Master's thesis

Time Series Analysis of Online Social Network Data and Content

Philipp Singer

Knowledge Management Institute
Graz University of Technology
Head: Univ.-Prof. Dr.rer.nat. Klaus Tochtermann

Supervisor: Univ.-Ass. Dipl.-Ing. Dr.techn. Markus Strohmaier
Advisor: Dipl.-Ing. Claudia Wagner

Graz, 26. September 2011

Masterarbeit

Zeitverlaufsanalysen von sozialen Netzwerkdaten und -inhalten am World Wide Web

Philipp Singer

Knowledge Management Institute
Technische Universität Graz
Vorstand: Univ.-Prof. Dr.rer.nat. Klaus Tochtermann

Begutachter: Univ.-Ass. Dipl.-Ing. Dr.techn. Markus Strohmaier
Betreuerin: Dipl.-Ing. Claudia Wagner

Graz, 26. September 2011

Abstract

Social networks have been an upcoming trend over the last few years and they have gained an established position in today's World Wide Web. With the rise of this new media movement, the analysis of social networking has become an essential part of web science. Recent research aimed to examine influences of social networks such as Twitter. These previous studies focused on structural properties of social networks as well as on the related content. Thus, the present thesis applies a framework presented by Wang and Groth [WG10] in order to detect bi-directional links between social and content network properties of individuals to two different Twitter datasets and an Irish bulletin board dataset named Boards.ie. The analysis of the two Twitter datasets suggests that the number of followers of a particular Twitter user has a strong influence on individual content network properties. Furthermore, the results show that the Twitter users in our datasets are becoming more active every day. The final study of this work on the Boards.ie dataset illustrates that the users of the forum are becoming more involved in the forum over time and seem to expand their repertoire of different topics used throughout their posts. This work should lead to a better understanding of influences between social and content network properties in social networks and can be considered as a stepping stone for further investigation.

Kurzfassung

Soziale Netzwerke sind einer der aufkommenden Trends der letzten Jahren gewesen und haben sich eine wichtige Position im World Wide Web erarbeitet. Auf Grund des Anstiegs dieser Bewegung neuer Medien, ist die Analyse sozialer Netzwerke ein integraler Teil der Web-Forschung geworden. Kürzliche Forschung hat sich auf die Bestimmung diverser Einflüsse in sozialen Netzwerken wie Twitter konzentriert, wohingegen sich solche Analysen nicht nur auf strukturelle Eigenschaften von sozialen Netzwerken konzentrieren, sondern auch auf den zugehörigen Inhalt. Diese Arbeit wendet folglich ein Framework von Wang und Groth [WG10] zur Erkennung bi-direktionaler Verbindungen zwischen Social- und Content-Netzwerk Eigenschaften von Usern auf zwei verschiedene Twitter Datensätze und einem Datensatz eines großen Forums namens Boards.ie an. Die Analyse an den zwei Twitter Datensätzen zeigt, dass die Zahl der Follower eines Twitterers einen großen Einfluss auf verschiedene Content-Netzwerk Eigenschaften hat. Darüber hinaus legen die Ergebnisse nahe, dass die Twitter Nutzer in unseren Datensätzen jeden Tag zunehmend aktiver werden. Die letzte Studie dieser Arbeit behandelt den Boards.ie Datensatz und stellt fest, dass sich die Nutzer des Forums im Laufe der Zeit immer mehr am Forum beteiligen und sie scheinen auch ihr Repertoire an unterschiedlichen Themen in ihren Beiträgen zu erweitern. Diese Arbeit ist ein Schritt hin zu einem besseren Verständnis der Einflüsse zwischen Social- und Content-Netzwerk Eigenschaften in sozialen Netzwerken und stellt ein Sprungbrett für weitere Untersuchungen dar.

TU Graz
Graz University of Technology
Senat

Deutsche Fassung:
Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008
Genehmigung des Senates am 1.12.2008

EIDESSTATTLICHE ERKLÄRUNG

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 entnommene Stellen als solche kenntlich gemacht habe.

Graz, am ……………………………. ………………………………………………..
(Unterschrift)

Englische Fassung:

STATUTORY DECLARATION

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.

……………………………. ………………………………………………..
date (signature)

Danksagung

Diese Masterarbeit wurde im Jahr 2011 am Knowledge Management Institute an der Technischen Universität Graz durchgeführt.

Zu Beginn möchte ich mich sehr herzlich bei meinem Begutachter Dr. Markus Strohmaier bedanken. Er stand mir während des Verfassens der Arbeit immer mit Rat und Tat zur Seite und hat mich unterstützt, wo er nur konnte. Seine Einwürfe und sein Feedback haben die Arbeit erst zu dem gemacht, was sie jetzt ist, und ohne seine exzellente Unterstützung wäre diese Arbeit nicht möglich gewesen.

Ein weiteres ganz großes Dankeschön geht an meine Betreuerin Claudia Wagner, die mir im Laufe der Arbeit laufend Feedback und Einwürfe gegeben hat und vor allem in der schwierigen Anfangszeit eine enorme Stütze gewesen ist. Trotz Zeitmangel hat sie sogar am Ende eine Nacht durchgearbeitet, um mir möglichst schnell ein finales Feedback liefern zu können.

An dieser Stelle auch ein herzliches Dankeschön an Shenghui Wang and Paul Groth, die mir das Framework dieser Arbeit zur Verfügung gelassen haben und die bei entsprechenden Fragen auch immer eine große Hilfe waren. Weiters möchte ich in diesem Zusammenhang John Breslin and Matthew Rowe danken, die mir den Boards.ie Datensatz ermöglicht habe.

Da diese Arbeit auch den Abschluss meines Masterstudiums darstellt, möchte ich mich hiermit auch bei meinen drei Studienkollegen Christoph, Jochen und Martin bedanken, mit denen ich das Studium seit dem ersten Semester bestritten habe. Vielen Dank für die Freundschaft und auch die exzellente Zusammenarbeit in den verschiedensten Kursen in unserem Studium.

Dankeschön auch an alle Leute, die sich die Mühe gemacht haben meine Arbeit, oder Teile davon anzusehen und mir Feedback zu geben.

Schlussendlich möchte ich mich auch noch bei meiner Familie und meiner Freundin Angelika bedanken. Ihr habt mich immer unterstützt, wenn es einmal nicht so gut lief und wart ein großer Motivator für mich, sowohl während meiner bisherigen Studienzeit, als auch während der Verfassung dieser Arbeit. Vielen Dank!

Graz, im September 2011 Philipp Singer

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Social networks have been an upcoming trend over the last few years and they have gained an established position in the World Wide Web. Alexa [Ale11], a comprehensive web service collection of user information on various websites, lists Facebook in second place and Twitter, a microblogging service containing up to 140 characters long status reports, in ninth place. In September 2010, Twitter CEO Evan Williams announced that Twitter already had 145 million registered users, and that it is also adding about 300.000 new users every day making it one of the most growing social networks in the world nowadays [Mas10].

Since social networking conquered the web, its analysis has become a compelling and essential part of recent web research. Especially semantic and pragmatic aspects of social networks offer exciting research topics. Conferences like ICWSM[1], WWW[2] or ISWC[3] cover many issues regarding the social semantic web every year. Twitter is also well-known for the use of so called hashtags (see section 2.2.1.3) and thus the dynamics of the Twitter network in particular have become increasingly interesting. Section 2.2.1 provides more details about the social network Twitter.

The main idea of this master's thesis is based on the work of Wang and Groth [WG10], who established a framework with which to measure the dynamic bi-directional influence between content and social networks. The authors highlight that the social semantic web often provides a social network as well as a content network of a user. It is interesting to discover connections between those networks and to observe as they change. Twitter offers an exciting potential for the study of this bi-directional influence. On the one hand, the social network of Twitter users consists of their individual followers and followees, while, on the other hand, their tweets offer a powerful content network, especially in combination with hashtags, retweets or replies. The research of Wang and Groth will be discussed in-depth in section 2.4.

[1] http://www.icwsm.org/2011/index.php

[2] http://www.www2011india.com/

[3] http://www.iswc.net

Over the last few years, several researchers were analyzing various datasets. Cha et al. [CHBG10], for example, carried out further research on social media by investigating user influence in a large Twitter dataset. Suh et al. [SHPC10] also worked with a Twitter dataset, but placed their focus on discovering factors that have an impact on retweeting. Matthew Rowe [Row11] tried to find connections between subscriber counts and several behavior features in a YouTube dataset. These papers agree with the fact that there are several influences between properties of social media. However, at the same time, they also show different results regarding the in-degree of these networks. A detailed description of these works can be found in section 2.2.2.

It is important to gain more insight into social networks and the influences between social and content network properties in order to detect factors and their impact on other parts of the network. An example would be: If Twitter developers want the retweet functionality to be used more often, they can take a look at different influence analyses in order to discover features and properties that influence and promote retweets.

Sample research questions of this thesis are: Does the rise of the number of followers of users improve their activity on posting tweets or content features such as hashtags? If a user is more popular, will her tweets be more often retweeted or will more users reply to the tweets? The experiments of this thesis illustrate that the number of a user's followers (representing the popularity) has the highest influence on other properties. The first Twitter experiment states that a high number of followers improves activity in posting tweets and getting more often retweeted. Furthermore, the second Twitter experiment also points out that the user is more often retweeted, when she gets more followers. Social media (SM) such as Twitter provides more and more users and content each second. These types of questions are of immense interest to many researchers due to the points outlined above. However, business could also benefit from the results by gaining a better understanding of the connections between content and social aspects of social networks [WG10].

1.2 Objective

One of the secondary objectives of this thesis is to present an overview of current Twitter research and social network analysis in general. Furthermore, time series analysis with all its facets should also be illustrated. The main target of this work is the collection of datasets and the analysis of bi-directional influences of different network properties of the corresponding social and content networks. These results should then be presented, evaluated and interpreted accordingly so that the background can be understood.

The following research questions constitute the main objectives of this work:

- How do social and content network properties influence each other in social media such as a public timeline based Twitter dataset? The analysis based on this question should reveal how different properties of the social and content part of social media influence each other. The dataset that should answer this question is crawled using the public timeline of Twitter.
- How do social and content network properties influence each other in social media such as a user list based Twitter dataset? The analysis of this question is similar to the question above. Nevertheless, another dataset, which is based on a public user list by a Twitter user, will be selected this time.
- Do the number of own replies to other posts and the number of replies by other users influence a user to become more of a generalist or specialist on a public board such as Boards.ie and if so, how? This question should give insight into the procedures of a public board like, in this case, Boards.ie. The analysis should reveal if there is a relationship between a user's reply-behavior and the topical focus of a user on a message board.

1.3 Contribution

The present thesis will provide the following overall results:

- This work expands the empirical investigations by Wang and Groth [WG10] by applying their framework to additional datasets based on Twitter and a public bulletin board called Boards.ie.
- In addition to the findings by Wang and Groth [WG10] the results highlight that there also exist many bi-directional influences between social and content network properties on Twitter.
- The work also illustrates that the framework by Wang and Groth [WG10] can be applied to the Boards.ie dataset in order to provide insight into the influence on generalist and specialist behavior.

1.4 Thesis Outline

In general, the thesis consists of five chapters. This introduction is followed by chapter 2, which gives a deeper insight into present related work regarding the issue of this work. The beginning of the related work chapter (see section 2.1) provides a deeper insight into social network theory. Afterwards, section 2.2 covers current research about social networks. This section includes an overview of Twitter and corresponding literature, related work regarding social network research focusing on the discovery of influences and finally, a detailed description of semantic and pragmatic network analysis. Furthermore, time series analysis and multilevel autoregression are presented in section 2.3 and in addition the work by Wang and Groth [WG10], which provides the basic idea of this work, is described in section 2.4.

Chapter 3 lists and describes all the produced datasets of this work. It is followed by chapter 4, which explains the actual experiments done with these crawled data sets. This chapter also covers the results and interpretations of the different experiments. Finally, chapter 5 concludes the thesis with a summary and suggestions for future work.

2 Related Work

2.1 Social Network Analysis

This chapter gives an overview of relevant literature in the field of social network theory. At first, some fundamentals about graphs are provided. As a second step some properties of graphs and networks, which are important for this work are explained. Finally, fundamentals about social networks are described.

2.1.1 Fundamentals of graph theory

Diestel [Die05] describes a graph as a pair G = (V,E) of sets such that $E \subseteq [V]^2$. The elements of V are the so called *vertices* (also called *nodes* or *points*, noted as v) of the graph G and the elements of E are the *edges* between the nodes. The edges between two nodes can be directed or undirected. Directed means that the edges are only pointing in one direction, whereas in an undirected graph, the edges are pointing in both directions [New03]. Diestel [Die05] furthermore states that the most-used way to draw a graph is by drawing a point for each node and showing the edges of the graph by drawing a line between two corresponding vertices. The set of all vertices of a graph G is referred to as V(G) and the set of all edges as E(G). It is also possible to state all edges in E at a node v with the notation of E(v).

Figure 2.1 shows a sample social undirected graph by Wasserman and Faust [WF94]. The graph could represent a social network component of Twitter, where each node refers to a user, and the edges represent the status, if a user is a friend (follows the other one and is followed by the other one) with another.

Graphs can be distinguished by its number of different types of vertices. If there is only one type of nodes in a network, it is called a *one-mode* (*unipartite*) network [New03]. Figure 2.1 is a sample for a one-mode graph, because all the nodes represent users. On the other hand, it is possible, that there are different kinds of nodes in a graph. Diestel [Die05] states that a graph G = (V,E) is called *r-partite*, if V admits a partition into r classes such that every edge has its ends in distinct classes. It is important that there is no link between two nodes of the same class in such graphs. So every edge has its ends in different classes. The most common *r-partite* graphs are *bipartite* graphs, where

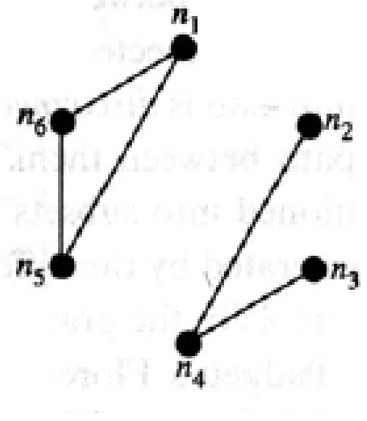


Figure 2.1: Sample Social Graph [WF94]

two dissimilar types of nodes are present. These networks are also called *two-mode* networks. For this work, only *one-mode* networks are important, because all of the social and content networks of the datasets presented in chapter 3 have only one type of nodes. If the social and content networks would be combined, they could represent *two-mode* networks.

2.1.2 Properties and metrics of graphs and social network analysis

There are several metrics for network analysis available. This section covers some of them and focuses on properties, which are relevant for the experiments of this work as described in chapter 4.

The following two metrics are used to better describe graphs.

Complete graph

According to Diestel [Die05] two vertices x,y are *neighbors*, or *adjacent*, if there exists an edge between x and y in the graph G. If all vertices of G are pairwise adjacent, then G is *complete*. This means that the graph has a single component. In the other case, when there is no edge for each pair of nodes available, the graph has multiple components. Figure 2.1 shows a sample graph with two components. So this graph is *incomplete*. If G = (V,E) is a non-empty graph, the set of neighbors of a node v in G is denoted by $\mathrm{N}_G(\mathrm{v})$.

Degree

According to Diestel [Die05] the degree N$_G$(v) = d(v) of a vertex v is the number |E(v)| of edges at v, which is equal to the number of linked neighbors of the node. Equation 2.1 shows the equation for the average degree of the complete graph.

$$d(G) = \frac{1}{|V|} \sum_{v \in V} d(v) \tag{2.1}$$

For a given vertex v also the *degree centrality* can be measured. The simplest definition according to Wasserman and Faust [WF94] is that the central actors must be the most active in the sense that they have the most ties to other actors in the network or graph. The degree centrality is equal to the degree of v divided by the maximum possible degree. So the degree centrality C$_D$(v) for the vertex v is [WG10]:

$$C_D(v) = \frac{d(v)}{n-1} \tag{2.2}$$

For directed graphs, there is for every vertex as well an *in-degree* (represented by the edges pointing to the node) and an *out-degree* (represented by the edges pointing to another node) [New03]. As a result of this, those graphs also have an average in-degree, an average out-degree, an in-degree centrality and an out-degree centrality.

The following properties are an excerpt of many different properties for social network analysis.

Betweenness centrality

According to Newman [New03] the *geodesic path* is the shortest path through a network from one node to another. It is possible that there exists more than one geodesic path between two nodes. The *betweenness centrality* of a vertex v now is the number of geodesics between other nodes that run through v. The main idea behind the betweenness centrality is that an actor is central if it lies between other actors on their geodesics [WF94]. Goh et al. [GOJ$^+$02] also state that "*the betweenness centrality is commonly used in sociology to quantify how influential a given person in a society is*". Wang and Groth [WG10] state that the betweenness centrality of a vertex is defined as the fraction of all shortest paths that pass through it over all shortest paths in the network. Equation 2.3 states the formula for the betweenness centrality, where σ_{st} is the number of shortest paths from v_s to v_t ($v_s, v_t \in$ V) and $\sigma_{st}(v)$ is the number of shortest paths from v_s to v_t that run through v [WG10].

$$C_B(v) = \sum_{v_s \neq v \neq v_t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.3}$$

Clustering Coefficient

Watts and Strogatz [WS98] introduced and defined the *clustering coefficient* in 1998 as follows. When a vertex v has k_v neighbors, then at most $k_v(k_v - 1)/2$ edges can exist between them. This case occurs when every neighbor of v is connected to every other neighbor of v. The clustering coefficient C_v describes now the number of actual links between the neighbors divided by the number of possible links between the neighbors of a vertex v. Due to the fact that the number of possible edges between the neighbors is different for directed and undirected graphs, there are two distinct equations. Equation 2.4 shows the clustering coefficient for directed graphs and equation 2.5 for undirected graphs [WG10].

$$C_v = \frac{|\{e_{jk}\}|}{k_i(k_i - 1)} : v_j, v_k \in N_G(v), e_{jk} \in E(G). \tag{2.4}$$

$$C_v = \frac{2|\{e_{jk}\}|}{k_i(k_i - 1)} : v_j, v_k \in N_G(v), e_{jk} \in E(G). \tag{2.5}$$

The average clustering coefficient of the complete graph or network is defined by Watts and Strogatz [WS98] as the sum of all local clustering coefficients divided by the number of vertices:

$$C = \frac{1}{n}\sum_{i=1}^{n} C_{v_i} \tag{2.6}$$

Watts and Strogatz [WS98] also state that these statistics have intuitive meanings for friendship networks. The local clustering coefficient C_v reflects the extend to which friends of v are as well friends of each other and the average clustering coefficient C measures the cliquishness of a typical friendship circle.

2.1.3 Fundamentals of social networks

Wasserman and Faust [WF94] describe a social network as a set of actors and the relations between those actors. "*Social network analysis provides a precise way to define social concepts, a theoretical alternative to the assumption of independent social actors, and framework for testing theories about structured social relationships*". Social network analysis anyway focuses more on the relations of the network than the actors in general. The methods provide explicit formal statements and measures of social structural properties. Furthermore, it would be impossible to describe measurements of structures and systems without relational concepts. Social network analysis is based on several key concepts. The following part describes some key concepts based on the definitions by Wasserman and Faust [WF94].

Actor

As already stated, social network analysis is interested in the relationships between *actors* of a network. The nodes of the network (graph) are called actors. In our sample Twitter network shown in figure 2.1 the nodes represent Twitter users, which are the actors of the network. The networks in this work focus on a collection of actors of the same type, which is a one-mode network. Some methods allow one to look at actors of conceptually distinct types of levels, or from different sets. More informations about the different modes of a network are provided in section 2.1.1.

Relational Tie

The actors of a social network are linked to one another by social ties, which represent the relationships of these actors. The definition of a tie is that it establishes a linkage between a pair of actors. The type of the tie can vary from network to network. In the sample Twitter network shown in figure 2.1 the ties represent friendships of the actors.

Dyad

At the most basic level, a relationship is described by a tie between two actors. A *dyad* exists of two actors and the tie between them. Social network research focuses now generally on the analysis of a dyad, to understand ties among pairs. So the dyad is most of the time the basic unit for statistical social network analysis.

Triad

Sometimes it is also necessary to study the relationships between more than two actors (*dyad*). Many social network analysis focuses on a *triad.* A triad is a subset of three different actors and the existing ties between them. The most important observations of triads are, if it is *transitive* or *balanced.* Transitive means that if an actor i "likes" actor j and actor j likes k, then actor i also likes actor k. Balanced, on the other hand, means that if actor i and j like each other, then they should be similar in their evaluation of some third actor k and if actor i and j dislike each other, then they should differ in their evaluation of a third actor k.

Subgroup and group

The analysis of different *subgroups* of a graph is an important aspect of social network analysis. A subgroup can be defined as any subset of actors and the ties between those actors. On the other hand, a *group* is "*the collection of all actors on which ties are being measured*". It is always necessary that a social network has a finite set of actors.

Relation

A *relation* is defined as "*the collection of ties of a specific kind among members of a group*". For example, the set of friendships over a group of Twitter users as shown in figure 2.1. For any group of actors it is possible to measure different kinds of relations.

Those different terms provide a core working vocabulary for discussing and analyzing social networks.

2.1.4 Categories of network data

In the last years, large-scale network data has become more important than ever. People are not only interested in analyzing small networks, like a network of employees, but also make observations about large network datasets. Easley and Kleinberg [EK10] have published a list of different sources and categories of large-scale network data that people have used for research. The categories are not distinct, and a single dataset can easily exhibit characteristics from several of them.

- **Collaboration Graphs:** Those graphs record who is working with whom in a specific domain. An example of a network of this kind is stated in section 2.4.2, where scientific co-authorships are analyzed.
- **Who-Talks-To-Whom Graphs:** Focus on the communication factor between a large set of actors.
- **Information Linkage Graphs:** Focus on data which stands out both in its scale and in the diversity of what the nodes represent. The data consists of billions of little pieces of information, with links wiring them together. Not only the information is of interest, but also the social and economic structures that stand behind the information. (i.e. snapshot of the web)
- **Technological Networks:** Represent technological aspects, where nodes are physical devices and edges are physical connections between them.
- **Networks in the Natural World:** Graph structures also occur in biology and other natural sciences. In this case, the representation can vary in the scale, ranging from the population level down to the molecular level.

2.2 Social Media Research and Twitter

Social network analysis has become a very popular and often covered topic in the last years. It generally views social relationships in terms of network theory consisting of nodes and edges. Using graph algorithms (see section 2.1.1 for further informations about graph theory), "*social network analysis characterizes the structure of social* *networks, strategic positions in these networks, specific sub-networks and decompositions of people and activities*" [Sco00]. Social network analysis has been applied to different Web 2.0 platforms (e.g. Facebook, Twitter, Wikis), but also to the whole Web, ontologies and the Semantic Web [WG10].

The roots of social network analysis can be found in the 1930's and 1940's where the psychiatrist Jacob Levy Moreno "invented" the *sociogram.* Wasserman and Faust [WF94] point out that this innovation was the beginning of *sociometry*, which can be seen as the precursor to nowadays social network theory. A sociogram is a picture, where persons are represented as points and the relationships between pairs of nodes are represented by lines, which link two corresponding points. Sociometry can be understood as the "*measurement of interpersonal relations in small groups*" or more detailed as "*the study of positive and negative affective relations, such as liking/disliking and friends/enemies among a set of people.*" To represent relational data so called *sociomatrices* are often used. The two dimensions for a sociomatrix are on the one hand the *sending actors* in the rows and on the other hand the *receiving actors* in the columns.

A very prominent experiment regarding social network analysis was performed by Travers and Milgram [TM69] in 1969 and was titled "An Experimental Study of the Small World Problem". The authors describe the small world problem exactly the following way:

> *The simplest way of formulating the small world problem is "what is the probability that any two people, selected arbitrarily from a large population, such as that of the United States, will know each other?" A more interesting formulation, however, takes account of the fact that, while persons a and z may not know each other directly, they may share one or more mutual acquaintances; that is, there may exist a set of individuals, B, (consisting of individuals b_1, b_2 ... b_n) who know both a and z and thus link them to one another: More generally, a and z may be connected not by any single common acquaintance, but by a series of such intermediaries, a-b-c- ... -y-z; i.e., a knows b (and no one else in the chain); b knows a and in addition knows c, c in turn knows d, etc.*

The experiment by Travers and Milgram [TM69] was built upon the goal of defining a single target person and a group of starting persons and then to generate a acquaintance chain from the starting persons to the target person. Each person out of the starting group received a document and should send it by mail to the target person. The persons got some information about the target and should only send the document to people they know and another restriction was also that the persons should always send the letter to the next link in the chain. The results show that overall 217 out of the 296 starting persons actually sent the document to friends and 64 out of 296 persons (29%) reached the target. Furthermore the analysis shows that chains based on geographic information reached the target's hometown or geographic near locations readily, but as soon as they reach this location they often circulate before they reach the final destination. Next to the complete chains it was also possible that the chains were incomplete, mostly based on the fact that the person who received the document did not send it on. The total 64 letters that reached the target were only sent by 26 different people, so common channels emerged for the complete chains. Another interesting concept is the "6 degrees of separation", which will be described in section 2.2.1.1.

