10.1038/30918. URL: http://dx.doi.org/10.1038/30918 (cit. on p. 49).

Weller, Katrin, Evelyn Dröge, and Cornelius Puschmann (2011). "Citation Analysis in Twitter : Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences." In: *1st Workshop on Making Sense of Microposts*, pp. 1–12. URL: http://www.researchgate.net/profile/Katrin%5C_Weller/publication/228405783%5C_Citation%5C_Analysis%5C_in%5C_Twitter.%5C_Approaches%5C_for%5C_Defining%5C_and%5C_Measuring%5C_Information%5C_Flows%5C_within%5C_Tweets%5C_during%5C_Scientific%5C_Conferences/links/02e7e51d16aec47490000000.pdf (cit. on p. 8).

Wen, Xidao, Yu-Ru Lin, et al. (2014). "Twitter in academic conferences: Usage, Networking and Participation over Time." In: *Proceedings of the 25th ACM conference on Hypertext and social media - HT '14*. New York, New York, USA: ACM Press, pp. 285–290. ISBN: 9781450329545. DOI: 10.1145/2631775.2631826. URL: http://dl.acm.org/citation.cfm?id=2631775.2631826 (cit. on pp. 1, 7).

Wen, Xidao, Denis Parra, and Christoph Trattner (2014). "How groups of people interact with each other on Twitter during academic conferences." In: *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW Companion '14*. New York, New York, USA: ACM Press, pp. 253–256. URL: http://dl.acm.org/citation.cfm?id=2556420.2556485 (cit. on p. 7).

Weng, Jianshu et al. (2010). "TwitterRank: Finding Topic-sensitive Influential Twitterers." In: *Proceedings of the third ACM international conference on Web* *search and data mining - WSDM '10*. New York, New York, USA: ACM Press, p. 261. URL: http://dl.acm.org/citation.cfm?id=1718487.1718520 (cit. on p. 10).

Wu, Shaomei et al. (2011). "Who says what to whom on twitter." In: *Proceedings of the 20th international conference on World wide web - WWW '11*. New York, New York, USA: ACM Press, p. 705. ISBN: 9781450306324. DOI: 10.1145/1963405.1963504. URL: http://dl.acm.org/citation.cfm?id=1963405.1963504 (cit. on p. 9).

Yin, Zibin et al. (2012). "Discovering patterns of advertisement propagation in Sina-Microblog." In: *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy - ADKDD '12*. New York, New York, USA: ACM Press, pp. 1–9. ISBN: 9781450315456. DOI: 10.1145/2351356.2351357. URL: http://dl.acm.org/citation.cfm?id=2351356.2351357 (cit. on p. 13).

Zhao, Dejin and Mary Beth Rosson (2009). "How and Why People Twitter: The Role that Micro- blogging Plays in Informal Communication at Work." In: *Proceedinfs of the ACM 2009 international conference on Supporting group work - GROUP '09*. New York, New York, USA: ACM Press, p. 243. URL: http://dl.acm.org/citation.cfm?id=1531674.1531710 (cit. on p. 6).

Zubiaga, Arkaitz et al. (2011). "Classifying trending topics: A Typology of Conversation Triggers on Twitter." In: *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*. New York, New York, USA: ACM Press, p. 2461. URL: http://dl.acm.org/citation.cfm?id=2063576.2063992 (cit. on p. 11).