"*Content analysis is a long used tool to determine the presence of certain words or concepts within texts and with content analysis tools social researches can make inferences about the content of the texts [WG10].*" When speaking of such content analysis the term *semantic networks* plays an important role. According to Monge and Contractor [MC03] this term is used for over three decades in the "*cognitive science, social psychology, and artificial intelligence literatures.*" The authors furthermore point out that in this literature the concept of semantic networks often refers to networks that are generated from texts. The nodes represent the words or ideas and the ties between the nodes represent the relations between these words or ideas. These networks are often generated by the use of automated software. A more modern concept of semantic networks was introduced by Monge and Eisenberg in 1987 [ME87]. Monge and Contractor [MC03] point out that the new aspects of the concepts of semantic networks by Monge and Eisenberg [ME87] was that they focus more on the "*shared interpreations that people have for message content*". The traditional semantic networks had a focus on the texts, whereas the new concepts focused on interpretations by people of these texts.

Monge and Contractor [MC03] define a semantic relation S_{ij} the following way. Person i for example can point out interpretations of a statement. These interpretations can then be coded into several different categories (M1, M2, ...). For each single individual these categories would be used as attributes. The semantic relation the can be defined as "*the degree to which two people share dissimilar interpretations on one or more such attributes of the mission.*" According to the authors this can be stated the following way:

$$S_{ij} = (M1_i - M1_j) + (M2_i - M2_j) + ... \tag{2.7}$$

Monge and Contractor [MC03] now state several properties of the semantic network based on the defined semantic relation:

- Densely connected network: Based on shared interpretations the nodes of the network are closely tied to each other.
- Semantic network clique: A group of people that have similar interpretations with each other but not with nodes outside this clique.
- High degree centrality: A person who has the same interpretations as many other persons.

- High betweenness centrality: A person that shares interpretations with others, but these others do not necessarily have to directly share them.

Monge and Contractor [MC03] also point out that extensions to semantic networks are always possible.

Wang and Groth [WG10] point out that in the last years researchers have tried to produce automated content analysis tools, which will probably also be the main research goal in the next years. The authors state that "*these two different analysis fields have been studied and researched in a mostly parallel style and only until recently, social scientists took the chance to try to combine social network analysis and content analysis*". There has been little work on the combination of these two classes of analyses, especially in Semantic Web Context. The main paper used in this work is by Wang and Groth [WG10]. It is one of the first papers to combine the two analysis fields in Semantic Web context. This work also focuses on the temporal analysis over content and social networks. The combination of social and content analysis of temporal data is also a less frequently analyzed topic.

This section is now split into three parts. At the beginning section 2.2.1 gives an overview about characters of Twitter and some recent research regarding Twitter. Section 2.2.2 covers work about social networks analysis. This section focuses on social network analysis, which covers the combination of social and content analysis and tries to find influence aspects in especially large-scale networks. Afterwards, the third section 2.2.3 covers emergent semantics and pragmatics in social media, so this section will focus more on the content of social networks.

2.2.1 Twitter - a brief introduction

As already stated in section 1.1 Twitter[1] is one of the biggest and most influenced social networks at the time. It is a microblogging service, which allows users to post status messages with a maximum of 140 characters. These so called tweets should give an answer to the by Twitter asked question "What's happening"? Figure 2.2 shows Twitter's current web interface.

2.2.1.1 Followers and Followees

The social part of Twitter is built upon a system called *followers*. You can follow a user you are interested in, and can observe all her posted tweets. The users you follow, are called *followees*. Beside that, other users can follow you, and so they can see your timeline. These users are the so called *followers*. If a user does not want that other

[1] `http://www.twitter.com`

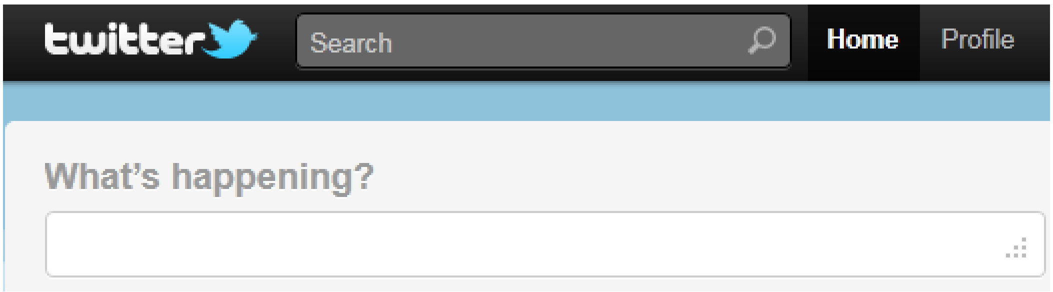


Figure 2.2: Twitter web interface

persons can follow and see all the *tweets* (a status report), it is possible to protect the own profile, and so only approved person can get insight to it. Figure 2.3 shows the current user interface for the individual timeline, showing all recent tweets by the users, whom the person is following.

The difference of the social aspect in Twitter to other social-networking sites like Facebook or MySpace is that it is not necessary to build a reciprocal relationship and this means that a user, who is followed by another person, does not need to follow the user back [KLPM10]. Anyway, many users on Twitter follow a follower blindly back, to establish a higher number of followers for both sides.

Kwak et al. [KLPM10] constructed a directed network, including information about followers and followees based on a large Twitter dataset to analyze its basic characteristics. Interestingly the results show that the follower distribution does not follow a power-law distribution, whereas most of the other social networks available on the web do follow a power-law distributions. Section 2.2.3 provides a deeper insight into the dynamics of social networks and also covers power-law effects. A big part of Twitter is the presence of many celebrities, who rapidly form online relations with their fans. Twitter includes 40 users (according to the work by Kwak et al. [KLPM10] in 2010) with more than a million followers and all of these users are celebrities or mass media. Further topological investigations by the authors point out that the majority of Twitter users who have fewer than 10 followers never tweeted or just tweeted once (some kind of trial) and so the median stays at one. It is shown that the average number of tweets by a user against her number of followers stays above the median. This indicates that there exist also outliners, who tweet much more than their number of followers would let us expect. Most of the top users with regards to the number of followers do not follow their followers back. Overall, Twitter shows a low level of reciprocity. Only 22,1% of the user pairs with a connection are reciprocal. The authors refer to such users as *r-friends*. Interestingly other studies regarding the reciprocity on other social networks show much higher values. Cha et al. [CMG09] report a reciprocity in

Flickr[2] of 68% and the authors of [KNT06] show that in Yahoo! 360[3] the bi-directional relationships make a total of 84%. Very interesting is also the high number of Twitter users (67,6%) who are not followed by any of their followees. This indicates that for these users Twitter represents an information tool.

Further analysis by Kwak et al. [KLPM10] provide insights into the degree of separation, which is a key indicator for the societal structure of a social network and it is based on the "six degrees of separation" experiment by Stanley Milgram [Mil67] (also see section 2.2 for further information). This famous experiment is also a motivation for many other researchers who analyze social networks. Stanley Milgram reports in his work that any two people in the world can be connected on average within six hops from each other. Watts and Strogatz [WS98] describe social and technological networks as "small-world" networks, because they also have small path lengths in many cases. Kwak et al. [KLPM10] now show that the average path length in Twitter is 4.12. The authors point out that this is very short for a large network like Twitter and the low percentage of reciprocity. This observation indicates that many Twitter users don't follow others for the social aspect, but more of getting information.

Another interesting tendency when looking at social networks is *homophily*. McPherson et al. [MSLC01] describe homophily as "*the principle that a contact between similar people occurs at a higher rate than among dissimilar people*". They also point out that "*the pervasive fact of homophily means that cultural, behavioral, genetic, or material information that flows through networks will tend to be localized*". Kwak et al. [KLPM10] took a look at homophily in Twitter on two aspects: geographic location and the number of r-friends' followers. The results show that users who have 1.000 or fewer followers are likely to be geographically close to their corresponding r-friends and also have a similar popularity. So when looking at bi-directional relationships in Twitter some level of homophily can be observed.

Huberman et al. [HRW08] analysed how relevant a list of friends is to the number of followers and followees based on a Twitter dataset. The authors define a friend as a person a user has directly communicated with at least two times. It is pointed out that the number of friends is a more accurate feature to determine a user's popularity than the number of followers. The results show that a user has just a few friends in comparison to her number of followers and followees. So a Twitter user is just directly communicating with a short number of the members of her network. This indicates that there are two actual networks existing in Twitter: 1) a dense network built by the followers and followees and 2) a sparser network built by the actual friends. Users with many friends post more posts on Twitter than those with fewer friends.

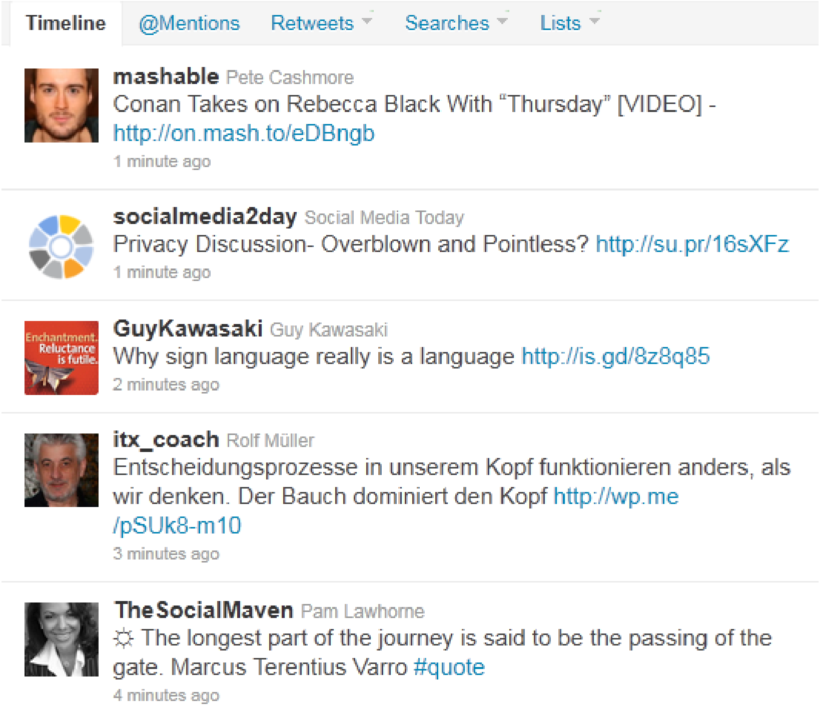
[2] `http://www.flickr.com/`

[3] `http://pulse.yahoo.com/`

2.2.1.2 Retweets, Replies and Mentions

Twitter offers the users a service, which is called *retweet*. As the name states, it is possible to copy another person's tweet. Most times it is copied without any further comments or personal opinion. The idea of retweets was originally created by users and furthermore implemented by some 3rd party clients. Twitter added the service in 2010. The Twitter web interface offers three different tabs for retweets [Twi11c]:

- Retweets by others
- Retweets by you
- Your tweets, retweeted



mashable Pete Cashmore
Conan Takes on Rebecca Black With "Thursday" [VIDEO] - http://on.mash.to/eDBngb
1 minute ago

socialmedia2day Social Media Today
Privacy Discussion- Overblown and Pointless? http://su.pr/16sXFz
1 minute ago

GuyKawasaki Guy Kawasaki
Why sign language really is a language http://is.gd/8z8q85
2 minutes ago

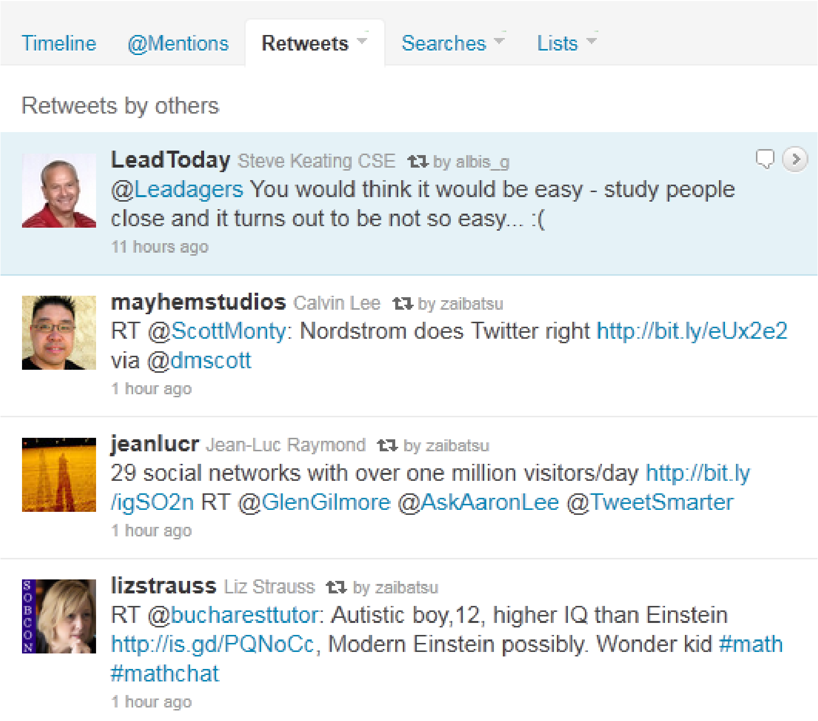
itx_coach Rolf Müller
Entscheidungsprozesse in unserem Kopf funktionieren anders, als wir denken. Der Bauch dominiert den Kopf http://wp.me/pSUk8-m10
3 minutes ago

TheSocialMaven Pam Lawhorne
☼ The longest part of the journey is said to be the passing of the gate. Marcus Terentius Varro #quote
4 minutes ago

Figure 2.3: Twitter user timeline

Figure 2.4 shows the web interface for the tab "Retweets by others". The interface also shows the person, who created the tweet in the first place. In [Twi11c] further informations can be found.

According to [Twi11b] on Twitter, it is also possible to directly *reply* to a Tweet posted by another user. Users are often saying a lot on Twitter, and if you want to say something back you can simply answer to that by using the reply functionality. The reply always begins with "@username", where username is replaced by the actual screen



Retweets by others

LeadToday Steve Keating CSE by albis_g
@Leadagers You would think it would be easy - study people close and it turns out to be not so easy... :(
11 hours ago

mayhemstudios Calvin Lee by zaibatsu
RT @ScottMonty: Nordstrom does Twitter right http://bit.ly/eUx2e2 via @dmscott
1 hour ago

jeanlucr Jean-Luc Raymond by zaibatsu
29 social networks with over one million visitors/day http://bit.ly/igSO2n RT @GlenGilmore @AskAaronLee @TweetSmarter
1 hour ago

lizstrauss Liz Strauss by zaibatsu
RT @bucharesttutor: Autistic boy,12, higher IQ than Einstein http://is.gd/PQNoCc, Modern Einstein possibly. Wonder kid #math #mathchat
1 hour ago

Figure 2.4: Twitter Retweets by others

name of the user you are replying to. It is also possible to just *mention* another user in the tweet, by adding "@username" to the text of the Tweet. So a reply itself is also considered as a mention. Figure 2.3 and Figure 2.4 show the available tab on the top for showing the mentions in the Twitter interface.

Next to the number of followers, a Twitter user's popularity can be also measured by retweet information. Kwak et al. [KLPM10] ranked Twitter users by the PageRank algorithm (see [BP98]) and also by the number of retweets to provide insights into the popularity. The ranking by the number of followers and by PageRank seems to be similar. On the other hand, the ranking on the number of retweets seems to differ from the other two rankings. The authors state that this indicates a "*gap in influence inferred from the number of followers and that from the popularity of one's tweets*". The ranking by retweets shows the rise of alternative media in Twitter.

Kwak et al. [KLPM10] furthermore state that retweets can also give good insights to how the information is spreading on Twitter. On Twitter, retweets provide the possibility for users that they acquire information from persons they don't follow directly. The authors constructed retweet trees and found that there exist interesting retweet patterns such as repetitive retweets and cross-retweets. Repetitive retweet means that the same tweet is retweeted repeatedly and cross-retweet describes the retweeting of each other. Further investigations regarding retweeting point out that any retweet reaches an average of 1000 Twitter users whatever the number of followers of the user is. As soon as a tweet is retweeted it gets almost instantly retweeted on the 2nd, 3rd and 4th hops away from the source. This is also shown by the constructed retweet trees.

Boyd et al. [BGL10] investigated conversational aspects of retweeting on Twitter. For the analysis, the authors used four different datasets. The first two datasets were crawled from the public timeline and provide quantitative context for the understanding of retweeting and the other two describe data, which is used more directly. The paper focuses on the restriction of the 140 characters per tweet and the corresponding problems with retweeting. The work shows that this constraint should not be seen as a limitation. It should be kept in mind that Twitter's own retweeting function solves the limitation problems. Many people shorten or modify the tweet in other ways to fit it into the 140 characters. Retweeting generally provides a simple possibility for produce, consume and share messages. People use retweeting for several reasons like spreading tweets to new audiences, entertaining a specific audience and much more. Twitter users retweet for others but also for social action. On the other hand, some users as well request a tweet to be retweeted by others when posting a status update and some users also retweet when people retweet messages that refer to them, what is called "ego retweets". A big problem of retweeting is that the retweet might not be an "*accurate portrayal of the original message*". When users alter the original tweet they might also alter the original meaning.

2.2.1.3 Hashtags

Golder et al. [GH06] state that the marking of content with descriptive terms is a frequently used way to organize and structure content. The terms are often also called keywords or tags. This tagging makes it easier to navigate, filter or search the content in the future. This is not entirely new, but the collaborative form has become very popular on the web and is called *tagging*.

Twitter itself does not support any tagging functionalities like for example Del.icio.us[4], so the users have developed an own tagging culture by adding a hash symbol (#) in front of short keywords [HTE10]. These keywords with their hashes are called *hashtags*. The first introduction to hashtags was made by Chriss Messian in a blog post [Mes07]. Huang et al. [HTE10] state that since this introduction of a new social tagging culture a complete new phenomenon, also called *micro-meme*, has been established. The tag selection on social tagging sites in general is most times an *a posteriori* approach. The participation in micro-memes is an *a priori* approach, because a single user would most likely not have written a tweet on a special topic, if she has not observed the micro-meme hashtag used by a different Twitter user.

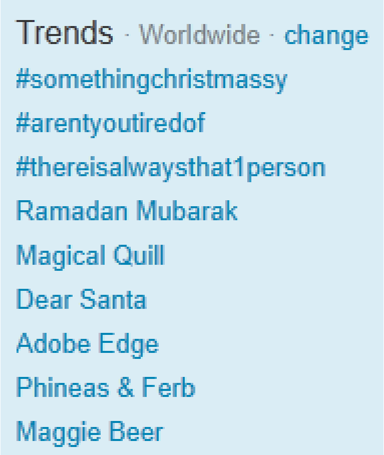
[4] http://www.delicious.com/

Both Figure 2.3 and Figure 2.4 show one tweet, which use different hashtags in their text to mark it.

In section 2.2.3 tags in general and collaborative tagging are covered in a greater detail. Furthermore, semantic and pragmatic aspects of tags are discussed in this section.

2.2.1.4 Trending Topics

Twitter collects keywords and hashtags, which are trending throughout the Twitter sphere and presents a top ten list of trending topics on the web interface. Figure 2.5 shows a sample list of Twitter trending topics. It is also possible to change the trend location to get a list of trending topics corresponding to the chosen location. It is not exactly known how Twitter is determining the trends, because they haven't made the algorithm public.



Trends · Worldwide · change

#somethingchristmassy

#arentyoutiredof

#thereisalwaysthat1person

Ramadan Mubarak

Magical Quill

Dear Santa

Adobe Edge

Phineas & Ferb

Maggie Beer

Figure 2.5: Twitter trending topics

Kwak et al. [KLPM10] compared Twitter trends with trends in other media. The results confirm that Twitter's role is a media for breaking news. Further investigations show that out of 41 million Twitter users, a very large number of users is participating in trending topics (about eight million), and about 15% participate in more than ten topics during four months. There seems to be core members who provide content over a long time for a certain topic (#iranelection for example). A trending topic also does not last forever, but it also does not disappear and never comes back. Most of the trending topics (about 73%) just have a single active period.

2.2.1.5 Twitter API

Twitter is providing developers an *Application Programming Interface* (API)[5], to be able to create applications, which are integrating with Twitter. It is possible to access all the features, which are offered by Twitter and which are already described in the past sections. For normal users the API requests are limited to 200 per hour, but it is possible to apply for a developer account, which has 20.000 allowed requests per hour. Due to the fact, that there are a lot of third-party applications using Twitter features up to date, Twitter doesn't give away a lot of these developer accounts any longer.

The Twitter servers have to handle up to 3 billion calls every day just to the API and 75% of all the traffic, comes from third-party applications [DuV10]. Third-party applications can range from standalone programs like Tweetdeck[6] to web mashups like Twittervision[7], which also combines different APIs (Twitter and Google Maps).

2.2.1.6 Further Twitter Research

Teevan et al. [TRM11] explored the search behavior on Twitter. In this paper the authors compare the search behavior on Twitter with the behavior on web search engines like Google or Bing. The studies are based on a dataset consisting of search queries of the same users for Twitter and web search. As a first step of this work a questionnaire on 54 Twitter users regarding the motivation for Twitter search is presented. The results show that people search Twitter for discovering temporally relevant information and information related to people. The search query analysis reveals that especially search queries containing micro-memes, Twitter users and celebrity names are popular. In comparison to web search queries the Twitter queries are generally short, but they contain longer words, more specialized syntax and more references to people. Twitter search seems to be generally used for monitoring content, whereas web search is most of the time used to develop and learn about a topic. Nevertheless, some persons issue the same query to both Twitter search and web search. The results also reveal that "*Twitter queries are more common, repeated more and change less than web queries*". Further investigations regarding the search results reveal that Twitter results include more social content and events, whereas the web search results contain more facts and navigation. The language that is used by Twitter results and web results is also very different.

Another example for a further research topic is the Twitter timeline. For example, Shamma et al. [SKC10] investigated applications of existing methods to discover the structure and content of media events on Twitter.

[5] `http://dev.twitter.com/`

[6] `http://www.tweetdeck.com/`

[7] `http://twittervision.com/`

Further research about the influence on Twitter is provided in the next section 2.2.2.

2.2.2 Social Network Analysis of Social Media

According to Cha et al. [CHBG10] the study of influence is not a new research field, but has a long history in the fields of sociology, communication, marketing and political science. Understanding influence plays an important role in many business areas, and influence is also important for the function of a society. Studying influence patterns can as well help to identify and understand trends or innovations. Whereas the study of influence is often very difficult, there has been a lot of research about it. Many researches speak about so called influentials, a minority group of users, who have a high influence on others and also persuade them. A more modern view states that the influence is based on the "*interpersonal relationship among ordinary users*" and the "*readiness of a society to adopt an innovation*".

In social media directed links could represent anything from intimate friendship to common interests. These links determine the flow of information and as a result of that indicate a user's influence on others. Cha et al. [CHBG10] measured the user influence in Twitter on a large dataset. This work has a very similar aspect as the work done in this thesis, whereas it rather focuses on the influence of a Twitter user on another Twitter user. The authors define the Twitter network as a news spreading medium and study the types and degrees of influence within the network. The influence is based on the fact that an individual might have the "*potential to lead others to engage in a certain act*". The authors highlight three "interpersonal" activities on Twitter. First, users can interact by following other users and so have the ability to read their tweets (see section 2.2.1.1). Second, users can share these tweets by other users by retweeting them (see section 2.2.1.2). Finally users can response to other people's tweets (see section 2.2.1.2). So they study the in-degree (number of followers), retweet (number of retweets) and mention (number of mentions) influence. The different aspects of their studies show that they focus on social and content network features and try to find influences between them.

As already mentioned Cha et al. [CHBG10] collected a huge Twitter dataset and they tried to gather all possible information of all Twitter users for their dataset. The two Twitter datasets of this work (see section 3.1 and section 3.2) just pick out a small sample of Twitter, because we had not the crawling possibilities to gather a huger dataset. As a result of this crawling process they had a Twitter dataset consisting of six million users. For the comparison of user influence the Sperman's rank correlation coefficient was used. As a first analysis step Cha et al. [CHBG10] tried to identify the top influentials out of the dataset on the three different aspects. The analysis showed that across all the three measures, the top influentials were mostly well-known public figures or websites, but there was nearly no overlap between the three top lists.

Figure 2.6 shows the marginal overlap of the top-100 lists. This indicates that the three measures focus on a different type of influence. Regardless, it is interesting to see, that the discovered influential Twitter users in this work can be well compared to results of an influence measure called TunkRank[8]. TunkRank is a similar measure like Google's PageRank to measure the influence of Twitter users. The top list of the TunkRank measure has a lot of users in common as the top influential list of this thesis. Similar to this Weng et al. [WLJH10] also tried to extend the PageRank measure to indicate influence in Twitter. The authors found a high link reciprocity from a non random sample of Singapore Twitter users, and they argued that this high reciprocity is indicative for homophily. As a result of this they exploited this fact to measure the influence.

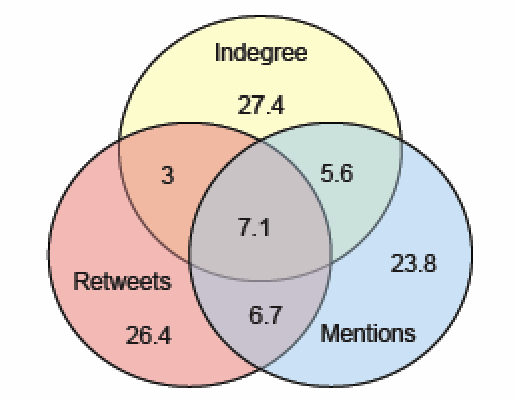


Figure 2.6: Venn diagram of the top-100 influentials across measures [CHBG10]

Cha et al. [CHBG10] further investigated the influence of the three different aspects. The analysis shows that among all users there exists a good correlation of all the combinations of the measures. More interesting results were shown at the correlation coefficients of the top 10% users (regarding the in-degree). The top users show a strong correlation in their retweet and mention influence. So if a user gets mentioned often he also gets retweeted often. Interestingly the in-degree was not related to the other measures at all. So it was shown that the in-degree, representing the user's popularity, is not highly related to other measures. So the in-degree itself does not provide a lot of information about the influence regarding retweets and mentions, of a user. In chapter 4 a similar analysis is done at the smaller datasets of this thesis.

Further case studies done by Cha et al. [CHBG10] focused on the dynamics of influence across topics and time. The results show that most influential users can retain a "*significant influence over a variety of topics*". A temporal analysis shows how different types of influentials interact with their audience. News organizations spawned a huge

[8] http://www.tunkrank.com

amount of retweets among their tweets over a variety of topics, whereas celebrities got a lot of mentions from their audience. The authors also point out that influence needs great personal involvement.

Bongwon Suh et al. [SHPC10] performed a more detailed analysis of factors impacting retweets on Twitter. So they examined a number of features that might affect the *retweetability* of tweets. The base of the analytics of this work was collected by crawling 74 million tweets with help of the Twitter API using the public timeline, which was used for quantitative content analysis associated with retweeting. So this dataset is smaller than the dataset by Cha et al. [CHBG10], but it is based on the same strategy by collecting a set of tweets and analyse them. The dataset is based on a different strategy than the Twitter datasets of this work (see section 3.1 and section 3.2), because they are based on information about a group of seed users. The authors also collected a smaller dataset consisting of 10.000 tweets and traced the retweet count for each tweet. This dataset was used to perform an exploratory data analysis using *Principal Components Analysis* (PCA) and *Generalized Linear Modeling* (GLM) and the purpose of this data analysis was to understand the features, which are correlated with retweeting. For both datasets, the following features were extracted:

- URL (number of URLs in a Tweet)
- Hashtag (number of hashtags in a Tweet)
- Mention (number of mentions in a Tweet)
- Follower (number of users who follow the author of a tweet)
- Followee (number of friends that the author is following)
- Days (number of days the author's Twitter account exists)
- Status (number of tweets made by the author)
- Favorite (number of favorited tweets by the author)
- Retweet (number of retweets of a given tweet)

This list of features correlates with the features analyzed by Cha et al. [CHBG10] and it also correlates with a lot of properties analyzed in this work, which are discussed in chapter 4.

As a first step of the analysis Suh et al. [SHPC10] performed PCA with the smaller dataset with all the nine features, where possibly correlated features are transformed into a smaller number of factors called principal components. The results show that two of the resulting factors seem to distinguish tweets by the profiles of the tweet author and further that retweets are correlated with these factors. The first factor is interpreted by the authors as "*capturing the degree to which tweet authors are broadcasters*". So this factor represents the number of followers, number of followees and number of total tweets of the author. The second factor is interpreted as "*a content factor separating tweets that contain URLs and hashtags from those tweets, which include mentions*".

The second analysis by Suh et al. [SHPC10] focused on discovering associations of the first eight features with retweeting using a generalized linear model similar to the analysis done in this work (see chapter 4). The studies were performed again on the smaller data set and approved the results of the PCA analysis. Hashtags and URLs have "*significant effects on the retweet probability*" and mentions have "*a marginally significant negative association with retweeting*". Further number of followers and followees are strongly predictive of retweet probability. The number of status is marginally negative effective, and the number of favorites is statistically not relevant.

Based on these results Suh et al. [SHPC10] further investigated relevant feature patterns on the large Twitter dataset. This confirmed investigation approved the above described findings. URLs and Hashtags are a significant factor impacting retweetability, whereas the domain of the URL and the type of hashtag matter. Also the number of followers and followees is associated with the retweetability, but the relationship of the number of followees and retweets is not as strong as that with followers. The number of past tweets by an author interestingly does not influence the retweetability. Overall, the analyses show slight different result than the research done by Cha et al. [CHBG10], who point out that the in-degree (number of followers) does not correlate with the retweetability.

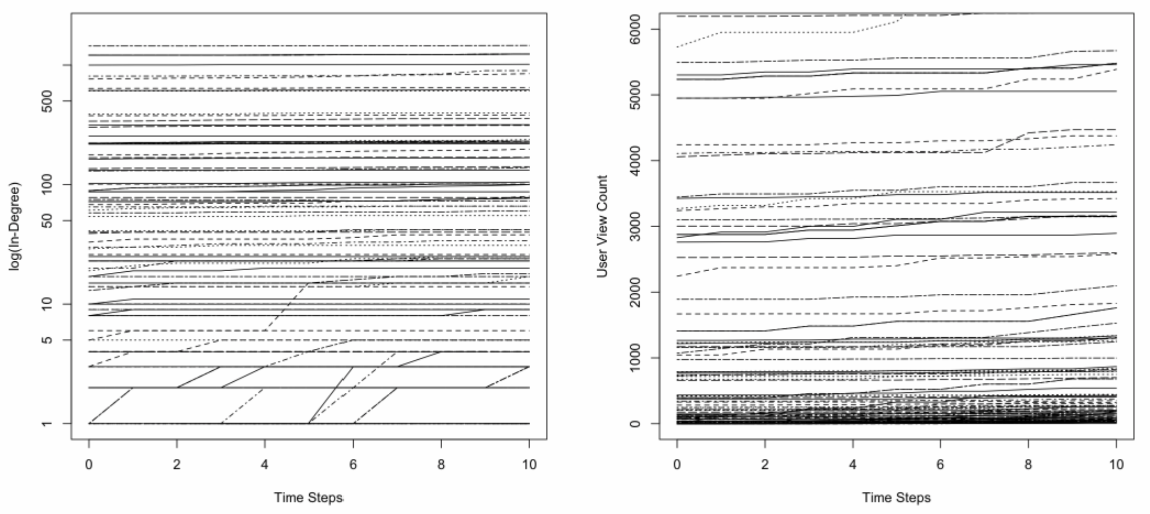
Matthew Rowe [Row11] made similar explorations about the correlation between the subscriber counts and several behavior features on a YouTube[9] dataset. The subscriber counts represent the in-degree of a user and is similar to the in-degree on Twitter, which is stated by the number of followers. So again this analysis combines social and content features and tries to find different influence aspects throughout them. As a second step of this work the audience levels of users were forecasted based on observed behavior. Both the influence and forecasting analysis were done by the help of *Multiple Linear Regression Models* (see section 2.3 for further information), which is the same technique as used in this master thesis.

The dataset used for the analysis in the paper by Matthew Rowe [Row11] consists of information about 200 YouTube videos, randomly chosen out of a 2000 YouTube collection, which was gathered by getting recent YouTube videos. For these videos, information was collected over a 10-day time period. So this approach is very similar to the one presented in this work (see chapter 3 for further information). The data was then divided into a training/testing split using an 80/20 random split, where the former set provides the data for the features and their change over time. Following social and content properties were collected for the dataset:

[9] http://www.youtube.com

- In-degree (number of subscribers of a user)
- Out-degree (number of other users the user is subscribed to)
- User view count (number of unique views of videos)
- Post count (number of unique videos uploaded)
- Post view count (number of times a video has been watched)
- Favorite count (number of times a video has been favorited)

Figure 2.7 shows the change of the in-degree and as well the change of the view count over time of the 80% training split. So it becomes obvious, that the in-degree of the user and the view counts increase with time.



(a) Time vs In-Degree

(b) Time vs View Count

Figure 2.7: Analysis of 80% training split [Row11]

Matthew Rowe [Row11] explored what the correlations between in-degree and other features are. By using a Multiple Regression Model the results show that only the post view count shows a significant correlation with the in-degree of a user. The author claims that this correlation suggests that when an uploaded video by an user gets more views, they also gain more subscribers. This result is nearly identical to the results shown above done by Cha et al. [CHBG10], stating that the in-degree does not have a high impact on other properties, whereas the view count is not possible to measure on Twitter and thereby has not been analyzed in this work. Anyway, the results are somewhat in conflict with the work done by Suh et al. [SHPC10], who point out that the in-degree has a high influence on the retweetability. As already stated, the continued analysis focused on forecasting the audience levels. For that the best features for the in-degree prediction were required. The feature selection identified that an optimal combination of features includes all features without the user view count. Based on that result the forecasting was done by using, on the one hand, all features and on the other hand, just the best features. The analysis shows that the forecasting process with the best features achieved statistically significant performance over the process with all features.

Further investigations about the influence among social networks have been done by Anagnostopoulos et al. [AKM08] focusing on the Flickr[10] social network. They analyzed the tagging behavior of users for a time span of 16 months. The final number of seed users in the dataset was about 800.000. Because a lot of users out of this dataset did not use any tags at all, the authors restricted the dataset to those who used tags, getting a final dataset of 340.000 users. This dataset consists of a giant component with a size of 160.000 users, a second one with size 16 and 165.000 isolated users. The authors used techniques called the shuffle test and the edge-reversal test for their analysis. Anagnostopoulos et al. [AKM08] point out that "*the shuffle test is based on the idea that if the influence does not play a role, even though an agent's probability of activation could depend on her friends, the timing of such activation should be independent of the timing of other agents*". The edge-reversal test is performed by reversing the direction of all edges and run logistic regression on the new graph. The results with the analysis of the two statistical techniques show that there is no significant influence behavior regarding the tagging behavior of Flickr users. This result shows that there is much less social influence on Flickr. This can be possibly be explained by the fact that most users just want to share their personal photos or pictures out of websites, so the social ties between the users are loose.

2.2.3 Semantic and Pragmatic Network Analysis of Social Media

When speaking about emergent semantics and pragmatics of social networks one of the most important and interesting aspects is the usage of tags. In section 2.2.1.3 a brief introduction to tagging and hashtags was already provided. This section will explain tagging in a greater detail.

According to Golder and Huberman [GH05] tagging can be understood as the marking of content with descriptive terms, which are also called keywords. This is an often used way of organizing content, to have the possibility for easier navigation, filtering or search in the future. The keywords of this process are the so called tags.

The authors state that in contrast to a taxonomy, tagging is neither exclusive nor hierarchical and as a result of this fact, tagging can have many advantages over

[10] `http://www.flickr.com`

hierarchical taxonomies. In taxonomies there exist many different possibilities to organize the content and a user has always to decide about the hierarchy. So sometimes a search requires a discovery of distinct folders, whereas in tagging systems a keyword based search just returns the result, which covers the corresponding tags.

The basic tagging style of providing simple keywords as tags has also been adopted in different ways. The most prominent adoptions are the so called hashtags, where an asterisk (#) is placed in front of a tag. These hashtags are used in Twitter. More details about hashtags can be found in section 2.2.1.3.

So with these tagging concepts and the great development of Web 2.0 platforms collaborative tagging becomes very popular. According to Tonkin et al. [TCM+08] it is often also called social tagging, social indexing or social classification. It describes a practice where users can assign keywords (tags) to content resources. These resources are most of the time web based, and the assigned tags are visible to other users of the system. The vocabulary of the tags is in contrast to traditional classification uncontrolled. This fact makes it sometimes hard for researchers to analyze the behavior and find semantics. "*The popularity of tags is determined by their level of use*", and people often speak about the "tag cloud" where the most popular tags are included [MM06].

Hotho et al. [HJSS06a] point out that social resource sharing systems use a knowledge representation, which is called *folksonomy*. The word folksonomy is a combination of the words "taxonomy" and "folk", which should express a conceptual structure created by people. Formally, a folksonomy is a tuple $F := (U, T, R, Y)$ where U, T, and R are finite sets, whose elements are called users, tags and resources. Y is a ternary relation between them (i.e. $Y \subseteq UxTxR$). The elements of Y are called tag assignment.

Wagner and Strohmaier [WS10] also speak about social awareness streams, which are relevant when analyzing social network data. A social awareness stream is an important part of social networks like Facebook, Google+ or Twitter. When a user is logging into such a system, she usually sees a stream of messages of persons she is interested. Social awareness streams include different features provided by the system itself, but basically of messages, including URLs, tags, words and so on. Sometimes the syntax of these messages is also changing by innovation by users (like hashtags). "*This has made social awareness streams complex and dynamic structures*". The authors now introduce a tripartite model of social awareness streams, which they call a *tweetonomy*. It consists of messages, users and the content of messages and is based on the already described existing tripartite structure of folksonomies. The tweetonomy extends the definition of a folksonomy by adding qualifiers to the users, messages and resources to add additional types. The qualifier for the users represents different ways in which users can be related to a message. The qualifier for the messages represents the different types of messages, which are supported by a social awareness stream and the last qualifier for the resources represents different types of resources, which can be included in a social awareness stream. Figure 2.8 shows an example of a simple tweetonomy.

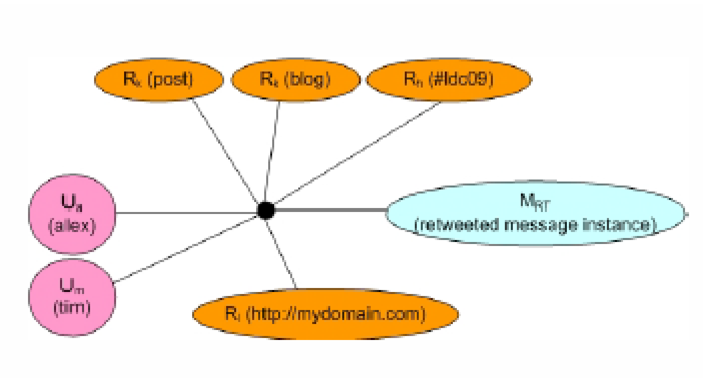

R_k (post)
R_l (blog)
R_n (#ldc09)
U_a (alex)
M_{RT} (retweeted message instance)
U_m (tim)
R_i (http://mydomain.com)

Figure 2.8: Sample tweetonomy [WS10]

The authors furthermore point out that it is also possible to aggregate the social awareness streams depending on the scope and task of investigation. Researchers typically have to decide, which part of a social awareness stream they want to analyze. There are three basic aggregations possible: resource streams (consisting of all messages containing one or more specific resources and all resources and users related with these messages), user streams (consisting of all messages which are related with a defined user set and all resources and further users who are related to these messages) and message streams (consisting of all messages of a given type and their related resources and users). Furthermore, the streams can also be restricted to a given time window.

Wagner and Strohmaier [WS10] performed experiments based on different properties (mainly diversity measures) on different datasets. The authors chose different social awareness streams for the topic "semantic web" which were recorded within the same time period. The four different semantic web datasets are split into a hashtag stream, a keyword stream, a user list stream (the same as used in this thesis described in section 3.2) and a "wefollow" user directory stream. The results show that hashtag streams are in general more robust against external events (like New Years Eve), while, on the other hand, user list streams are more vulnerable to such interruptions. The results also show that in a user stream of experts for a certain topic it seems to be the case that resources, which co-occur with many different hashtags, tend to be important for the main topic of the user group. The authors state that a possible explanation for this is that experts use a fine-granular vocabulary when talking about the topic. The results indicate that hashtag-resource transformations have the possibility to reduce the non-informational noise of social awareness streams, and they can provide meaningful semantic models of the corresponding domain stream aggregation.

According to Tonkin et al. [TCM$^+$08] collaborative tagging has become more and more appealing for researchers. It is possible to produce useful folksonomies with the help of discovered semantic concepts out of a huge set of tags. These folksonomies could replace traditional ontologies. It is anyway relevant to research the community influence on the tagging behavior. So the analysis of user and tag behavior is often an significant starting point for further research.

The uncontrolled vocabulary of tags is, as already mentioned, a severe problem for research. Golder and Huberman [GH05] speak in their highly influential paper about the structure of collaborative tagging about three major problems: polysemy, synonymy and basic-level variation. In the following paragraphs, these three problems will be described with a reference to the explanation by Golder and Huberman [GH05].

Polysemy means that a word can have many related senses. A basic example is the word "window", which can refer to a hole in the wall, a pane of glass that resides within it or also a window in an operating system. Polysemy is especially for the search in tagging systems a serious problem, but it is also difficult to emerge semantics out of concepts.

Synonymy, on the other hand, is referred to multiple words, which have the same meaning. This is an even more difficult problem to approach as a researcher. It also presents a difficulty for tagging systems, because "*the possible inconsistency among the terms used in tagging can make it very difficult for one person to be sure that all the relevant terms have been found.*" An example for this problem is that the words "auto" and "vehicle" have the same meaning.

The *basic level variation* describes a fundamental problem of categorization. It says that several terms, which describe an item, differ in their specificity ranging from very general to very specific. The problem now is that different users observe terms on a different level of specificity, because the users may have a different access to the domain.

To counteract these problems it is often important that there is a general agreement across objects of the users in a system [GH05]. Halpin et al. [HRS07] describe that a system has to become *stable* to solve the problems. Stable means that users have developed a consensus about which tags to use for a certain case. Macgregor and McCulloch [MM06] also state that traditional controlled vocabularies are often not adequate for online resource discovery. They as well describe a fundamental obstacle, which prevents the wider deployment of these controlled vocabularies. The authors exactly explain that "*the obstacle is that the spreading of digital libraries and the Web precedes the ability of any one authority to use traditional methods of metadata creation and indexing*".

Golder and Huberman [GH05] analyzed some aspects of tagging behavior on Del.icio.us[11].

[11] `http://www.delicious.com`

The analysis was performed on two different Del.icio.us datasets. The first one covers all URLs, which appeared in the popular section of Del.icio.us in a certain time period. The dataset contains all bookmarks ever posted to each of these URLs. The second dataset contains a random sample of 229 users and all bookmarks ever posted by these users. The first analysis was performed to identify user activities and tag quantity. The results show that different users use a different number of tags throughout their list of bookmarks. Some users use many tags, whereas other users just use a few. More interestingly, there is not a strong relationship between the number of bookmarks and the number of tags the users use. As a further step, the authors analyzed the kind of tags users use. The following list describes the kind of tags:

- Identifying what (or who) it is about
- Identifying what it is
- Identifying who owns it
- Refining categories
- Identifying qualities or characteristics
- Self reference
- Task organizing

The research also shows that many URLs get most of their bookmarks very quickly, but some URLs also get just a few bookmarks for a long time until they get rediscovered and achieve a high popularity. Empirically it was also shown that usually after the first 100 bookmarks of a URL, each tag's frequency becomes stable. The authors speak about two reasons why this stabilization might occur: imitation and shared knowledge. "*Del.icio.us users may imitate the tag selection of other users.*" Del.icio.us also offers a service which recommends tags, which are used by many other users for the same URL. However, imitation does not explain everything. Shared knowledge among the users of the system may also be the reason for them to use the same tags. As a conclusion Golder and Huberman [GH05] say, that "*the stability they have shown demonstrates that tagged bookmarks may be valuable in aggregate as well as individually, in performing this larger function across the web*".

A good indicator for the stability is also the appearance of a power-law distribution of the tags. Such distribution simple describes that a few tags occur often, whereas the majority of the tags occur drastically less often. Research regarding this topic by Halpin et al. [HRS07] on a Del.icio.u dataset showed that tag distributions of popular sites often follow a power-law distribution exactly (see figure 2.9), whereas the power-law effect is not so distinct at not so often tagged individual sites. So generally the tag distributions of Del.icio.us tend to follow a power-law distribution.

Further indicators that can determine the stabilization of a tag distribution are for example the Kullback-Leibler-Divergence [HRS07], Skew, Kurtosis or the standard deviation [HTE10].

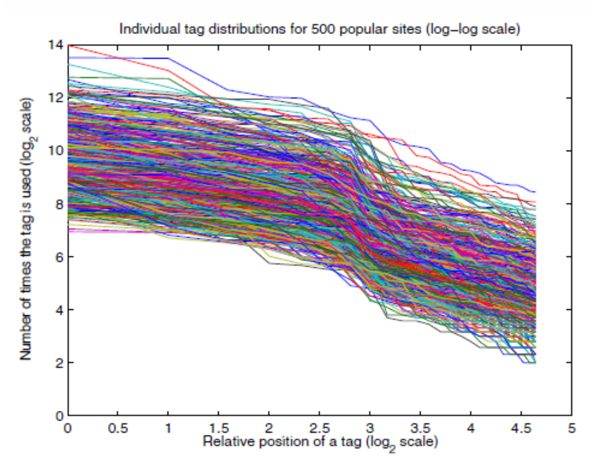


Figure 2.9: Frequency of tag usage, based on relative position. The plot uses double logarithmic (log-log) scale: the horizontal scale gives the logarithm base 2 of the relative position (where the most used tag is in position 1, the second most used tag is in position 2 and so on), while the vertical scale gives the logarithm of the frequency of use. [HRS07]

These above described papers focus on the understanding of the structure and dynamics of social tagging systems. However, also the emergent semantic structures (folksonomies) build important data sources for Semantic Web Applications [MCM$^+$09]. Markines et al. [MCM$^+$09] have tried to answer a key question for the harvesting of semantics of such systems: "*How to extend and adapt traditional notions of similarity to folksonomies, and which measures are suited for applications such as navigation support, semantic search and ontology learning?*" The authors have built an evaluation framework to compare different similarity measures. Because the triple representation of a folksonomy is unsuitable for similarity measures, the dimensionality of the triple space has to be reduced to a two-mode view. The framework focuses on resource-resource and tag-tag similarity, so the aggregation is done across users. The authors use different aggregation methods: projection, distributional, macro-aggregation and collaborative. Furthermore, the following similarity measures are considered: matching, overlap, jaccard, dice and mutual information. For further details about the aggregation methods and similarity measures see [MCM$^+$09].

Markines et al. [MCM$^+$09] used a benchmark dataset of BibSonomy[12] for the evaluation. BibSonomy allows users to directly input relationships between tags. These relation-

[12] http://www.bibsonomy.org/

ships were used to predict user-defined relationships between tags. The results show that mutual information performs better than the other measures with distributional aggregation. For the collaborative aggregation approach, "*it is difficult to determine a clear ranking between the measures*". As a further step, the authors have performed an evaluation based on external grounding. For the tag similarity evaluation, WordNet was used and the results show that in this case macro-aggregation performs the worst with the exception of matching and mutual information. The collaborative aggregation method provided better accuracy for tags. For resource similarity, the authors used the URL collection of the Open Directory Project[13]. The results in this case show that the distributional information does not have a large impact. Mutual information is again the best-performing measure and collaborative aggregation improves the accuracy. Furthermore, Markines et al. [MCM$^+$09] also discuss the scalability problem of the measures and point out that mutual information has quadratic complexity. Macro and collaborative aggregation measures, on the other hand, "*can compensate a loss in accuracy with a huge scalability gain*". Overall, this paper has given an overview of different similarity measures and the decision of dimension design.

A further insight to the chapter of tag similarity measures is provided by Cattuto et al. [CBHS08] who analyze three different measures of tag relatedness: tag co-occurrence, cosine similarity of co-occurrence distributions and FolkRank. The authors point out, that in most studies, the choice of similarity measures in done in an ad-hoc fashion, so they want to provide a deeper insight into these measures to provide a better understanding of the different measures. The tag co-occurrence can be measured on a tag-tag co-occurrence graph, which is a weighted and undirected graph. The cosine similarity is measured in tag-tag co-occurrence distributions. The third measure of this work is the FolkRank, which is based on the idea of the PageRank algorithm (see [BP98]). The FolkRank states that "*a resource which is tagged with important tags by important users becomes important itself*". See [HJSS06a] for a more detailed explanation of the FolkRank algorithm. The experiments of this paper are conducted on a Del.icio.us dataset. A first insight states that the cosine similarity provides more synonyms than other measures in most of the cases. Cattuto et al. also observe that the tags "java" and "python" could be considered as siblings. They point out that a possible explanation for this could be that the cosine similarity is measuring in the global context, whereas the co-occurrence measurement and as well the FolkRank measure in the same post. The cosine similarity also provides tags that belong to a broader class of tags, which are not strongly correlated with rank. Furthermore, the authors also provide a formal validation based on semantic grounding (as in [MCM$^+$09]). They have used WordNet and to measure the similarity the taxonomic shortest-path length and the Jiang and Conrath distance measure is used (see [JC97] for further details). The evaluation shows again that tags obtained by cosine similarity seem to be synonyms or siblings of the original tag and the two other measures provide tags that are more

[13] `http://www.dmoz.org/`

general. So the authors conclude their work with the following application areas of the different measures:

- For synonym discovery the cosine similarity is the best measure.
- For building a concept hierarchy FolkRank and co-occurrence relatedness are the measures to favor.
- For discovery of multi-word lexemes FolkRank is the best measure.

Benz et al. [BGH$^+$08] have also tried to better understand the semantics of tags and the tagging process. For their analysis, a Del.icio.us and Flickr dataset was used. Del.icio.us can be understood as a broad folksonomy (each user can tag any resource), whereas Flickr is a narrow folksonomy (only the user who owns the resource is allowed to tag it, or she can allow a list of friends to tag it). Because nowadays a lot of users are registered on several social tagging systems, it is interesting to understand the similarities and differences in the tagging behavior across the different folksonomies. This is called cross-folksonomy analysis and has become an interesting part in recent work. Benz et al. [BGH$^+$08] took a subset of both Del.iciou.us and Flickr where the users' profiles were correlated. Figure 2.10 shows a histogram illustrating distribution of tags that appear in both Delicious and Flickr (intersection), and those that appear only in Delicious (disjoint).

A first observation by the authors shows that a less frequent used tag in Del.icio.us is less likely to appear in Flickr. To establish semantically similar tags, tag context similarity is used. For an example, the tag "apple" is used. It is shown that the users with the highest similarity to "apple" use the tag in both Del.iciou.us and Flickr to refer to the same concept. On the other hand, when choosing the user with the lowest similarity, it is shown that this user uses "apple" in Del.icio.us to refer to the computer company and in Flickr to refer to the fruit. Further investigations use the cosine similarity, which is a useful measure for determining tag similarity (see above). This time the analysis was performed on the 10.000 most popular tags in Flickr. A first observation by the authors is that the tag context similarity also provides good results for a narrow folksonomy like Flickr. The tag context similarity measure is also useful for disambiguating terms. Furthermore, Benz et al. [BGH$^+$08] embedded the representation of the tag-tag space with the cosine similarity measure in a three-dimensional space using the software OntoGen[14]. Finally a further feature is added to the analysis of the tag-tag co-occurrence graph. When users annotate their resources, they place the tags in a specific order. The authors now tried to analyze if the tag ordering has an impact on the semantic. Three different similarity measures were used to analyze this ordering and the results generally show that the tag order has a relevant semantic value, but further work is required to determine the different characteristics

[14] http://ontogen.ijs.si/

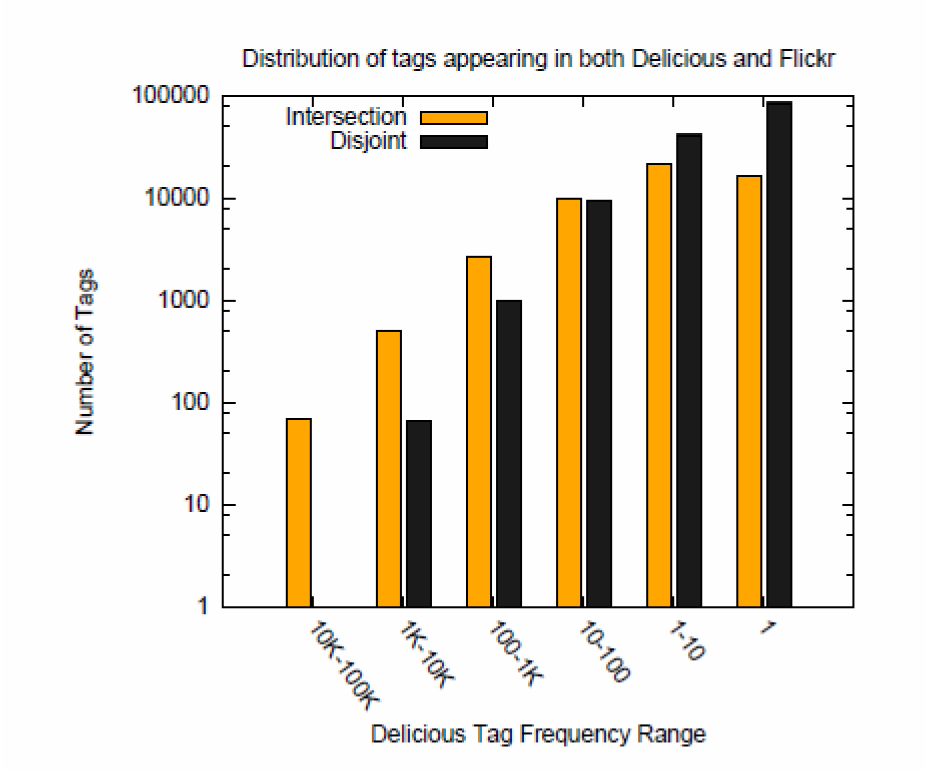


Figure 2.10: Distribution of tags that appear in both Delicious and Flickr (intersection), and those that appear only in Delicious (disjoint) [BGH+08]

in a greater detail. Overall, this work can be concluded by stating that tags in Flickr represent more visual meaning (describing the photos in a visual way), whereas the tags in Del.icio.us have a bias towards a technical meaning.

In a further paper by Benz et al. [BHS10] the authors analyzed the potential of self-emerging ontologies from folksonomies. This is based on the already mentioned evidence for underlying semantics in such evolving structures. An algorithm (based on [BH07]) is extended so that it creates a hierarchical organization scheme, and that it can capture the semantics and the diversity of the shared knowledge. The authors use the word *diversity* to state that the self-emergent ontology integrates different views on the data. To do these analyses a Del.icio.us dataset is used. As a first step the algorithm should identify synonym tags and by doing so the vocabulary of the folksonomy should "shrink" by merging all similar tags. The resulting structure is called *synsetized folksonomy.* As a second step ambiguous tags should be discovered. Ambiguous means that different users use different tags to refer to the same semantic concept. Ambiguity and synonymy are major problems of collaborative tagging and have been also mentioned at the beginning of this section. To discover the ambiguity the synsetized folksonomy is taken as the base structure and then the goal is to iterate through all synsets and to check whether the current synset provides different meanings. The ontology learning algorithm then can be operated on the resulting structure. To evaluate the results the authors chose two gold-standard bases: WordNet and Wikipedia. The evaluation shows that the extended algorithm provides ontologies, which resemble more closely to both gold-standard ontologies. So identifying synsets and resolving ambiguity is a great way to better reproduce the diversity of shared knowledge. Figure 2.11 shows a summary of the results obtained by the best parameter settings.

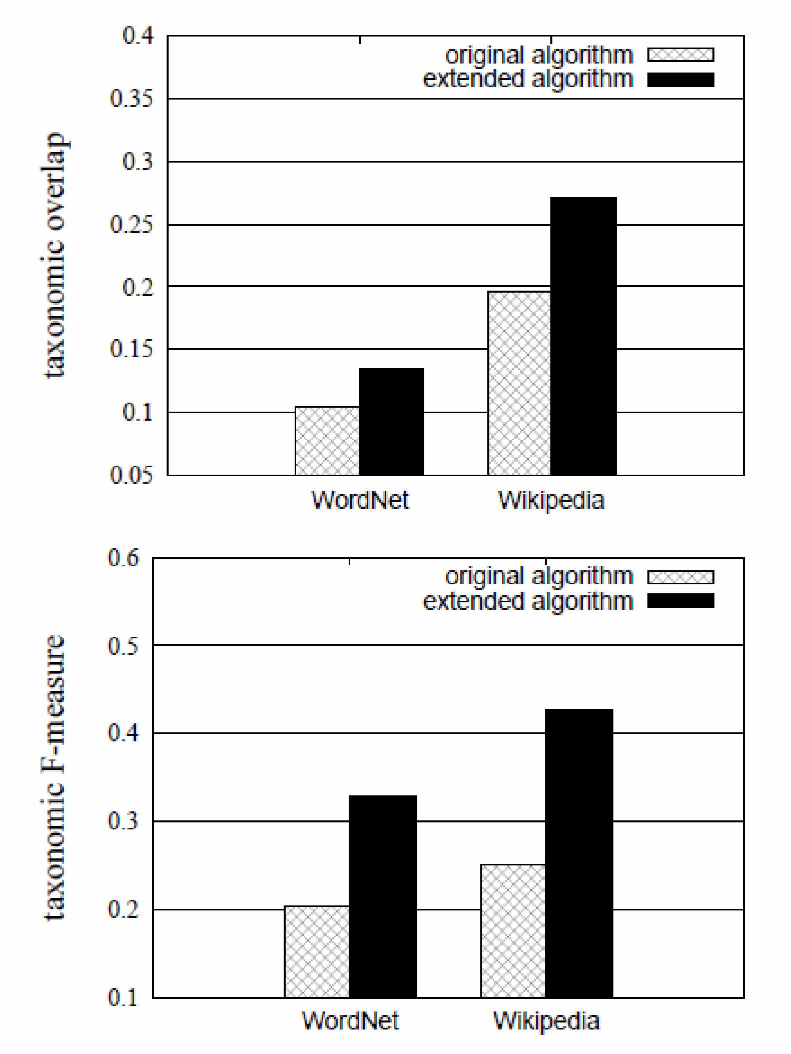


Figure 2.11: Experimental results of comparing the learned ontology with the reference ontologies from WordNet and Wikipedia [BHS10]

Regarding the emergent semantics of social tagging systems, Hotho et al. [HJSS06b] discovered topic-specific trends within folksonomies. For this target, the authors took snapshots of the corresponding folksonomy at different time points. As a first step a ranking is needed and then a ranking method is used to focus on a specific topic. For the analysis, the already described ranking algorithm FolkRank was used. A topic-directed FolkRank computation was used. Furthermore, this measure of change in popularity should give insight into the trends in a certain community in the folksonomy. The folksonomy to analyse was again Del.icio.us. The interesting aspect of this experiment is that it can be done regardless of the underlying type of resources of the collaborative tagging system, which makes it also interesting for multimedia applications. The results indicate that the early users of Del.icio.us were more critical and idealistic, because they used more tags like "activism" or "war". With increasing time the popularity of such critical tags faded and the tags tended to form a more uniform distribution. So over time the tags converted towards more mainstream topics. Furthermore, some more topic-related analyses were conducted. The authors also compared their results with observations done by Dubinko et al. [DKM$^+$06] who used a metric of interestingness to determine trends. Hotho et al. [HJSS06b] conclude their comparison by stating that the measure of interestingness seems to be more useful for short-term observations on particular folksonomy elements, but it is more sensitive to momentary changes in the folksonomy than the FolkRank. Overall, the authors conclude their work by stating that topic-specific trends can be discovered in folksonomy-based collaborative tagging systems.

Further research regarding tag similarity and tag relatedness will be provided in the following parts of this section to identify the pragmatics of tagging.

So far, we know more about the structure and dynamics of folksonomies and also have reviewed aspects of emergent semantics of social tagging systems, a next step is to know more about the pragmatics of social tagging systems. Strohmaier et al. [SKK10] tried to find the underlying user motivations for tagging and furthermore how they influence resulting folksonomies and tags. To analyze the motivation it is useful to distinguish between two types of user motivation for tagging. On the one hand, there are users, who are motivated by categorization. These users view tagging as a way "*to categorize resources according to some high-level characteristics*". They want to organize their content to have better navigation possibilities in the future. On the other hand, there are users, who are motivated by description. Such users want to "*accurately and precisely describe the resources they use*". The main target of these users is to provide better possibilities for later searching. Figure 2.12 shows the distinction of categorizers and describers by illustrating tag clouds of users. Anyway, tagging in the real world is very likely motivated by a combination of categorization and description. The authors now studied three different questions: 1) How can we measure the motivation behind tagging? 2) How does users' motivation for tagging varies across and within different tagging systems? and 3) How does tagging motivation influence resulting folksonomies?

To answer the three different questions Strohmaier et al. [SKK10] gathered many

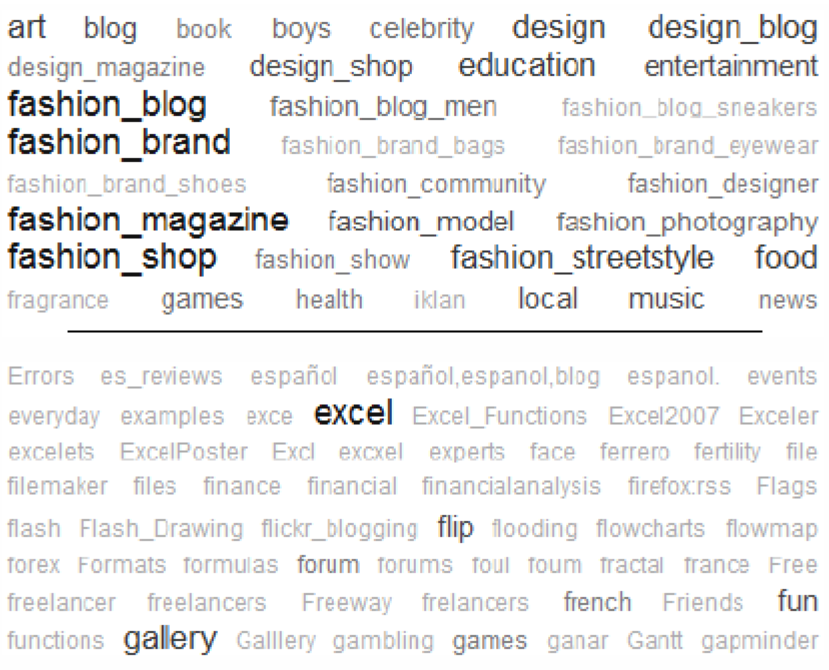

art blog book boys celebrity design design_blog
design_magazine design_shop education entertainment
fashion_blog fashion_blog_men fashion_blog_sneakers
fashion_brand fashion_brand_bags fashion_brand_eyewear
fashion_brand_shoes fashion_community fashion_designer
fashion_magazine fashion_model fashion_photography
fashion_shop fashion_show fashion_streetstyle food
fragrance games health iklan local music news
Errors es_reviews español español,espanol,blog espanol. events
everyday examples exce excel Excel_Functions Excel2007 Exceler
excelets ExcelPoster Excl excxel experts face ferrero fertility file
filemaker files finance financial financialanalysis firefox:rss Flags
flash Flash_Drawing flickr_blogging flip flooding flowcharts flowmap
forex Formats formulas forum forums foul foum fractal france Free
freelancer freelancers Freeway frelancers french Friends fun
functions gallery Galllery gambling games ganar Gantt gapminder

Figure 2.12: Sample tag clouds for categorizers (top) and describers (bottom) [SKK10]

different datasets split into two main categories: "Synthetic Personomy Datasets" and "Real-World Personomy Datasets". *Personomy* can be understood as all the information available for a user inside a tagging application. The first datasets try to simulate the behavior of users, who are mainly motivated by description and the second ones include datasets of popular tagging systems. To determine the tagging motivation two measures were used by the authors. To detect categorizers it is possible to see the activity of tagging as an encoding process. So categorizers would try to "*maintain high information value in their tag vectors*". This can be captured by using the conditional entropy. The second measure to detect describers is based on the fact that such users would generate tags that closely resemble the content of the resources. So describers aim to use a high number of different tags, which can be captured by the orphan ratio. The applied analyses show that "*tagging motivation of individuals differs within and across tagging systems and that users' motivation for tagging has an influence on resulting tags and folksonomies*". It is also shown that the agreement on tags among categorizers is significantly lower compared to tags among describers, which is also an interesting point for the problems described above in this section. This indicates that a general agreement among the users of a tagging system for the choice of their tags is an important way to counteract the corresponding problems of collaborative tagging systems. The authors, furthermore, describe that users, who are motivated by categorization generally produce fewer descriptive tags. This would indicate that not all tags are equally useful for different tasks. A possible example could be information retrieval. Moreover, this would indicate that it is nearly impossible to valuate the usefulness of tags without knowing the motivation of a user. The results also show interesting aspects for tag recommendation, showing that categorizers could benefit from such recommenders that suggest tags based on their individual tag vocabulary, and that describers could benefit more from tags that capture the content of the resource the best.

Next to these analyses based on the corresponding two measures, there are also many other measures available for detecting tagging motivation. Körner et al. [KKGS10] studied different quantitative measures for tagging motivation. Again, the authors tried to find answers to the question, why users tag, and they tried to understand the motivation behind tagging in a better way. Körner et al. [KKGS10] extend the two already covered measures (condition entropy and orphan ratio) with measures that try to cover quantitative measures like the number of annotated resources normalized by the number of complete resources of the user. The three new measures are: 1) Tag/Resource Ratio (total number of annotated resources divided by the number of total resources of a user) 2) Overlap Factor (relating the number of all resources to the total number of tag assignments of a user) and 3) Tag/Title Intersection Ratio (intersection of tags from the words of a resource's title). All the five measures focus on tagging behavior of users and not on the semantics of tags.

The experiment was built upon a Del.icio.us dataset. Based on the experiment the evaluation was split into a qualitative and a quantitative evaluation. The qualitative evaluation was used to guarantee the usefulness of the different measures. For this evaluation human persons were given the task to look at a subset of posts out of users' personomies and classify whether a given personomy better refers to a personomy of a categorizer or a describer. The results show that all measures are useful to determine the tagging motivation, but not all are equally useful. Especially, the Tag/Resource Ratio seems to be the best measure to cover the human evaluation. Figure 2.13 shows the accuracy for the different measures resulting from the user study and compares them to a random baseline (see [KKGS10] for calculation equation).

The quantitative evaluation by Körner et al. [KKGS10] should try to answer the questions, whether the distinction between categorizers and describers has an impact during tagging and how this distinction can be captured in the best way by one of the measures. This evaluation uses tag recommenders to analyze the influence on the tag decisions by users. Tag recommenders can be distinguished by two types. A folksonomy-based recommender looks at all tags available at the folksonomy for a single resource and ranks the corresponding tags based on their frequency. On the other hand, a personomy-based recommender looks at the tags used by the user in his personomy. For the evaluation, the folksonomy-based recommender was used for describers and the personomy-based recommender for categorizers. The results show that tags used by describers tend to be more similar to the tags by other describers, whereas categorizers

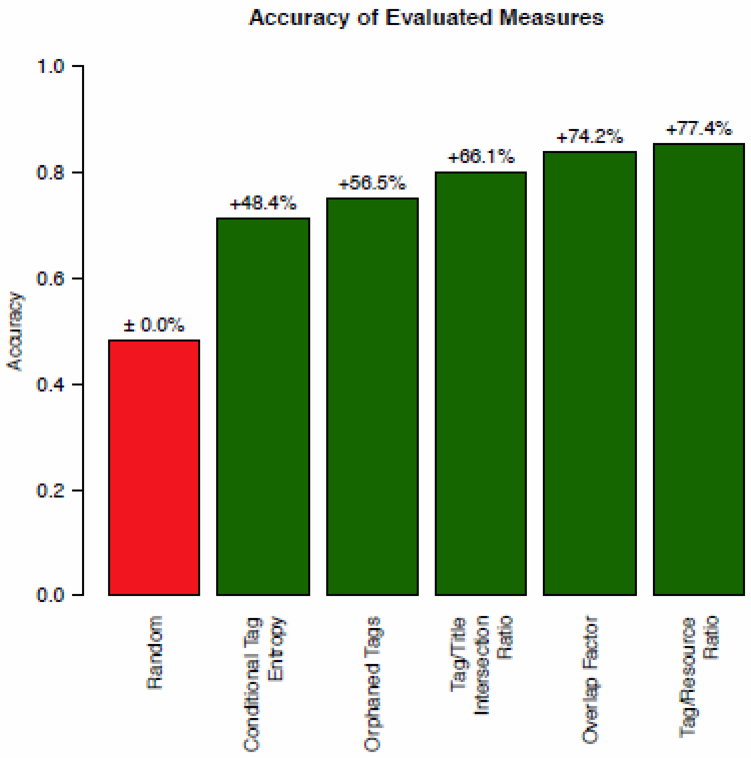


Figure 2.13: Accuracy for the different measures resulting from the user study [KKGS10]

seem to use an exclusive tagging vocabulary. These results are similar to the results shown by Strohmaier et al. [SKK10]. It is also shown that the Tag/Title Intersection Ratio and the Tag/Resource Ratio best predict user behavior. Overall, the quantitative evaluation shows that the tagging behavior of users is significantly based on the motivation behind tagging.

Körner et al. [KBH$^+$10] have done similar work, and they tried to find factors that influence the evolution of semantic structures in collaborative tagging systems. The authors again distinguish between categorizers and describers throughout their work. As a first step the paper addresses the problem of detecting tag relatedness. To solve this, five measures are used: co-occurrence count, three context measures (cosine similarity), and FolkRank (see [HJSS06a]). The above described papers by Strohmaier et al. [SKK10] and Körner et al. [KKGS10] cover several further measures for tag relatedness. The authors now cover the topic of pragmatics of tagging by looking at the motivation behind tagging. To measure tagging pragmatics several measures are used (some are already known from Strohmaier et al. [SKK10] and Körner et al. [KKGS10]): vocabulary size, Tag/Resource Ratio, Average tags per post, Orphan ratio.

The authors used a large Del.icio.us dataset and followed an incremental approach to determine the capability of each measure. As a first step all users are sorted in ascending order according to the measures so that e.g. the first user in the orphan ratio list is supposed to be the most extreme categorizer, whereas the last one should be the most extreme describer. Figure 2.14 shows the distributions of membership scores for each list (values close to 0 indicate strong categorizers and values close to 1 point to describers). A first obvious observation is that the distribution of the orphan list differs clearly from the other distributions. Nevertheless, the results show that the orphan ratio performs often worse than the other measures.

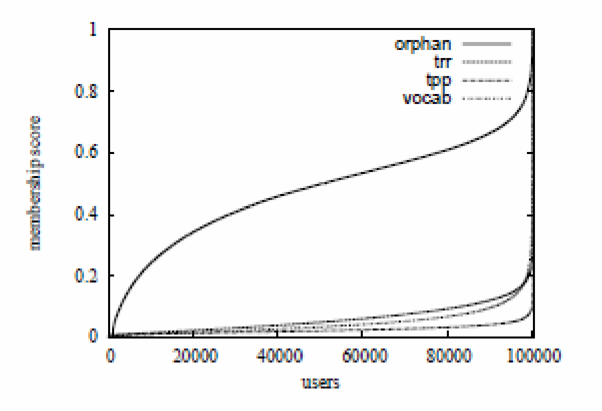


Figure 2.14: Distributions of the membership scores for each measure of tagging motivation [KBH$^+$10]

The main strategy of the work by Körner et al. [KBH$^+$10] was to gather subsets of users out of the dataset and analyze the suitability of each of the previously stated pragmatic measures. In summary, 64 partitions for each of the four measures were created. The results show that generally the more people are contributing to a tagging system, the better is the quality of semantic tag relations, which can be extracted from the produced folksonomy structure. Furthermore, the results suggest that subsets based on describers provide more precise inherent semantic structures than those subsets, which are based on categorizers. The authors claim that "*the effectiveness of current semantic measures for tag relatedness are influenced by factors originating outside of the semantic realm*". This suggests that it is important to know about the pragmatics of tagging, when someone wants to harvest semantics of tagging systems. Furthermore, the work indicates that selection strategies based on describers create smaller folksonomies, and these folksonomies produce meaningful semantics. Overall, the best result was established by taking a subset where 10% of the extreme describers (based on tags-per-post measure) are eliminated. It should also always be kept in mind that spammers can falsify the expected results.

Kern et al. [KKS10] also showed the above stated results on the motivation behind tagging. The authors sum the results up by stating that tagging motivation shows a significant variety, and that fact could play an important role for problems like tag recommendation and information retrieval.

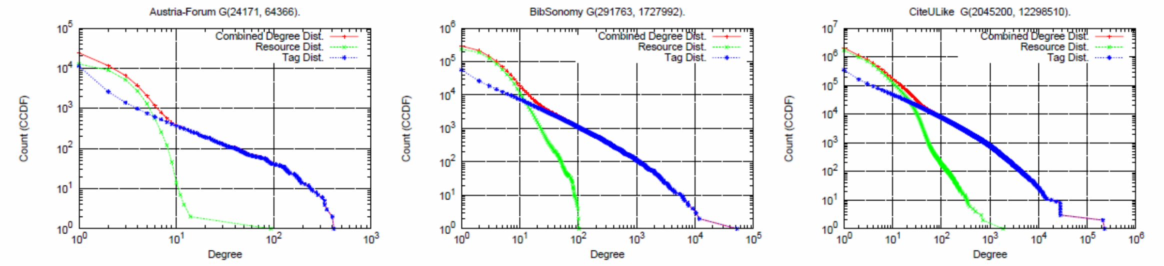
When talking about tagging, another interesting aspect is the understanding of the different levels of tag relatedness. Benz et al. [BKH$^+$11] have addressed this topic. It is generally important to know more about this topic, to identify hierarchical relationships between concepts. The paper presents a systematic analysis of several folksonomy-based notions of term abstractness. A measure for term generality aims to "*allow a differentiation of lexical entities $l_1, l_2, ...$ by their degree of abstractness*". The authors introduce several measures to determine the tag generality. A first idea is that more abstract tags occur more often, because there might exist more resources, which are relevant for this tag. So the first measure is frequency-based and simply counts the number of tag assignments. A second idea is that "*more general tags show a more even distribution*". This could be the case, because they might be used at a constant level and are used to annotate a broad spectrum of resources. So this can be measured with entropy, where more abstract terms will have a higher entropy (this entropy idea is also used in this thesis in section 4.4). Another idea is to use centrality measures, because they determine the importance of a vertex in a graph. The intuition behind this is, that more abstract terms should be more important, and so they should be more central. A final measure is based on the idea of statistical subsumption, saying that a tag subsumes another tag and to measure the generality by using the number of subsumed tags. Schmitz et al. [Sch06] state that a tag X subsumes a tag Y if $P(X|Y) >= t$ and $P(Y|X) < t$ based on a useful co-occurrence threshold t.

To evaluate the four different measures, Benz et al. [BKH$^+$11] compared them against a ground truth (established datasets). They chose several core ontologies and taxonomies for their evaluation. For determining the performance of the measures a Del.icio.us dataset was used. As a first step the tag-tag co-occurrence graph was created and all tags with a degree of less than two were removed. Furthermore, a tag-tag similarity graph was derived by using the Resource-Context-Similarity (see [CBHS08]). Furthermore, the authors derived two measures from a taxonomy. These measures allow a comparison of the abstractness level between terms, which occur in disconnected parts of the taxonomy graph. The two measures are 1) the shortest path to the taxonomy root and 2) the number of subordinate terms (done with an experiment with human persons). An observation of the work is that measures based on frequency, entropy or centrality in the tag co-occurrence graph show a good agreement on information provided by standard taxonomies. Another observation of the experiments is that measures, which are based on tag similarity graphs, provide the worst results. Overall, the tag-tag co-occurrence graphs furnish the best "taxonomic" information, but also the probabilistic model of subsumption performs well. The work as well shows that popularity seems to be a good way to determine the generality of a given tag.

Pragmatic aspects can also be an interesting aspect for the evaluation of folksonomies. Helic et al. [HST^{+}11] have done work regarding the question to what extend folksonomies are pragmatically useful for navigating social tagging systems. The main idea of their work is to use hierarchical structures learned by folksonomy algorithms as background for decentralized search. Decentralized search means that a user only has local knowledge about the network structure. So this can be understood as a way that users at any given page only know about the links from that page and do not know about links from other pages into the system. So overall decentralized search can be seen as a very natural model of navigating tagging systems. An algorithm in decentralized search on a network starts its search at a special node and then tries to reach another determined destination node, whereas this search is constantly based just on local knowledge. The performance of such decentralized search algorithms are based on the quality of the hierarchical background knowledge. The authors focus on the navigation of tag-tag graphs in their work. Their pragmatic folksonomy evaluation framework is based on the following steps: 1) folksonomy induction, 2) classification of searchable networks, 3)modeling navigation, 4) defining evaluation metrics (length of the shortest path), 5) simulation and finally 6) evaluation.

The work by Helic et al. [HST^{+}11] uses four different folksonomy induction algorithms on five different social tagging datasets for validation. A theoretical evaluation of the folksonomies shows that the existing algorithms for folksonomy induction produced folksonomies are theoretically useful for decentralized search. Anyway, it was shown that not all folksonomies provide the same results. Those which are based on tag similarity graph algorithms are more useful than those which are produced by hierarchical clustering algorithms. The pragmatic analysis shows that the used existing algorithms produce folksonomies, which are better for exploratory navigation than a random baseline folksonomy. Using tag similarity graphs to produce folksonomies is a great way to support exploratory navigation. Hierarchical clustering again shows weaker results. So the pragmatic evaluation agrees with the theoretical evaluation. The authors claim that hierarchical clustering seems to lack additional information about the dataset, which can be provided by using tag similarity graphs or centrality ranking. Overall, it can be said that folksonomies can provide a useful background knowledge for exploratory navigation. Furthermore, the authors state that future folksonomy research needs to look more and more on the pragmatic aspects and evaluation in addition to semantic evaluation in order to "*examine the usefulness of folksonomies for different tasks*".

Helic et al. [HTSA10] made further analyses regarding the navigability of social tagging systems. They tried to answer the question, whether tag clouds are useful for navigation. The authors point out that most of the social tagging systems designers think that tag clouds are a useful tool for navigation. The authors did different experiments to provide answers to the question. The paper focuses on tag-resource bipartite graphs, because such graphs represent a natural way users are adopting tag clouds for navigation. A first network-theoretic approach shows that there exist short paths between nodes in a social network, and that people are able to navigate through the network having only local knowledge of the network, which was already shown in [HST$^+$11] (see above). For further experiments, the authors chose three different datasets (Austria-forum[15], BibSonomy[16] and CiteULike[17]). First, the authors investigated the usefulness of tag clouds for navigability without taking interface restrictions into account. This step shows that the phase of adoption of a social tagging system is highly relevant for the usefulness of tag clouds for navigation. Furthermore, it is pointed out that social tagging networks follow a power-law distribution (see beginning of this section) and such networks are navigable. Figure 2.15 shows the tag, resource and degree distributions for each of the three datasets.



(a) Austria-Forum (b) BibSonomy (c) CiteULike

Figure 2.15: Tag, resource and degree distributions for the three datasets [HTSA10]

As a second step Helic et al. [HTSA10] added a first interface restriction to the experiments: limiting the tag cloud size. It is shown that this restriction does not influence the network to a large extend and therefore, limiting the size to useful sizes does not influence the navigability. As a third step the experiments take pagination into account. Limiting the out-degree of hub nodes in a power-law network lets the giant component (containing the majority of the nodes of a network) collapse and therefore, also destroys the navigability regarding tag clouds. These results implicate that in theory tag-resource networks provide efficient navigability, but popular interface decisions like pagination (combined with reverse-chronological listing of resources) can hurt the navigability. Based on the results of the experiments Helic et al. [HTSA10] illustrated a way to generate more efficiently tag clouds for navigation in collaborative tagging networks, which are not so vulnerable to the pagination effect. Anyway, the authors also suggest that engineers, who want to construct tag cloud algorithms for navigation purposes have to take semantic and navigation penalties into account and try to find a balance between them.

[15] http://www.austria-lexikon.at/

[16] http://www.bibsonomy.org/

[17] http://www.citeulike.org/

2.3 Time series analysis

This section gives an overview about time series. At the beginning, a description about time series data is presented (section 2.3.1). The following sections cover one way of analyzing time series data. First regression models are introduced in section 2.3.2. Next selected types of regression models are discussed in detail by explaining multiple linear regression models (section 2.3.3), autoregressive models (section 2.3.4) and at the end multilevel regression models (section 2.3.5). The aim of this section is to provide an explanation of the statistical tools used in this work. Section 4.1 explains the exact model used for the experiments of this thesis.

2.3.1 Time series data

Genshiro Kitagawa defined time series data in his book "Introduction to Time Series Modeling" [Kit10] as follows:

> *A record of phenomenon irregularly varying with time is called time series.*

Classic time series examples are meteorological data like temperature or rainfall; economic data like stock prices and also medical data [Kit10]. Figure 2.16 shows a sample time series of the yearly average global temperature deviations. It is clearly visible that there is a trend of rising global temperature deviations, so the time series shows an upward trend [SS06].

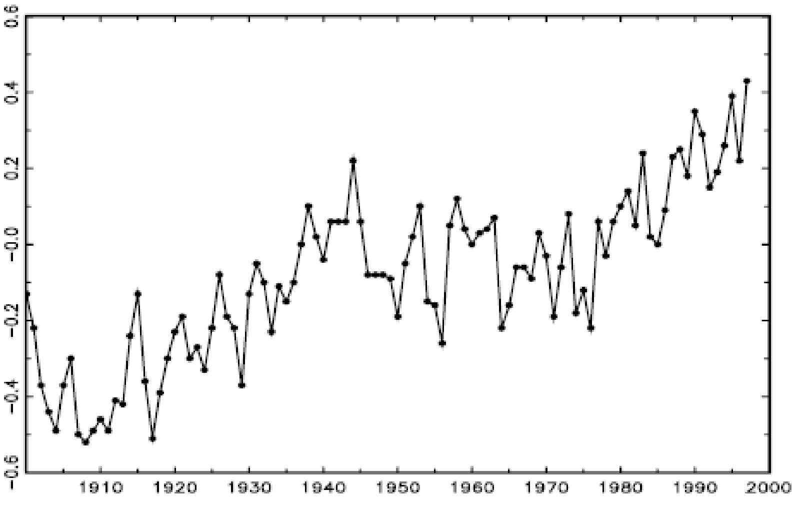


Figure 2.16: Yearly average global temperature deviations in degrees centigrade. [SS06]

Shumway and Stoffer [SS06] claim that the analysis of experimental data, which is based on observations at different time points, "*leads to new and unique problems in statistical modeling and inference*". The analysis of such time series is furthermore important for many different scientific areas. Kitagawa [Kit10] points out that it is also necessary to carefully examine graphs of the data as a first step of the time series analysis. So it is easier to identify the next step of the analysis and find appropriate strategies for statistical modeling.

Kitagawa [Kit10] describes different classifications of time series. Following categories are the most relevant for this thesis:

Continuous and discrete time series

Data which is recorded continuously is called *continuous time series*, whereas data, which is observed at certain intervals of time, is called *discrete time series.*

Univariate and multivariate time series

If there is only a single observation at each time point, the time series is *univariate.* On the other hand, the data is called *multivariate* time series, if the data is obtained by simultaneously recording two or more phenomena.

Stationary and nonstationary time series

If the recorded phenomenons are varying irregularly over time, the time series is called *stationary.* If the stochastic structure of the time series itself changes over time the time series is called *nonstationary.*

Linear and nonlinear time series

If the time series is an output of a linear model it is called *linear time series* and in contrast if it is the output of a nonlinear model it is called *nonlinear time series.*

The primary objective of time series is to find appropriate mathematical models that describe the sample data plausible [SS06]. As a further step prediction can be proceeded, where the future behavior of time series can be estimated based on the correlations over time and among the variables [Kit10].

2.3.2 Regression Models

According to Rawlings et al. [RPD98] modeling is generally used to develop mathematical expressions, which describe the behavior of a random variable of interest. This variable can be the number of deaths from murder, the price of one liter milk in the world market or the average global temperature. This variable is always the *dependent variable* (Y). A subscript on Y identifies the unit in a greater detail. For example, the country from which the numbers of deaths from murder are recorded.

The modeling aims to describe how the true "*mean of the dependent variable $\varepsilon(Y)$ changes with changing conditions*". Furthermore, it is possible to add other variables, which can provide further information about the behavior of the dependent variable. These variables are called *independent variables* and are denoted by X. They can be described as predictors or explanatory variables. All models also include unknown constants, which are called *parameters*. Such variables are denoted by Greek letters, control the behavior of the model and must be estimated from the data.

As stated in section 2.3.1, models with linear parameters are called *linear models*. Rawlings et al. [RPD98] state that when the preliminary study of a process or prediction is the primary objective, the models usually fall into this category. The parameters are simple coefficients or functions of the independent variables. On the other hand, more realistic models are nonlinear models, where the model is nonlinear in the parameters. Nonlinear models often can be transformed into linear models, and occasionally they cannot be transformed. This work focuses on linear models.

The simple linear model involves a single independent variable, and this model states that "*the true mean of the dependent variable changes at a constant rate as the value of the independent variable increases or decreases*" [MJK08] [RPD98]. The functional relationship between the true mean $\varepsilon(Y_i)$ of Y_i and X_i is the equation of a straight line and written as [RPD98]:

$$\varepsilon(Y_i) = \beta_0 + \beta_1 X_i \tag{2.8}$$

In this equation β_0 is the intercept (the value of $\varepsilon(Y_i)$ when X = 0) and β_1 is the slope of the line [RPD98].

Rawlings et al. [RPD98] state that "*the observations on the dependent variable Y_i are assumed to be random observations from populations of random variables with the mean of each population given by $\varepsilon(Y_i)$*". To take the deviation of an observation Y_i from its mean $\varepsilon(Y_i)$ into account a random error ϵ_i is added, which is shown in equation 2.9.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2.9}$$

Rawlings et al. [RPD98] point out that the subscript i describes the individual observed unit, where i ranges from 1 to n. X_i denotes the n observations on the independent variable and they are supposed to be measured without error. Y_i and X_i are paired observations, and it is always assumed that these variables are measured on every observation unit. According to Montgomery et al. [MJK08] the error term ϵ_i is used, as already mentioned, to take "*the deviations of the actual data from the straight line*" into account. ϵ_i is seen as a statistical error, and so it is defined as a random variable. Typically, it is assumed that ϵ_i is normally distributed with mean zero and variance

σ^2. This is stated as $N(0, \sigma^2)$. The variance is assumed constant, and so it does not depend on the value of the independent variable.

2.3.3 Multiple Linear Regression Models

When using regression models it is often necessary to use more than one independent variable to determine the behavior of the dependent variable. So "*the linear additive model can be extended to include any possible number of independent variables*" [RPD98]:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_m X_{im} + \epsilon_i \tag{2.10}$$

Rawlings et al. [RPD98] describe that the subscript notation seen in equation 2.9 is extended in this equation to identify each independent variable and its regression coefficient. This extension is added to each X and β. There is a total of m different independent variables; and that means - including the parameter β_0 - we have to estimate a total of $m + 1$ parameters. The distinct β parameters are often called *partial regression coefficients*, because they provide information of the effect of one independent variable on the dependent variable Y_i with the assumption that all the other independent variables do not change [MJK08].

According to Rawlings et al. [RPD98] it is also possible to state the multiple linear model in matrix notation. Equation 2.10 can be expressed by the following four matrices:

Y This is a $n \times 1$ column vector consisting of the different observations of the dependent variable.

X This is a $n \times (m + 1)$ matrix consisting of a column of ones (1) and this column is followed by the m column vectors, describing the observations of the different independent variables.

β This is a $(m + 1) \times 1$ vector consisting of the parameters.

ϵ This is a $n \times 1$ vector for the errors.

So the multiple linear model can now be written in the following way [RPD98]:

$$Y = X\beta + \epsilon \tag{2.11}$$

In matrix form the equation can be written as [RPD98]:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{13} & ... & X_{1m} \\ 1 & X_{21} & X_{22} & X_{23} & ... & X_{2m} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & X_{n1} & X_{n2} & X_{n3} & ... & X_{nm} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_m \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix} \tag{2.12}$$

2.3.4 Autoregressive Models

John M. Gottman [Got09] says that "'*if we think about the problem of trying to predict the future from our knowledge of the past, it will become clear that to solve this problem we must assume that something does not change very much.*" The author furthermore points out that this is often an appropriate way to model time series. An *autoregressive model* is a model where the model goes back p time units in the regression to have the ability to predict. This model is noted as $AR(p)$ and the parameter p determines the order of the model. The attempt of this model is to estimate an observation as a weighted sum of previous observations, which is the number of the parameter p.

So in a time series case it should be allowed that the dependent variable is influenced by past values of the independent variables and as a further refinement by its own past values [SS06]. Section 2.3.3 already described multiple regression models. Gottman [Got09] states that in a time series, it is for example possible to pick a point x_3 and to use the two previous values in time (x_2 and x_1) as dependent variables. The point x_3 now has to be slided along the time series until a set of three neighboring points are found. These points are x_t and the two previous points in time x_{t-1} and x_{t-2}. Now the multiple regression can be stated as an autoregression [Got09]:

$$x^{(t)} = a_1 x^{(t-1)} + a_2 x^{(t-2)} + \epsilon^{(t)} \tag{2.13}$$

This model is called an autoregressive model of second order, because the model goes back two time units [Got09]. An autoregressive model with the order p can generally be written as [SS06]:

$$x^{(t)} = a_1 x^{(t-1)} + a_2 x^{(t-2)} + ... + a_n x^{(t-p)} + \epsilon^{(t)} \tag{2.14}$$

Shumway [SS06] and Gottman [Got09] point out that in this equation $x^{(t)}$ is the stationary and $a_1, a_2, ..., a_n$ are the parameters of the model. The number p is representing the number of observations. It is assumed that $\epsilon^{(t)}$ is the noise with Gaussian distribution. It has zero mean and variance σ^2_ϵ.

According to Gottman [Got09] the above autoregressive model is usually written as the deviations from the mean:

$$x^{(t)} = \sum_{i=1}^{p} a_i x^{(t-p)} + \epsilon^{(t)} \tag{2.15}$$

Often, the model is also added by a constant c. In most of the literature (i.e. [SS06] or [Got09]) this constant is omitted, to provide a simpler model. The equation added with the constant is written by [WG10]:

$$x^{(t)} = c + \sum_{i=1}^{p} a_i x^{(t-p)} + \epsilon^{(t)} \tag{2.16}$$

This thesis focuses on a simple model, which calculates each variable $x^{(t)}$ independently and furthermore just includes values from the last time unit. This model is denoted as AR(1) and written as: [WG10]

$$x_i^{(t)} = c_i + a_{1i} x_1^{(t-1)} + a_{2i} x_2^{(t-1)} + ... + a_{mi} x_m^{(t-1)} + \epsilon_i^{(t)} \tag{2.17}$$

Wang and Groth [WG10] mention that in these models, each dependent variable $x_i^{(t)}$ at the time point t is modeled as a linear combination of the independent variables at the last time unit $t-1$. Each independent variable is weighted by a coefficient, which tells how the variation in the independent variable at time $t-1$ is correlated to the dependent variable at time t. I this case i denotes the total number of variables to model, and it ranges from 1 to m. The coefficients can state the influence among different variables over time, which is the main target of this thesis and provides a good tool to measure these effects.

2.3.5 Multilevel Regression Models

Andrew F. Hayes paraphrases in his primer on multilevel model citehayes the problem of cross-level analysis as an assumption "*that some progress has been made, but that for the most part the field continues to exist as islands of researchers with their theories, aware of but lacking interest in the residents of neighboring islands.*" Ritchie and Price [RP91] already speculated in 1991, that many communication researchers fail to cross levels of analysis ("*their apparent inability to or disinterest in parting the seas*" [Hay06]), because the statistical techniques are not appropriate to cross-level research. Pan and McLeod [PM91] pointed in the same year out, that statistical methods already existed, which allow communication researchers to examine independent and interactive effects of variables, which are measured at different levels on communication-related outcomes and so the crossing of levels of analysis was possible. The authors named contextual analysis and multilevel modeling as example methods. Andrew F. Hayes [Hay06] states in his primer, that perhaps it was a problem for many researchers, that such methods were not implemented in available software. He points out that this problem maybe was and is the reason that many researchers don't use these methods for their analysis. Anyway, Hayes also states that he thinks that the most plausible explanation is a lack of awareness for these statistical methods.

Andrew F. Hayes [Hay06] describes in his work an example, where doctor-patient interactions were observed. In this sample, the patients are spread across a lower number of doctors and as a result of this fact the patients are nested within the doctors. This nesting means that each patient is correlated to one doctor, but each doctor is responsible for many different patients. This nesting represents a standard feature of multilevel data. Each patient nested to the same doctor is most likely influenced similarly by the values, which characterize the doctors. This nesting now can be problematic for standard single-level regression methods, because they don't cover the "non-independence" between observations, which are characteristic for multilevel data. This dataset includes some properties, which describe doctor-patient relations, such as the length of the consultation. Such variables are all regarded as level-1 variables. Additionally, a value called "doctor business" was measured, which describes the average number of patients a doctor sees each day. This property is now a level-2 variable, because it is just an attribute of a doctor, which is a level-2 unit and as already described the patients are nested under the doctors. The author states that a level-2 variable does not change between all the level-1 units, which are nested under the same level-2 unit. Furthermore, level-1 variables can change between the level-1 units, which are nested under the same level-2 unit.

Based on the explanations by Skrondal and Rabe-Hesketh [SRH04] this example can now be summarized by saying that the elementary units are regarded as level-1 units and the clusters are regarded as level-2 units. The authors point out that it is also possible that the cluster itself is nested into a higher level, and this may result in a three-level structure. As already mentioned in above example, the units, which belong to the same cluster, share the same cluster-specific influences. It is, however, not expected, that all cluster-specific influences are included as covariates in the analysis. This problem is based on the fact that often only limited knowledge about the dataset, and the relevant covariates is given. The result of this is "*a cluster-level unobserved heterogeneity leading to dependence between responses for units in the same cluster after conditioning on covariates.*" In multilevel regression, this unobserved heterogeneity is modeled by adding random effects to the fixed effects. These random effects can be divided into two types: *random intercepts* (representing unobserved heterogeneity in the overall response) and *random coefficients* (representing unobserved heterogeneity in the effects of the independent variables on the dependent variable).

With regard to this thesis, Hayes [Hay06] states that multilevel modeling is also useful when a level-2 unit is an individual in a study, and the properties are measured repeatedly. So it is very useful in the study of time series. The measurement occasion is the level-1 unit, and it is nested under an individual, the level-2 unit. Researchers often focus on the change of these repeated measurements. According to the author multilevel modeling nowadays is often the tool of choice for the analysis of such longitudinal data.

Andrew F. Hayes [Hay06] lists in his work some basic questions, researchers want to answer with multilevel modeling of time series:

- How much on average do individuals change over time?
- What is the rate of change?
- Do individuals differ with respect to how much they change?
- What predicts how much or how quickly people change?

Hayes [Hay06] furthermore points out that multilevel analysis of time series has all the advantages of standard multilevel analysis. There is no requirement that every individual's measurement begins and ends at the same time, the distance between the different measurements does not need to be equally, and the number of measurements does not need to be the same for all individuals. Overall, Andrew F. Hayes [Hay06] states that "*in a longitudinal multilevel model, each level-2 unit is measured repeatedly on the same variable and the focus of the analysis is on estimating change in the outcome variable over time and predictors of that change.*"

With this knowledge about multilevel modeling equation 2.17 now can be rewritten as [WG10]:

$$x_{i,p}^{(t)} = c_i + a_i^T x_p^{(t-1)} + \epsilon_i^{(t)} + b_{i,p}^T x_p^{(t-1)} + \epsilon_{i,p}^{(t)} \tag{2.18}$$

Wang and Groth [WG10] state that in this equation $x_p^{(t)} = (x_{i,p}^{(t)}, ..., x_{m,p}^{(t)})^T$ is defined, and it represents a vector, which contains the variables for an individual p at time t. Furthermore, $a_i = (a_{i,1}, ..., a_{im})^T$ represents the fixed effect coefficients and $b_i = (b_{i,1}, ..., b_{im})^T$ represents the random effect coefficients. To compare the fixed effects to each other, the variables in the random effects regression equations need to be linearly transformed to represent standardized values. This can, for example, be done by subtracting their mean and division by their standard deviation. As a result of this the fixed effects can be analyzed as "*the effect of one standard deviation of change in the independent variable on the number of standard deviations change in the dependent variable.*"

2.4 The influence between content and social networks

This section gives a summary of the work done by Wang and Groth [WG10], which was inspiration for this work. The authors proposed a framework to measure the bi-directional influence of social and content network properties over time. The networks are characterized using network properties on both social and content networks (see section 2.1.2 for further details). The bi-directional influence of these properties is measured by using a set of multilevel time-series regression models (see section 2.3) to create a so called *influence network*, which states how the chosen properties of the network influence each other over the time the dataset is available. Following contributions are the output of the paper by Wang and Groth [WG10]:

- A framework for measuring the bi-directional influence between social and content network properties
- A multilevel time-series regression model
- Results on use cases described in section 2.4.2

2.4.1 Influence Framework

As described by Wang and Groth [WG10] the influence framework is based on three different stages to measure the bi-directional influence between a social network and the corresponding content network. The framework focuses on the influence of social network properties over time. As a result of this the framework needs a dataset, which varies over time. The three stages of the framework are the following:

- Network generation
- Measuring the different network properties
- Time series analysis

In the following sections the different stages are described based on the explanations by Wang and Groth [WG10]:

2.4.1.1 Network generation

The framework needs a dataset consisting of a social and content network, which varies over time. So the first step is to generate this dataset. The social network is built by informations about users, who interact over time. Based on this data a *series of social networks*, which differ over time, is established and the content of these seed users over time is collected. This series of content is then forming the content network over time.

Pieces of the content, nevertheless, should be similar to each other. On Twitter this could be for example the tweets the users produce over time. The network generation is the most domain-specific part of the three stages, because it must be decided how the social network with the user relations and how the content should look like.

In chapter 3 the datasets used in this thesis are thoroughly described.

2.4.1.2 Measuring network properties

As soon as the dataset, consisting of social and content networks over time, is collected, the network properties have to be defined and then measured. The properties have to vary over time to be suitable for the framework. It is important that the meaning of the properties for each different domain and dataset have to be taken into account. It is possible to use any possible network property that varies over time for this framework.

Detailed informations about general social network properties are listed and explained in section 2.1.2.

2.4.1.3 Multilevel time series regression models

The framework has the aim to model the longitudinal influences between the social and content network properties. The output of the stage, described in section 2.4.1.2, forms a time series. The output of this third stage is the set of statistics generated by fitting the regression models. Furthermore, a diagram is produced, which visually shows the influences of the variables. This diagram is called the influence network (see figure 2.17).

Detailed descriptions of time series analysis and multilevel regression models are provided in section 2.3. The exact model used in this paper is described in section 4.1.

2.4.2 Performed use cases

Wang and Groth [WG10] performed two use cases based on their framework. The first one is a simple use case based on the influence between co-authors of academic papers and the topics they address. The second use case focuses on the influence between social status of online forum participants and their political attention. This use case is also extended by some newly defined variables to provide more specific answers to questions about this domain. In this section only the first use case of the work is covered to give an idea, what possibilities the framework provides. The own experiments of this thesis are described in chapter 4.

As stated in section 2.4.1 each use case is based on three general stages. The different stages of the first use case covering the influence between co-authors of academic papers and the topics they address are the following based on Wang and Groth [WG10]:

Data Collection

The data is based on a corpus of meta data about the World Wide Web Conference from the Semantic Dogfood repository (see [MHHD07]). The metadata includes the program of the conference, the paper metadata and the organization metadata. The metadata covers a time span of four years and hence it is a good dataset to perform time series analysis on it. The data was stored in RDF files for each considered year.

Generation of social networks

The authors chose the co-author networks as the social networks for the analysis. For each individual year of the corpus, the co-author pairs for each article were extracted. A co-author pair states a shared authorship on a paper and if for example a work has three different authors there are three possible co-author pairs, representing the shared authorship between two persons. From these results, a weighted undirected graph was built, where authors are represented by nodes and the shared authorships of an article are represented by edges between the author nodes. The edges are furthermore weighted by the number of co-authorships between the corresponding authors. For each individual year, the degree and betweenness centrality, representing how active the author is in sharing authorships with others, and the clustering coefficient, providing a measure how close a group is, are measured.

Generation of content networks

The extracted content of the corpus describes the topics discussed at the conference each year. The keywords the authors have denoted in their papers provide the different topics. Keywords, which contained more than one word, were divided into different keywords and also were stemmed (reduce words to their root form). Based on this, a weighted undirected graph was built. A node in the network is a keyword/topic, and the edges are the co-occurrences between two keywords. The edges are again weighted by the number of co-occurrences, and the graph is produced for each year. The properties for the content network are the degree and betweenness centrality.

Generation of the influence network

The five network properties for social and content network described in the above steps are used for the creation of the influence network. The multilevel time series regression models are now used to study the bi-directional influence between these properties over time. In a greater detail, the models study the influence between the topics of a conference (content network) and the shared authorships of papers (social network). Figure 2.17 shows the resulting influence network of this use case.

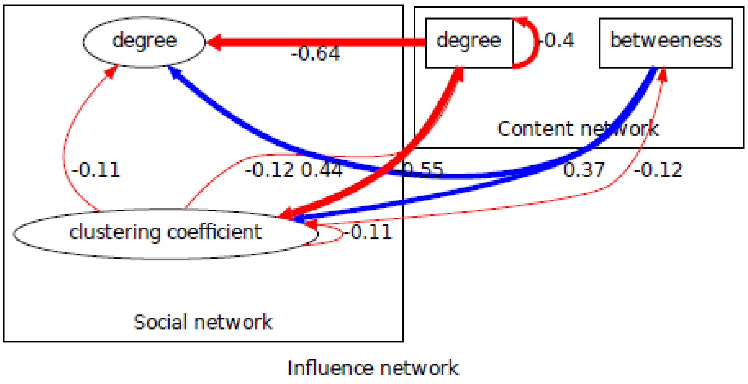


Figure 2.17: Influence Network for the WWW conference [WG10]

The figure has directional edges over time and only shows influence effects, which are significant. If an edge between two properties is existing, this can be read as the value of one property at some time t has a positive (blue arrow) or negative (red arrow) effect on the value of the other property at time $t + 1$. For example, the degree property in the content network at time t has a negative effect on the degree centrality and the clustering coefficient in the social network at time *t + 1*. Wang and Groth [WG10] point out that a plausible interpretation for this is that "*after a burst of the collaboration on a trendy and popular topic, the topic is becoming less popular and the shared authorship between more authors is dying*". The degree centrality on a topic also has a large negative effect on itself. Wang and Groth [WG10] argue that this is the case, because "*a popular topic of the conference in one year is very likely to be less popular the next year*". Positive effects can be seen on the figure by looking for blue edges. The betweenness centrality of the content network has strong positive effects on the degree centrality and clustering coefficient in the social network. The authors point out that a possible explanation for these effects could be, that "*if a topic bridge the gap between other topics in one year, it is possible that in the next year more authors focus on a collaboration for this topic.*"

3 Data Sets and Data Collection

Part of this work is the crawling of two different Twitter datasets and a Boards.ie dataset. The problem with Twitter datasets is, that only a few are available for free and most of them do not provide temporal information about social network data. Nevertheless, we have not found any suitable datasets, which collect social and content networks of different Twitter users over time. So we had to crawl it by ourself.

The crawler code is written in Java using the twitter4j library[1] for Twitter API access. The complete social and content networks of each dataset were crawled each day and over a time period of one month. The data was stored in a local MySql database. The crawling process was started each day about the same time, which was at about 12.00 pm CET. Due to API restrictions, the completeness of data can not be guaranteed. Twitter allows accessing some percentage of their live-stream via their API. So the start of the crawling process can vary some hours each day. Sometimes the Twitter API was not accessible over some time period, so that on some days some users could not be retrieved. How this is handled will be described in each of the use case sections in chapter 4. When retrieving tweets from a user, the Twitter API restricts up to date the resulting set by 3200 tweets [Twi11a], so if the user has more than 3200 tweets overall, it was impossible to get his complete timeline.

The following sections will describe the crawled datasets in a greater detail. Each section is split into a description of the data collection and a description of the dataset.

3.1 First Twitter Dataset

This dataset represent the first Twitter dataset of this work and is used for the first Twitter experiment described in section 4.2. The following two sections describe at first the data collection and then the dataset itself.

[1] http://twitter4j.org/

3.1.1 Description of the data collection

The first dataset is based on 1500 randomly chosen seed users, using the public timeline method of the Twitter API. The decision using the public timeline method and not using just random user IDs is based on the fact, that the timeline method is biased towards active Twitter users. Normally, it would be recommended, to use random ids, because random datasets should generally not be biased. For this analysis, we decided to collect more active users for the dataset because time series analysis can present more valuable results when the users of the networks provide more information at the different time points..

The users were collected according the following rules: they tweet generally in English language, have at least 80 followers, 40 followees and 200 tweets. These conditions are necessary to ensure that our sample set consists of active Twitter users. For the sake of simplicity we focus on the English-speaking popularity of Twitter. Furthermore, the user accounts should not be protected, to be able to get all information needed. To determine, if a user tweets generally in English, the language of the tweet received from the public timeline method has been analyzed, using the Java library textcat[2]. As a second step the language property of the relevant Twitter account is checked through the Twitter API, if the user has selected English as her main Twitter language. These two checks give a good likelihood, that the chosen user tweets mostly in English.

In Appendix A.1 more details about the starting seed users can be found. Appendix A.2 shows details about the crawled users each day of the complete crawling time span of this dataset. As already mentioned in chapter 3, the Twitter API sometimes had temporal unavailability. Also some users have deleted or protected their accounts during the crawling timespan, so that it was not possible to get information about them any longer. It was also possible, that some users unprotected their accounts after a while again, so that it was again possible to crawl their data. Table A.2 shows the relevant information about each day and also shows a note, if some temporal API problems occurred. The large difference of the crawled users on the first day and from the original count of seed users (1500) happened, because the random seed user sample was crawled about a week before the first day of the dataset crawling process, so in that time some seed users deleted or protected their accounts. How this is handled will be described in section 4.2.

3.1.2 Description of the dataset

The complete dataset is built by the individual datasets of each user, including following parts:

[2] http://textcat.sourceforge.net/

- Social Network: All followers
- Social Network: All followees
- Content Network: Up to 3200 recent tweets

The social networks, consisting of the followers and followees, are represented by edges. For each tweet the complete status text is stored. Furthermore each tweet is attached by a value, stating, if it is a retweet, and by the count of actual retweets.

3.2 Second Twitter Dataset

This dataset represent the second Twitter dataset of this work and is used for the second Twitter experiment described in section 4.3. The following two sections describe at first the data collection and then the dataset itself.

3.2.1 Description of the data collection

The second dataset is based on the users of a personal Twitter list from Stefano Bertolo, who is a professional in knowledge representation, information retrieval, natural language representation and many similar topics. Stefano Bertolo has set up a user list[3] on his Twitter profile[4], which includes 137 Twitter users, who tweet about topics around the semantic web.

Out of this semantic web user list all users were extracted and these user form the group of seed users for this dataset. The decision to take these Twitterers from the user list as sample for this dataset is based on the fact, that Stefano Bertolo has a good insight into the topic of semantic web, because he is a researcher at the EU and is responsible for this topic. So it is very likely that the users really tweet about semantic web topics throughout their timeline. A further very important criterion is that it is very likely that the users are heavily linked together. So there might exist many edges within the group of users of the semantic web list.

Based on these possible edges between the users, this dataset differs from the first dataset, which is described in section 3.1, in the collection of the followees and followers. The first dataset consists of all edges linking from and to the seed users, whereas the second dataset only consists of edges between the seed users. So the main target of the second dataset is to create a closed network of users, which might be linked together with a lot of edges. Anyway, the selection of the seed users is not based on the criteria

[3] `http://twitter.com/#!/sclopit/semweb`

[4] `http://twitter.com/#!/sclopit`

of linked together users, but it is based on a sharing of the same topic they tweet about.

In Appendix B.3 more details about the seed users can be found. Appendix B.4 shows details about the crawled users each day of the complete crawling time span of this dataset. As mentioned in section 3.1, the crawling process of the second dataset had the same problems as the first dataset. The Twitter API sometimes had some temporal unavailabilities and there is also the possibility that some users got deleted or protected during the timespan of the crawling process. It also could be that some users unprotected their accounts after a while, so that it again more possible to crawl their data. Table B.4 describes the different crawled users for each day and there is also a note added, if some temporal API problems occurred during the crawling process. The table shows that it was possible to crawl information of 134 users at the beginning of the one month crawling period. So out of the 137 users of the user list three users have protected their accounts, so that it was impossible to access their information through the Twitter API or there were simply no links to or from other seed users. How this is handled will be described in section 4.3. Throughout the complete crawling process each day it was possible to crawl information about each of the 134 users.

3.2.2 Description of the dataset

The complete dataset is built by the individual datasets of each user, including following parts:

- Social Network: Followers of the set of seed users
- Social Network: Followees of the set of seed users
- Content Network: Up to 3200 recent tweets

The social networks, consisting of the followers and followees, are represented by nodes and edges. For each tweet the complete status text is stored. Furthermore each tweet is attached by a value, stating if it is a retweet, and by the count of actual Retweets.

3.3 Boards.ie Dataset

This dataset represent the Boards.ie dataset of this work and is used for the Boards.ie experiment described in section 4.4. The following two sections describe at first the data collection and then the dataset itself.

3.3.1 Description of the data collection

This third dataset is based on a collected subset of a popular Irish bulletin board called Boards.ie[5]. The dataset was originally used for a competition[6], where the participants had to do something interesting with the data. In [BB08] it is stated that the completely dataset contains of ten years of discussions and is available in SIOC (Semantically-Interlinked Online Communities) format. The data is built upon a top-down link structure, so the top-level site document links to different users and also to the top-level forums. A forum links to sub-forums and threads, which then link to the different posts. There is also a linked list between the posts, so that it is stated which post replies to another and where quoting has been done. There are also FOAF (Friends of a Friend) files, describing a social network based on user's buddy lists, but for this work, the social network is extracted in a different way and is described later in this section. The complete dataset contains around nine million documents and uses about 50 GB of disk space. As a result of this the original data is split into several parts, so that it is possible to just download the data u really need.

The analysis of the third dataset (MySQL database storage) in this thesis focuses on properties calculated for the year 2006. The social network is built upon information about posts and replies. At first, all posts in the time period between the beginning of July 2005 and the end of 2006 were extracted. As a second step for each post out of this subset the reply information was collected to determine the hierarchical structure of the post. If the post is a reply to another post by a user, an edge between these two users was drawn. If a has replied to a post made by himself, the edge was discarded.

Based on these calculated social edges the in- and out-degrees for the corresponding users were calculated. The analysis needs information about the degrees for each beginning of a month in the year 2006. So to build the social network in the beginning of a month all edges in the past six months were collected and then the in- and outgoing ties were summed to calculate the degrees. The result of this calculation is a social network for each month (12 times 01.01.2006,..., 01.12.2006).

For the content information for each beginning of a month in 2006 (same 12 dates as for the social network) a topic distribution for each user is available. These topic distributions have been calculated by my advisor Claudia Wagner. The topic distributions state the probability of the posts of a user in the last six months (same as for social network calculation) to be assigned to a topic. In total, there are 650 different topics available. The complete sum of the probabilities of a month and a user is always 1 and if the probability for a topic is zero, then the topic is not listed in the corresponding table of the database. Overall, there are topic distribution information about 29.886 distinct users.

[5] `http://www.boards.ie/`

[6] `http://data.sioc-project.org/`

In Appendix C.5 more details about the distinct users of the topic distributions can be found. Appendix C.6 shows details about the number of different users for each month of the topic distributions. How the case will be handled, that there is not information on every month for some users and that maybe no social network information is available for some users out of the topic distributions, will be described in section 4.4.

3.3.2 Description of the dataset

The complete dataset is built by the individual datasets of each user, including following parts:

- Social Network: In-Degree for each user posted in the corresponding time span
- Social Network: Out-Degree for each user posted in the corresponding time span
- Content Network: Topic distributions for 29886 distinct users

4 Experiments

In this chapter the experiments for the three datasets are presented and discussed. At the beginning section 4.1 covers details about the experimental setup. The technological features and tools are described and furthermore also the exact regression model is stated. In the following sections each individual experiment is explained. Finally, results are discussed and potential conclusions are drawn.

4.1 Experimental setup

The aim of the experiments in this thesis is to analyze the bi-directional influence between social and content networks. The different datasets described in chapter 3 are already built to provide a content and a social network for distinct individuals, and they are provided as a time series. For a period of time exist different time points, where properties are measured. Sometimes it is necessary to bring the measurements of the datasets in another view, so that they can be used for the regression analysis. The exact procedure of this step is described in the *Preparation Phase* part of this section.

As already stated in chapter 3 all three different datasets are stored in a MySQL[1] database. To access the data and calculate the important properties Java[2] is used and the calculations are stored back into the MySQL database.

Because this work is inspired by the work by Wang and Groth [WG10], the same framework for the regression analysis is used. The framework is written in Python[3] and uses the library rpy2[4], which provides a low-level interface to R[5] for Python. The package is used for statistical and mathematical calculations in the framework. The framework is based on different steps, which are the following:

[1] `http://mysql.com/`
[2] `http://java.com/`
[3] `http://www.python.org/`
[4] `http://rpy.sourceforge.net/rpy2.html`
[5] `http://www.r-project.org/`

Preparation phase

As first step the available time-series data has to be prepared for the analysis. The input for this step is a text file (txt) containing the time series data. Each row of this file contains the individual, the date of the time-series and the social and content network properties. Each content of a row is space separated. A sample input containing time-series data of two users measured on two days (15.03. and 16.03.) could be stated by a text file as shown in figure 4.1. The "OutDegree" and "InDegree" values represent the social network properties, and the "Numtweets" value states the content network property. If one single user is not present in one collection date, all his properties have to be zero at the corresponding day.

```
User Day OutDegree InDegree Numtweets
1 1503 12 10 3
1 1603 13 12 2
2 1503 22 30 10
2 1603 25 33 13
```

Figure 4.1: Sample input for the preparation step

In addition to this file, another text file has to be provided, which states which property columns are social network properties, which are content network properties and which column states the individual.

At the beginning of the preparation phase, the framework normalizes the property values and saves the result in a new file. This normalization is done by using R and subtracting the mean of the vector and dividing the result by the standard deviation of the vector. According to Andrew Gelman [Gel08] this is a common method in applied regression. The subtraction of the mean improves the interpretation of main effects, and the division by the standard deviation brings all values to a common scale. This rescaling helps to interpret the regression coefficients more directly. As already stated in section 2.3.5 the fixed effects can be analyzed as "*the effect of one standard deviation of change in the independent variable on the number of standard deviations change in the dependent variable*" [WG10]. The output of the standardization for the sample input presented in figure 4.1 can be seen in figure 4.2:

Finally, this step is finalized by preparing the standardized data for the regression analysis in the next step. This is done by extending each line of the normalized dataset with the properties of the last time point, whereas the first time point of the data is just used for the next time point, because there is no information about data before it. This step is needed, because the regression model always needs information about the independent variables at time $t-1$ for the dependent variable at time t (see section 2.3 for further details). The used regresson model for the experiments is defined in the

```
User Day OutDegree InDegree Numtweets
1 1503 -0.925820099772551 -0.943249068338217 -0.747087367637628
1 1603 -0.771516749810459 -0.775560345078089 -0.933859209547035
2 1503 0.617213399848368 0.733638164263057 0.560315525728221
2 1603 1.08012344973464 0.985171249153249 1.12063105145644
```

Figure 4.2: Normalization of the sample input

analysis phase to make it clearer, why this preparation is exactly needed. The final output of the preparation phase for the sample input (see figure 4.1) now is the following text file (for each user the information is written in one line; the line break in this case is just due to formatting problems):

```
User Day OutDegree InDegree Numtweets l1.Day l1.OutDegree l1.InDegree l1.Numtweets
1 1603 -0.771516749810459 -0.775560345078089 -0.933859209547035
  1503 -0.925820099772551 -0.943249068338217 -0.747087367637628
2 1603 1.08012344973464 0.985171249153249 1.12063105145644
  1503 0.617213399848368 0.733638164263057 0.560315525728221
```

Figure 4.3: Final preparation of the sample input

Analysis phase

The analysis phase is needed to identify statistical significant influences between social and content network properties. The input for this step is the standardized and prepared data of the preparation phase (see figure 4.3) and the information about the social and content network properties. To perform the regression analysis the R package lme4[6] is used. This package allows the fitting of linear and generalized linear mixed-effects models. The framework uses the ideas and final model as described in section 2.3.5. The data always contains a time series, where information about properties of an individual at different time points t is available. So we can now define (as already stated in the corresponding section) $x_p^{(t)} = (x_{i,p}^{(t)}, ..., x_{m,p}^{(t)})^T$ a vector containing the measured properties for an individual at time t. In the experiment of this thesis the individual is always an user. The final equation used in this framework for all the three different experiments is [WG10]:

$$x_{i,p}^{(t)} = c_i + a_i^T x_p^{(t-1)} + \epsilon_i^{(t)} + b_{i,p}^T x_p^{(t-1)} + \epsilon_{i,p}^{(t)} \tag{4.1}$$

In this equation $a_i = (a_{i,1}, ..., a_{im})^T$ represents the fixed effect coefficients and $b_i = (b_{i,1}, ..., b_{im})^T$ represents the random effect coefficients [WG10].

[6] http://cran.r-project.org/web/packages/lme4/index.html

Equation 4.1 is now used to determine the bi-direction influence by using the lme4 package. Because a mixed-effect model is needed the function *lmer* out of the package is used. This function generally provides the functionality to fit a linear mixed model, a generalized linear mixed model or a nonlinear mixed model [BMB11]. The data for the function is provided by the standardized prepared data, and the formula is defined in [BMB11] the following way:

"*A two-sided linear formula object describing the fixed-effects part of the model, with the response on the left of a $\sim$ operator and the terms, separated by + operators, on the right. The vertical bar character "|" separates an expression for a model matrix and a grouping factor.* "

The formula for Equation 4.1 and the sample normalized and prepared data in figure 4.3 can now be written in the following way to use it in the R code:

```
OutDegree ~ 1  + l1.OutDegree + (l1.OutDegree- 1|User) + l1.InDegree +
    (l1.InDegree- 1|User) + l1.Numtweets + (l1.Numtweets- 1|User)
```

This above code fragment states the effects on the dependent variable "OutDegree". The constant of equation 4.1 is 1 and the random effects are subtracted by 1 and are nested under the level-1 individual "User". The subtraction by 1 is a way to prevent that for each separate dependent variable in the model an additional regression intercept would be estimated. The formula calculates one regression constant (the 1 at the start), which is not variably conditioned on person. The independent variables are always the variables at time $t-1$ and also the dependent variable at time $t-1$ becomes a predictor to determine the influences between the variable itself at different time steps. The values at time $t-1$ are often as well called "lag values". Note that this code just states the effects on the dependent variable "OutDegree". For the other variables, this can be done similarly.

For each dependent variable the output of the *lmer* function is a table for the random and fixed effects. For the fixed effects information about the estimate, the standard error and the t-test (a statistical hypothesis test to determine if the slope of a regression line differs significantly from zero) are available. To determine if the estimate is statistically significant Wang and Groth [WG10] chose to calculate, if the estimate is larger than two times the standard error. If the estimate would be equal or smaller than the standard error, it could not be guaranteed that it is a statistical significant influence, because the standard error could falsify the estimate result. If the estimate is now at least two times larger then it can be stated that the influence is statistical significant. This could also be done by using the t-values. Normally, it can be stated that the estimate is statistically significant if the t-value is larger than two, but to be sure a t-table should be consulted.

Finally, the resulting statistically significant influences are drawn into a "dot" file, which is used for layered drawings of directed graphs. This dot file then can be printed to a "pdf" file using Graphviz[7]. This graphical influence network now shows all the bi-directional influences between the different social and content network properties. If an arrow between two properties is existing, this can be read as the value of one property at some time t has a positive or negative effect on the value of the other property at time $t + 1$. Red arrows represent negative effects and blue arrows represent positive effects. A sample influence network can be seen in figure 2.17.

For social network analysis in Java the library JUNG[8] was used. This library was used to calculate network properties and to visualize the graph. Further plots of different distributions have been done by using Matlab[9].

4.2 First Twitter experiment

This experiment was performed on the first Twitter dataset of this thesis, which is described in section 3.1. To recap the specifications of this dataset the main characteristics are:

- 1500 randomly chosen Twitter users (using the public timeline)
- At least 80 followers, 40 followees and 200 tweets
- All followers
- All followees
- Up to 3200 recent tweets

The dataset was crawled over a time period of one month.

The main aspect of this dataset is, that it is a randomly chosen subset of the complete Twitter dataset based on 1500 seed users. This dataset consists of all edges linking from and to the corresponding seed users. So it should not be seen as an isolated network, but more as a part of a much greater network. As a result of this it is not very useful to calculate and analyze specific network properties (see section 2.1.2), because there is no accurate data available on the "world" outside of this sub-network.

For the study properties are required. Due to the structure of this network, properties are used, which can be measured for each seed user and include the hidden outer part of the network too. Each property is available at all observation dates of the time-series. The following descriptions list the different social and content network properties in detail:

[7] http://www.graphviz.org/

[8] http://jung.sourceforge.net/

[9] http://www.mathworks.com/

Social network properties

Each of the two social network properties is calculated for the present social network of the crawled dataset.

Out-degree

The out-degree of a user states how many other Twitter users a person is following. So this value is equivalent to the number of followees available in the dataset. It is measured by counting the edges that are pointing away from the corresponding user.

In-degree

The in-degree of a user states how many other Twitter users follow the corresponding seed user. This value is equivalent to the number of followers available in the dataset. It is measured by counting the edges that are pointing to the corresponding user.

Content network properties

The content network properties are calculated for each day of the timeframe. Because for each day all 3200 recent tweets are available in the dataset, only the tweet information of one specified day is required. This is done by going one day ahead (time point $t + 1$) and gather information for time t there. The reason for this is, that it is guaranteed that all tweets of the previous day are collected and there is only one possibility that the right information is not available, if a person tweets more than 3200 tweets on a day, what is not the case in this dataset.

Example: The content network properties of the observation date "02.04." should be calculated. This is done by looking at the tweet information of the time point "03.04." because all the tweets of the "02.04." are there available.

Hashtag ratio

This value expresses the number of used hashtags, normalized by the number of tweets. So it is calculated by going through all tweets on the corresponding day and counting the number of hashtags using regular expressions and then dividing that number by the number of total tweets that day.

Link ratio

This value expresses the number of used links, normalized by the number of tweets. So it is calculated by going through all tweets on the corresponding day and counting the number of Links using regular expressions and then dividing that number by the number of total tweets that day.

Retweet ratio

This value expresses the number of retweets out of the tweets, normalized by the number of tweets. It is calculating by counting how many tweets of the users are retweets and then dividing this number by the total number of tweets that day.

Retweeted ratio

This value expresses the number of retweets by another user of the tweets, normalized by the number of tweets. It is calculated by counting, how often the tweets of that day got retweeted by another Twitter user and then dividing this number by the number of total tweets that day.

Number of tweets

This value simply expresses the total number of tweets produced that day by the seed user. It is calculated by summing up all the tweets of that observation date.

As already mentioned in section 3.1 and stated in appendix A.2, it was impossible to get all data for every single seed user for each day. Even though multilevel regression analyses and time-series analyses can handle sparse data (see section 2.3.5), the decision has been made to only use individuals for this experiment, where observations at each time step of the time series are available. This should provide more accurate results. The final number of individuals, for which the analyses has been done, is 1188 users, which is still a high number and is above the originally planned number of 1000 users for this analysis.

The study uses the multilevel regression model and methods described in section 4.1. The social and content networks are binded together via the seed users. Figure 4.4 shows the resulting influence framework for the first experiment.

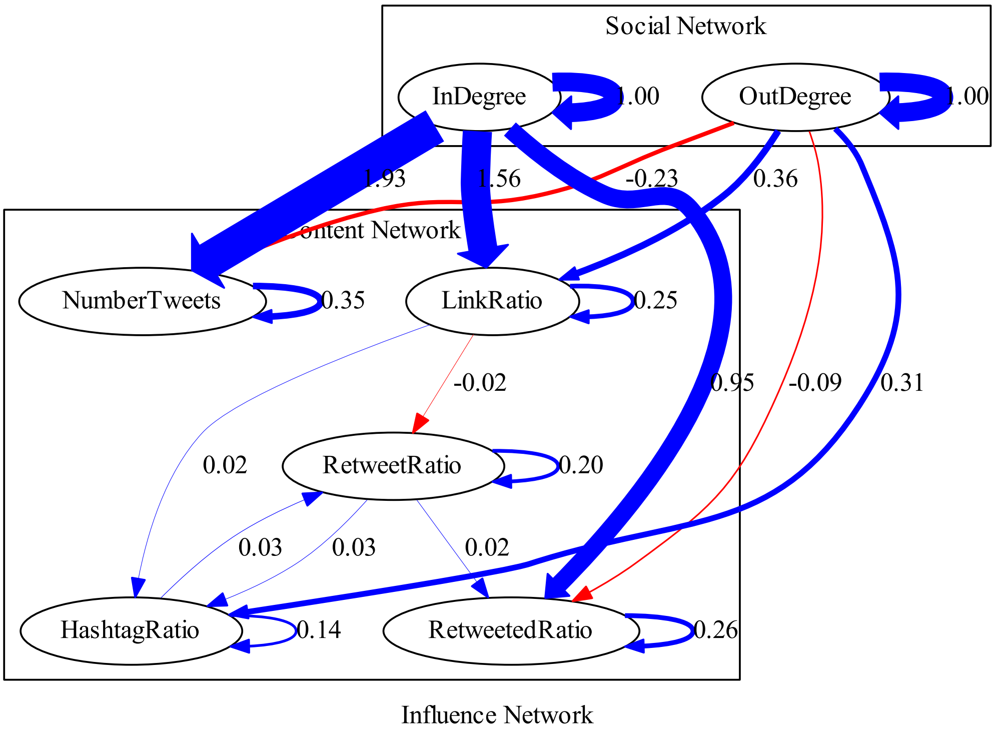


Figure 4.4: Influence network of the first dataset

The following table summarizes all the detected influences of the different properties. The influences are represented by the way that the properties in the first column have a corresponding influence on a value stated in the first row. An empty cell represents that there is no influence between the two corresponding properties. This table just illustrated the different arrows shown in figure 4.4 in a clearer way.

	Out-Degree	In-Degree	Link ratio	Hashtag ratio	Retweet ratio	Retweeted ratio	Number of tweets
Out-Degree	1.00		0.36	0.31		-0.09	-0.23
In-Degree		1.00	1.56				1.93
Link ratio			0.25	0.02	-0.02		
Hashtag ratio				0.14	0.03		
Retweet ratio				0.03	0.20	0.02	
Retweeted ratio						0.26	
Number of tweets							0.35

Table 4.1: Influences of the first experiment

When looking at the influence network it becomes clear that there is a statistically significant positive impact between all properties on themselves. This suggests that each of the properties one day is very likely to get higher the next day. This can be explained by the way Twitter works. Twitter always tries to push users to become more active on the platform. This results in following more users, being followed by more users and posting more tweets with more functionalities like hashtags, links or retweets. This observed effect is the strongest for the out- and in-degree values, which represent the social network properties. This seems to tell us that the social features of Twitter are working very well, and it is easy to attract more followers and followees as a Twitter user. Figure 4.5 compares this result with a plot, representing the average out-degree of all seed users each day and figure 4.6 represents the same plot with the in-degree. These two plots support the detected results by the regression analysis and show that the average out-degree and in-degree over all seed users it growing in time. When looking at these results it should be kept in mind that the seed users represent some sort of active users and have been chosen with minimum requirements. When looking at not so active users the results may differ.

No influence between the two social network properties is identified but there are a lot of influences between the social and content network properties. The in-degree has the strongest positive effects of all measures. It influences the link ratio, the retweeted ratio and the number of tweets. These phenomenons indicate that a user might get more motivated to write tweets and also use links in his tweets, when the Twitterer has a high number of followers. Furthermore, the tweets are more likely to get retweeted, because more followers are reading them.

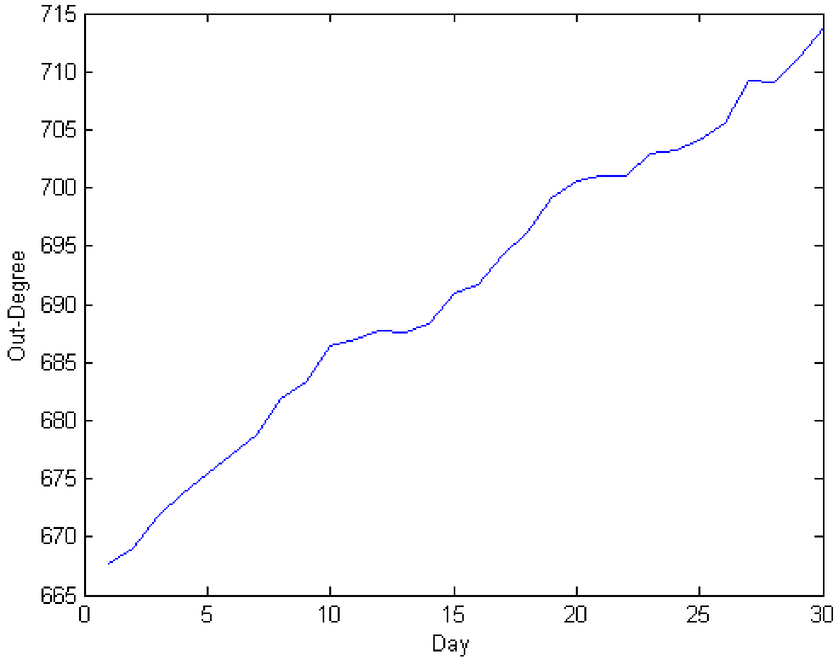


Figure 4.5: Average out-degree of all users each day

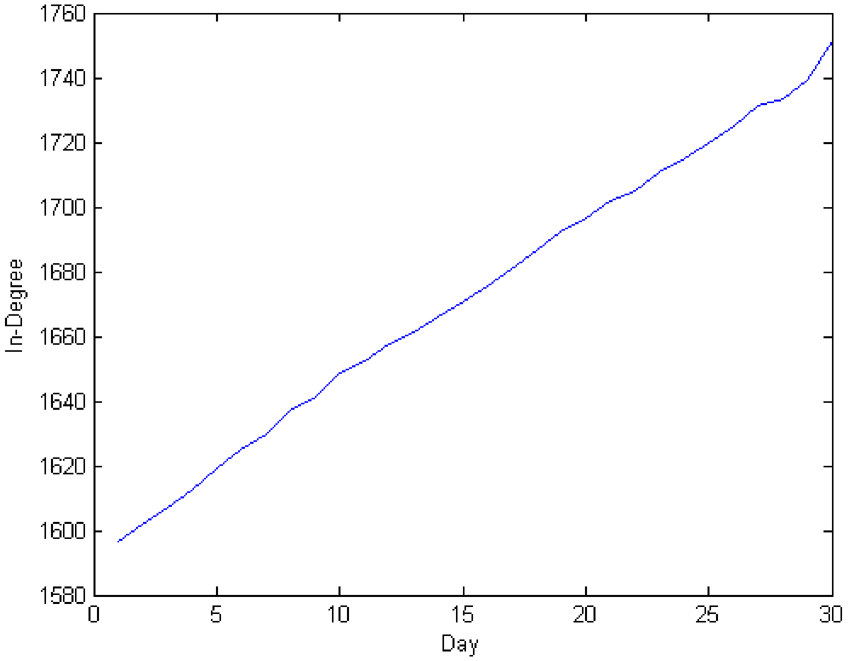


Figure 4.6: Average in-degree of all users each day

The out-degree, on the other hand, shows positive and negative influences. The positive effects point to the link and hashtag ratio. This can be explained by the fact that the higher the number of followees is, the higher the number of tweets a user is reading. As a result of this the user might get influenced by different links and hashtags and might use them more often in his own tweets. The out-degree also shows negative effects on the number of tweets and the retweeted ratio. One interpretation for this is that a user who follows many other users, is less focused on providing more content himself and thereby the number of tweets per day is shrinking and furthermore the tweets are also less likely to get retweeted. This result is the opposite of the result regarding the positive influences of the in-degree, where the user gets motivated to provide more content. In this case the user starts behaving more passive like a reader, rather than as a writer.

The influences between the content network properties themselves are not as strong as the ties between social network properties and content network properties. Furthermore, there are no influences between content and social network measures. This means that the content produced by the seed users does not affect their number of followers and followees. A weak positive impact can be seen between link and hashtag ratio. The link ratio, furthermore, has a weak negative effect on the retweet ratio and the hashtag ratio and the retweet ratio influence each other positive. Anyway, these connections are very vague, and so it is not practical to try to understand them, because they don't seem to be generally important.

Overall, the analysis of this experiment states that there is a powerful influence between social and content network properties but not vice versa. This tells us that if a user's social network (regarding followers and followees) is growing the user is most likely providing more content with more features in the future. On the other hand, it was shown that the content properties don't influence other properties heavily and that there are no strong negative effects over all, which states that the concept of Twitter regarding the user activity is working.

When these results are compared with the results stated in section 2.2.2 some differences can be found. It was shown that Cha et al. [CHBG10] did not find a correlation between the in-degree of a user and other measures. On the other hand, Suh et al. [SHPC10] showed that there is a correlation between the number of followers and followees with the retweetability of a user's tweets, but that the number of followers is much stronger related to the retweetability than the number of followees. This experiment approves the findings by Suh et al. [SHPC10] by getting similar results stating that the in-degree shows a strong influence on the retweeted ratio. Anyway, the out-degree is not correlated with the retweeted ratio in this study. It should be kept in mind that these three works differ distinctly in the datasets, and especially this work just focuses on a randomly small dataset of Twitter seed users. Furthermore, the strategies for choosing the datasets differ. The other two works used tweets (the content) as a starting point for creating their dataset, whereas this experiment focuses on seed users and their related content represented by their tweets. This thesis also tries to find correlations between all the measured properties instead of trying to get influences on a single property (like the retweetability).

The following experiment provided in section 4.3 extends this analysis by looking at a smaller number of seed users, who have something in common.

4.3 Second Twitter experiment

This second experiment was performed on the second Twitter dataset of this thesis (see section 3.2 for further information). The main characteristics of this dataset are recapped in the following list:

- 137 Twitter users
- Members of Stefano Bertolo's semantic web user list
- Followers of the set of seed users
- Followees of the set of seed users
- Up to 3200 recent tweets

This dataset is crawled on a daily basis over a time period of one month.

In contrast to the first dataset, this dataset focuses only on relations between the seed users and not on relations to the outside Twitter world. Even though there are many links between the different users, it was not a specific criterion to crawl this list of seed users. It is based on the fact that Stefano Bertolo has a good insight into the topic of semantic web and about important Twitter users for this topic. This experiment is designed to study potential influences between properties of a complete network with a specific domain. The main interest is to analyze network properties and furthermore the role of a user inside this network. The links to the outside Twitter region are discarded. Figure 4.7 illustrates a subset showing 33 vertices out of the dataset and the directed edges between them.

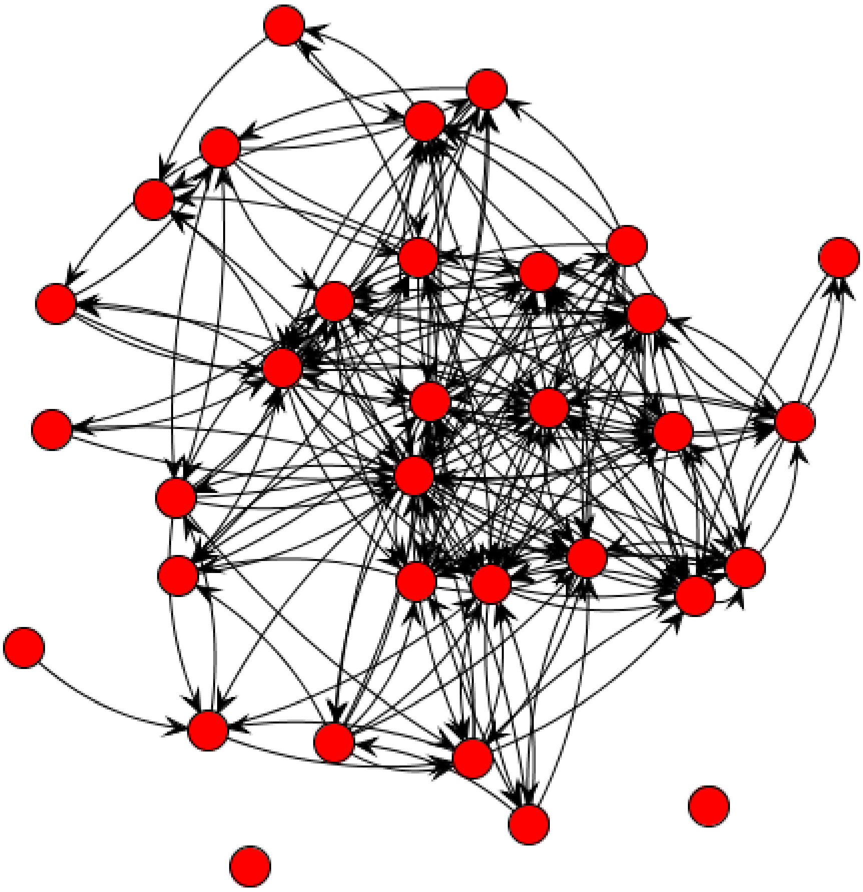


Figure 4.7: Sample of the second dataset

The analysis of this dataset of course needs different computed social and content network properties. The social network properties include also network specific values, whereas the content network properties are almost the same as in the first experiment (see section 4.2). The following list shows the different social and content network properties of this experiment:

Social network properties

Each of the following four social network properties is calculated based on the social network of the crawled dataset.

Out-degree centrality

The out-degree of a user states how many other Twitter users a person is following. So this value is equivalent to the number of followees available in the dataset. It is measured by counting the edges that are pointing away from the corresponding user. Finally, this value is divided by the number of other seed users in the dataset (n-1) to calculate the out-degree centrality.

In-degree centrality

The in-degree of a user states how many other Twitter users are following the corresponding seed user. This value is similar to the number of followers available in the dataset. It is measured by counting the edges that are pointing to the corresponding user. Finally, this value is divided by the number of other seed users in the dataset (n-1) to calculate the in-degree centrality.

Betweenness centrality

The betweenness centrality of a user states the number of geodesics between other nodes that pass through the vertex. It is also a measure to state the centrality of the user, and a user is central if it lies between other users on their geodesics. Further information can be found in section 2.1.2.

Clustering coefficient

The clustering coefficient generally states a measure of how many friends of a user are also friends of each other. It is a measure to determine the cliquishness of a friendship circle. Further information can be found in section 2.1.2.

Content network properties

The content network properties are all measured for the day of observation. Because for each day all 3200 recent tweets are available in the dataset, only the tweet information of the specified day is required. This is done by going one day ahead (time point $t + 1$) and gathering information for time t there. The reason for this is, that it is guaranteed that all tweets of the previous day are collected and there is only one possibility that the complete information is not available, if a person tweets more than 3200 tweets on a day, what is not the case in this dataset.

Example: The content network properties of the observation date "02.04." should be calculated. This is done by looking at the tweet information of the time point "03.04." because there all the tweets of the "02.04." are available.

Hashtag ratio

This value expresses the number of used hashtags, normalized by the number of tweets. So it is calculated by going over all tweets on the corresponding day and counting the number of hashtags using regular expressions and then dividing that number by the number of total tweets that day.

Link ratio

This value expresses the number of used Links, normalized by the number of tweets. So it is calculated by going through all tweets on the corresponding day and counting the number of Links using regular expressions and then dividing that number by the number of total tweets that day.

Favorite ratio

This value expresses the number of favorite tweets, normalized by the number of tweets. So it is calculated by going over all tweets on the corresponding day and counting how many tweets were favorited by other users and then dividing that number by the number of total tweets that day.

Retweet ratio

This value expresses the number of retweets out of the tweets, normalized by the number of tweets. It is calculating by counting how many tweets of the users are retweets and then dividing this number by the total number of tweets that day.

Retweeted ratio

This value expresses the number of retweets by another user of the tweets, normalized by the number of tweets. It is calculated by counting, how often the tweets of that day were retweeted by another Twitter user and then divided by the total number of tweets that day.

Number of tweets

This value simply expresses the number of total tweets produced that day by the seed user. It is calculated by summing up all the tweets of that observation date.

As already mentioned in section 3.2 not all 137 users could be crawled. However, figure B.4 shows that it was possible to crawl 134 users each day over the entire time period. So the analysis could be performed on all of these 134 users.

The study uses the multilevel regression model and methods described in section 4.1. The social and content networks are binded together via the seed users. Figure 4.8 shows the resulting influence framework for the second experiment.

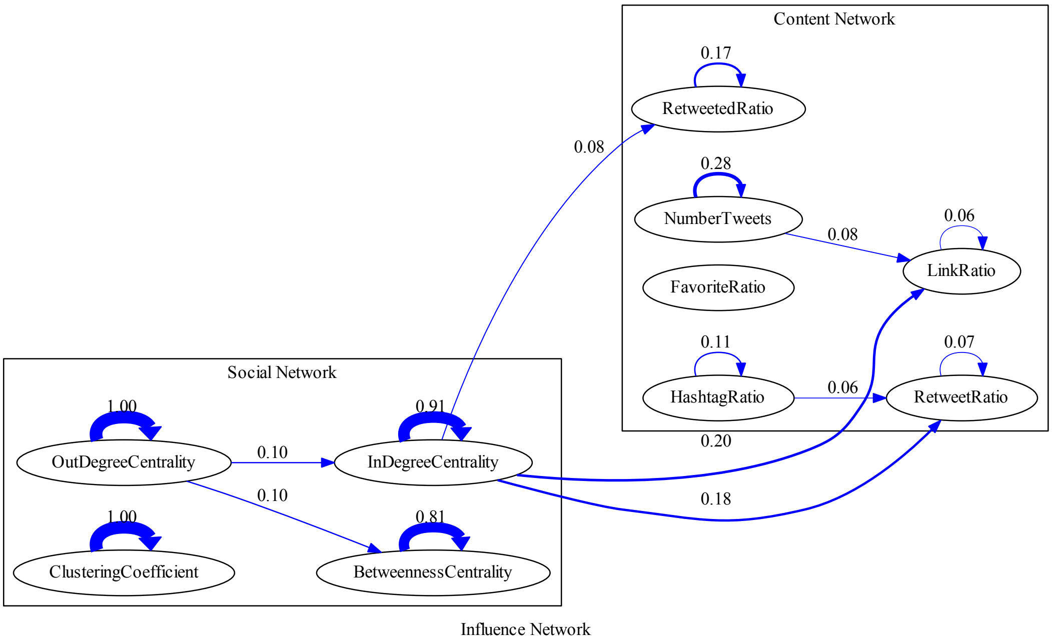


Figure 4.8: Influence network of the second dataset

Table 4.2 summarizes all the detected influences of the different properties. The influences are represented in such a way that the properties in the first column have a corresponding influence on a value stated in the first row (ODC = Out-Degree Centrality, IDC = In-Degree Centrality, BC = Betweenness Centrality, CC = Clustering coefficient). An empty cell represents that there is no influence between the two corresponding properties. This table just illustrated the different arrows shown in figure 4.8 in a clearer way.

	ODC	IDC	BC	CC	Link ratio	Hashtag ratio	Favorite ratio	Retweet ratio	Retweeted ratio	Number of tweets
ODC	1.00	0.10	0.10							
IDC		0.91			0.20			0.18	0.08	
BC			0.81							
CC				1.00						
Link ratio					0.06					
Hashtag ratio						0.11		0.06		
Favorite ratio										
Retweet ratio								0.07		
Retweeted ratio									0.17	
Number of tweets					0.08					0.28

Table 4.2: Influences of the second experiment

Like in the first dataset (see section 4.2)) it becomes clear that almost all properties show a positive impact on themselves, which is represented by blue arrows pointing to the same property where the arrow also starts. That means that all these properties are likely to increase over time. These results indicate that users increase their usage of content features and especially a growth of the social network properties can be observed. Overall, the influence network does not show any negative influences, which can be identified by the lack of red arrows. The only property, where no effect at all could be determined, is the favorite ratio.

As already mentioned, especially the social network properties show a strong influence on themselves. This can be interpreted as the fact, that this social community of semantic web Twitter users is getting more and more connected each day. The growing out-degree and in-degree centrality indicates that generally a rise in the number of edges between the users can be explored. This is also approved by looking at the actual number of edges of the dataset. At the first day 3.539 links exist, whereas the last day of the considered time frame the social network consists of a total of 3.559 edges. This change might not seem to be much, but it is actually a good number of new follower and followee relationship actions inside this already highly evolved network of semantic web Twitterers. The positive influence of the clustering coefficient on itself also suggests, that it is becoming higher the next day. This indicates that the friends of a node are as well friendshipping each other more on the next time step. So they show a higher cliquishness. The positive effect of the betweenness centrality on itself tells us, that a central user is likely to be even more central the next day. This means that the node lies between more other actors on their geodesics. There are only two influences between the social network properties, which are coming from the out-degree centrality: on the one hand, there is a positive effect on the in-degree centrality, which indicates that a person who is following a lot of other persons, is also likely to get followed by others (Twitter's inofficial rule: "I follow you, you follow me"). This indicates that the reciprocity is growing (see section 2.2.1.1). On the other hand. the second positive influence is on the betweenness centrality, so a user with a high out-degree centrality is also establishing his brokerage position in the network.

The in-degree centrality, which has positive effects on the retweeted ratio, retweet ratio and the link ratio, is the only social network property that influences content network properties. The most obvious effect is the one on the retweeted ratio, which can be explained that a user who has many followers is more likely to be retweeted, because more people are reading her tweets. The influence on the link ratio could be explained, that a user with a high followership, might have the desire to contribute more links inside this network. The same explanation can be given for the effect on the retweet ratio where the user might want to provide more tweets by simply retweeting more from other users.

There are only two effects visible in the influence network between content properties. The first one is a positive effect of the number of tweets on the link ratio. This simply means that a user who is writing a lot of tweets each day is likely to add more links to her tweets. The second one is a positive influence of the hashtag ratio on the retweet ratio. So a user who is using multiple hashtags might want to retweet more tweets of other persons.

Overall, this experiment shows us that the social "closeness" of the semantic web Twitter users is getting higher each day. The social properties are very likely to get higher the next day and individual users are getting more integrated into the network. Nonetheless, the influences between properties are very rare in the influence network. Only the in-degree centrality shows some interesting influences on other values and there are only two effects between content properties and none between content and social network properties. Nevertheless, it shows a different approach compared to the first experiment (see section 4.2)), because in this experiment, a complete sub-network of Twitter was analyzed, and the analysis showed that such partial networks can grow on their own without getting influenced by external Twitter users.

Some parallels to the first experiment (see section 4.2)) can be discovered. The in-degree value shows again the most interesting and strongest effects on the content network, which indicates that the number of followers is a very important motivation for Twitter users to add more content and also more content features. Like in the first experiment all properties influence themselves (except the favorite ratio), what indicates that users are becoming more active each day and the Twitter concept of trying to keep the users active works.

4.4 Boards.ie experiment

This experiment was performed with the Boards.ie dataset (see section 3.3. This dataset covers material from the large online board Boards.ie and in comparison to the other two datasets, it does not include Twitter data. The following list recaps the main characteristics of this dataset:

- Popular Irish bulletin board called Boards.ie[10]
- Analysis focuses on data of the year 2006
- Social network is built upon reply information
- Topic distributions form content information (total of 650 topics)

As already mentioned, this study just picks out information about the year 2006 of the dataset. This period is split into 12 monthly time steps, and each value is always calculated by information about the last six months. The corresponding dataset section (see section 3.3) explains this in a greater detail.

Like the second dataset, this dataset forms a "closed" social network, where no ties to the outside world are existing. On the one hand, the experiment tries to find influences between the two social network properties (in- and out-degree) and on the other hand, influences between the social network properties and the topic entropy of a user, calculated by the topic distributions of the user. It should be kept in mind that the topic entropy (see below for description) information itself does not constitute a real content network, but moreover, specifies relevant meta content information. Figure 4.9 shows a subset existing of 200 vertices out of the dataset and the directed social network edges between them.

The interesting thing about this sample subnetwork is, that a lot of users do not reply and a lot of other users do not reply to them at all. Nevertheless, some users are very active and respond a lot.

[10] http://www.boards.ie/

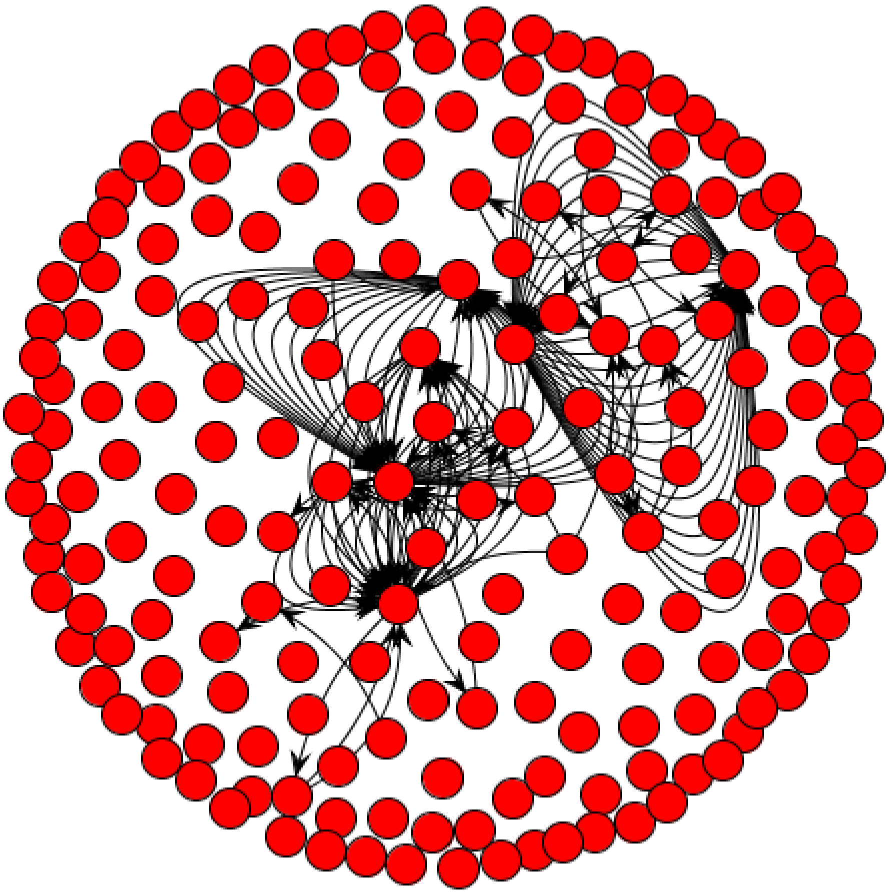


Figure 4.9: Sample of the third dataset

This analysis again is based on different social and content network properties. The main goal of this experiment is to find influences between the in- and out-degree to the topic entropy of a user. Following list shows the different social and content network properties of this experiment:

Social network properties

Each of the two social network properties is calculated with the present social network of the crawled dataset.

Out-degree

The out-degree of a user states how often the user is replying to posts by other users in the corresponding six-month period before the actual measure date. This is done by counting all hierarchical (nested) reply information, available for the user.

In-degree

The in-degree of a user states how often other users are replying to posts of the user in the corresponding six-month period before the actual date of interest. This is done by looking at the reply information and counting the number of replies to posts written by the user.

Content network properties

This analysis focuses just on one content property, which is the topic entropy.

Topic entropy

The topic entropy is defined as the Shannon entropy of the topic distribution of a user, which is available in the dataset. The topic distributions were calculated by my advisor Claudia Wagner using a *LDA model* (latent Dirichlet allocation). LDA is a generative probability model, which can be used as a topic model. The LDA model was learned on posts and the output provides topic annotations per user. The Shannon entropy was introduced by C. E. Shannon in the year 1948 [Sha48]. C. E. Shannon states that entropy plays a key role in information theory as measures of information, choice and uncertainty. It is defined as:

$$H = -\sum_{i=1}^{n} p_i * log_2 p_i \tag{4.2}$$

In equation 4.2 p represents the topic distribution vector of a person and n is the number of elements of this vector. In this analysis, we can use the entropy measure to determine, if a user is more of a generalist or a specialist. A generalist uses many different topics within the posts, whereas a specialist just focuses on a few topics. So the topic distribution of a generalist would be more equally distributed than the distribution of a specialist. If the topic distribution is more equally distributed the Shannon entropy is getting higher, whereas the entropy measure is lower, if there is a focus on just a few specific topics. This leads to the fact that a generalist will have a higher entropy than a specialist over his topic distribution.

Overall, there are topic distributions available for 29886 different users. These users form the group of seed users for this experiment. It was not always possible to calculate all corresponding properties for each user out of this group for each month, because they simply made no posts during the calculation time. If this is the case, the properties are simply zero. Because multilevel regression analyses and time series analyses can handle sparse data (see section 2.3.5) it was still possible to analyze this sparse data, and it is no problem for the regression analysis, that some users just have data for one or a few time points and are zero for the rest.

The study has been done using the multilevel regression model and methods described in section 4.1. The binding between the social and content networks is done simply by the information about the users. We know which user has which social network ties and furthermore we know which user has her content network at a specific time point. So it is easy to combine the corresponding properties. Figure 4.10 shows the resulting influence framework for the third experiment.

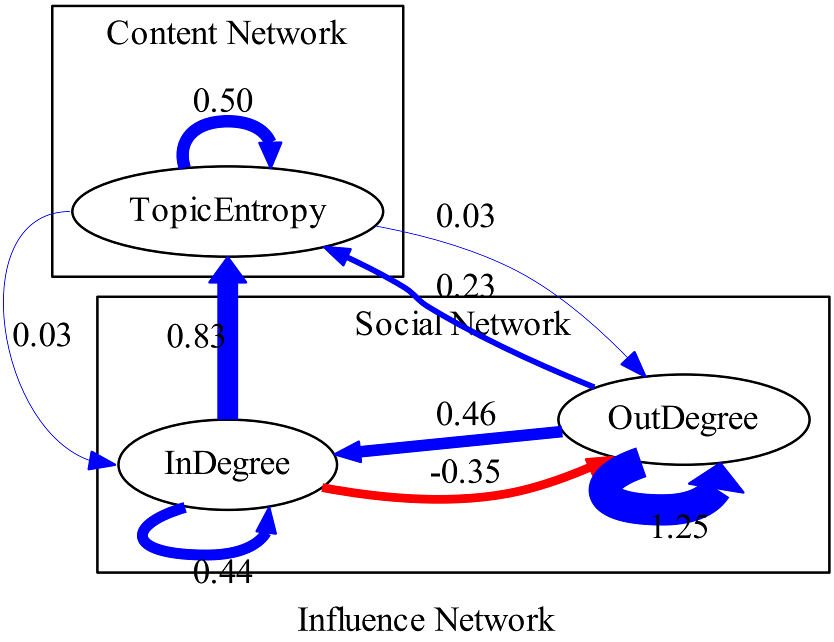


Figure 4.10: Influence network of the third dataset

The following table summarizes all the detected influences of the different network properties. The influences are represented in such a way that the properties in the first column have a corresponding influence on a value stated in the first row. This table just illustrated the different arrows shown in figure 4.10 in a clearer way.

When looking at the two social network properties (out- and in-degree) the influence network shows that both measures have a strong positive impact on themselves. The huge positive influence of the out-degree on itself means that a user is getting more active the next month and is replying to more posts. The influence of the in-degree on itself is not as strong, but still implies that more users are replying to posts made by

	Out-Degree	In-Degree	Topic entropy
Out-Degree	1.25	0.46	0.23
In-Degree	-0.35	0.44	0.83
Topic entropy	0.03	0.03	0.50

Table 4.3: Influences of the third experiment

the user. These two influences really show that the users tend to get more involved in the board and are posting more content. The positive influence of the out-degree on the in-degree is a result of this higher activity in the board. It is obvious that a user, who is providing more input to the forum by the form of making more posts, is likely to get more replies in the future because other users read more posts of the user. As a result of this, other users might reply more often to posts by the user. Interestingly the influence network states that a user with a high in-degree one month is expected to produce fewer replies the next month himself, which is shown by a negative influence of the in-degree on the out-degree. One possible explanation for this could be, that a user, who gets a lot of replies to posts made by him one month, becomes more inactive in providing replies made by himself the next month, because she becomes lazy.

The only content network property, which is the topic entropy, also shows a positive impact on itself. This suggests that a user, who provides content over several topics (general), is becoming even more general the next month. So, her posts don't focus so harshly on one or just a few topics, but moreover, she is extending her choice of the topics she is regarding in her postings. The in-degree also shows a positive influence on the topic entropy, which means that a user, who gets many replies, is extending her topic focus the next month and is becoming more general. This might be explained by the way the user gets influenced by other posts and might want to consider more of these topics in his posts. The out-degree also has a positive effect on the topic entropy, which can be explained similar to the influence of the in-degree on the topic entropy. The user replies to more posts, so she also reads more content from other people and also thinks more about these topics. So she might also include more topics the next month herself. The other direction just shows very low positive influences of the topic entropy on the in- and out-degree, which means that a user who is more general one month is also likely to post more and get more replies the next month.

Overall, this experiment shows that the users tend to get more involved into the Boards.ie forum over time, because they seem to post more replies and even get more replies by other users. It as well seems that the users are becoming more general posters, what means that they provide content over various topics and don't seem to focus just on a few topics. This also might be the fact, because the board covers a lot of different sections for distinct topics and so the user might write about different topics as well. Maybe these results would be different, when looking at a board, which has a higher focus on just a few topics and specialists are covering these topics. This experiment should anyway stand on its own, and it is difficult to compare it with the two other experiments, which cover Twitter data and focus on other aspects. It is anyway interesting to see that users of this board seem to get involved into several new topics each month and are expanding their horizons.

5 Conclusions

The main goal of this work was to study the bi-direction influence between social and content networks over time. This was done by applying a framework using a multilevel regression model introduced by Wang and Groth [WG10]. To study the influences in social networks three different datasets were used to provide distinct insights. The first two datasets are based on Twitter, a microblogging service. These datasets differ in their crawling strategy and network structure. On the one hand, the first dataset is a subscript of seed users taken from the public timeline, and it contains all social relations inside the subnetwork and also all links to the outside Twitter world. On the other hand, the second dataset consists of a group of seed users of a public user list by Stephano Bertolo representing Twitteres about the topic semantic web and this dataset only contains relations between the group of seed users. The analysis on both datasets shows that there exist many influences between social and content network properties, but not vice versa. This indicates that the social parts of a user in Twitter have a high impact on content features in the future. Furthermore, both results suggest that the number of followers of a user has the highest influence on other properties. The final observation that could be made, is that nearly all properties on both analysis have a positive influence on themselves, indicating that they will become even higher the next day and there are overall just very few negative effects to identify. The third and last dataset used in this thesis is based on a large public Irish bulletin board called Boards.ie. The analysis on this dataset was used to discover influences between the number of posts by a user and the number of replies by other users on the topic distribution representing the content part. The results indicate that the users of the forum are becoming more involved into the forum over time and also seem to extend their repertoire of different topics used throughout their posts.

5.1 Implications

The results presented in this work can provide a better insight to the influences in social networks for social analysis. The results indicate that the users in the corresponding networks represented by Twitter and Boards.ie tend to get more active during time. Furthermore, the number of followers in Twitter and the number of replies to posts made by the user in Boards.ie show the most positive influences to different content network properties. This suggests that the attention of other users is motivating users the most.

The analysis on Boards.ie shows that the users tend to become more general throughout their choice of different topics. This could be based by the fact, that Boards.ie provides a high number of different sub-forums covering a lot of different topics. For the hosts of such forums this could imply, that they have to be careful in designing the structure of their forum. If the goal is to have a very broad spectrum of topics, then there might be a lot of different sub-forums, but if the goal is to be a very specific forum, then there might just be a limited number of sub-forums.

5.2 Threats to validity and limitations

The three different experiments could answer the objectives stated in section 1.2. Nevertheless, it would be interesting to see if different datasets with the same crawling strategies would provide same results. It would be necessary to gather a good amount of different datasets or extend the datasets to a larger audience. Furthermore, the results on the Twitter datasets should be compared to datasets of other social networking sites like Facebook, Google+ or Del.icio.us. A big question is also, if the used multilevel regression model fits the best for the different analysis. It would be a good idea to extend the statistical approach to further techniques and compare the results to have better insight to the suitability of such methods.

The following list covers some possible threats to the validity of the experiments (based on [Pac04]:

- Lack of reliability on the independent variable due to variability: This is throughout the datasets not the regular case, because the values do not vary very much from one occasion to another. The properties are all based on information of social media. A sample case where this threat could happen would be if a person decides to delete all her followees on Twitter and then again decides to add them again. In this case the variable would vary.
- Lack of representativeness of the independent variable: The possibility exists, that some of the independent variables do not cover the construct of interest the best way. For example it could be that the number of used hashtags per day is not representative for the theoretical construct we want to analyze. Anyway, it is very likely that this is not the case in this work, because we could determine a lot of influences throughout the analyses. To clearly be sure it would be necessary to remove and add properties and compare the achieved results.
- Lack of impact of the independent variable: It is possible that a treatment does not produce a realistic impact on the seed users of the datasets in this work.

- Lack of reliability on the dependent variable: This can happen if the variation in measurement is too large. For this threat the same point as stated for the threat of a lack of reliability on the independent variable is the case. So it is very unlikely that this is a real threat for this work.
- Lack of representativeness of the dependent variable: The content validity has to be adequate. In this work the dependent variable always has an adequate content validity, because only properties were chosen, that were suitable for the representative content.
- Lack of sensitivity of the dependent variable: The measure has to be sensitive enough so that differences in the outcome can be detected. This seems to be the case in this work, because the differences can be detected.
- Nonrepresentative sampling: This could be very well the case. It is possibility that the seed users of the dataset do not represent the whole network very well. Another possibility could be that the sampling is too small, so it would be necesseray to expand the dataset (see section 5.3).
- Inappropriate use of statistical techniques: Maybe the used regression model is not the best to determine influences in the corresponding datasets. It would be useful to try alternations of the model and use different techniques (see section 5.3).
- Carryover effects: If one of the above mentioned threats occur, it is possible that the effects carry over, because this study uses repeated measures.

5.3 Future Work

This work should also provide motivation for further investigations. A first future aspect could be to apply the framework to further and maybe even larger datasets to get better insights. Based on these new datasets it would also be a good idea to extend the social and content network properties. As a result of this it would also be possible to compare the determined results better and have more possibilities to find implications. A further step to this work could be done by altering the used regression model. For example this could be done by condition the regression constant on person. Based on these new results a comparison to the results in this work could be done to decide which model provides the more plausible results. Furthermore it would be useful to explore in addition to regression models also other techniques to identify influences and then compare the results. A final idea is to develop a public user interface to give other people the possibility to analyze their own data.

Appendix

Details about the first Dataset

This appendix covers some more details about the first dataset.

A.1 Seed Users

The starting number of seed users for this dataset was 1.500. Table A.1 shows 20 sample seed users of this dataset. The complete userlist can be on this website[1]. It is possible, that some users may have changed their screen name in the meantime, but the userid of each user is unique.

UserID	Screen_Name
83515859	bizwingman
107203267	csnne1
116151006	wuyongshan
70260577	mitaLOVEEE
194379706	domo265
15809964	amyurbach
35379588	penguiiinn
71715193	BishopBronner
162763616	ShesTHATbreezy
36122815	TimmytheGiant
87205530	WasteAwareFood
95550902	PricelessLee_
22498558	bnb_tweets
65581669	fyratopia
227813434	Pretty_Ling
87721782	BarcardiiTee
22508290	kateyy___
60324533	Hauke_Borow
177917146	rizkysatriady
161170268	dwynnnn

Table A.1: 20 sample seed users of the first dataset

[1] http://www.student.tugraz.at/p.singer/dipl/seedusers.htm

A.2 Crawled users each day

Table A.2 shows the crawling process timespan of the first dataset. For each day the number of distinct crawled users is displayed.

Date	Crawled users	Note
15.03.2011	1459	
16.03.2011	1457	
17.03.2011	1365	Temporal API problems
18.03.2011	1456	
19.03.2011	1448	
20.03.2011	1443	
21.03.2011	1447	
22.03.2011	1447	
23.03.2011	1448	
24.03.2011	1436	
25.03.2011	1434	
26.03.2011	1430	
27.03.2011	1432	
28.03.2011	1437	
29.03.2011	1432	
30.03.2011	1436	
31.03.2011	1431	
01.04.2011	1433	
02.04.2011	1435	
03.04.2011	1435	
04.04.2011	1431	
05.04.2011	1427	
06.04.2011	1429	
07.04.2011	1426	
08.04.2011	1426	
09.04.2011	1422	
10.04.2011	1421	
11.04.2011	1421	
12.04.2011	1421	
13.04.2011	1414	

Table A.2: Crawl dates and crawled users for the first dataset

Details about the second Dataset

This appendix covers some more details about the second dataset.

B.3 Seed Users

The starting number of seed users for this dataset was 137. Table B.3 shows 20 sample seed users of this dataset. The complete userlist can be found on this website[2]. It is possible, that some users may have changed their screen name in the meantime, but the userid of each user is unique.

UserID	Screen_Name
7431072	yokofakun
14250157	andraz
14825144	LgComputer
14265466	francoisbry
17018622	robeng
16179709	witbrock
6637672	gromgull
21716274	mijopo
24332965	piellemme
7112242	novaspivack
13904112	zemanta
18277845	AnneJHunt
16639872	marko_grobelnik
17303350	semanticaweb
15916003	trueknowledge
14760739	evri
14130714	opencalais
14080345	Powerset
780290	PaulMiller
15729365	chipmasters

Table B.3: 20 sample seed users of the second dataset

[2] http://www.student.tugraz.at/p.singer/dipl/seedusers2.htm

B.4 Crawled users each day

Table B.4 shows the crawling process timespan of the second dataset. For each day the number of distinct crawled users is displayed.

Date	Crawled users	Note
26.04.2011	134	
27.04.2011	134	
28.04.2011	134	
29.04.2011	134	
30.04.2011	134	
01.05.2011	134	
02.05.2011	134	
03.05.2011	134	
04.05.2011	134	
05.05.2011	134	
06.05.2011	134	
07.05.2011	134	
08.05.2011	134	
09.05.2011	134	
10.05.2011	134	
11.05.2011	134	
12.05.2011	134	
13.05.2011	134	
14.05.2011	134	
15.05.2011	134	
16.05.2011	134	
17.05.2011	134	
18.05.2011	134	
19.05.2011	134	
20.05.2011	134	
21.05.2011	134	
22.05.2011	134	
23.05.2011	134	
24.05.2011	134	
25.05.2011	134	

Table B.4: Crawl dates and crawled users for the second dataset

Details about the third Dataset

This appendix covers some more details about the third dataset.

C.5 Seed Users

Table C.5 shows 20 sample seed users of this dataset. For the corresponding users, topic distributions are calculated, and these users form the network for the analysis of the third dataset.

UserID
51982
19695
2707
21128
53812
61425
61071
61057
60708
60723
60745
60769
60788
60830
8450
60575
60576
9729
60404
60358

Table C.5: 20 sample seed users of the third dataset

C.6 Users each month

Table C.6 shows the number of distinct users, for which information about the topic distribution are available.

Date	Distinct users
01.01.2006	8429
01.02.2006	8234
01.03.2006	8542
01.04.2006	8579
01.05.2006	8833
01.06.2006	8659
01.07.2006	8887
01.08.2006	9189
01.09.2006	8634
01.10.2006	8687
01.11.2006	8910
01.12.2006	8370

Table C.6: Dates and distinct seed users each month of the third dataset

Bibliography

[AKM08] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In Ying Li, Bing Liu, and Sunita Sarawagi, editors, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 7–15. ACM, 2008.

[Ale11] Alexa. Top sites. `http://www.alexa.com/topsites`, 2011.

[BB08] John Breslin and Ulids Bojars. Boards.ie sioc data competition. `http://data.sioc-project.org/`, 2008.

[BGH+08] Dominik Benz, Marko Grobelnik, Andreas Hotho, Robert Jäschke, Dunja Mladenic, Vito D. P. Servedio, Sergej Sizov, and Martin Szomszor. Analyzing tag semantics across collaborative tagging systems. In Harith Alani, Steffen Staab, and Gerd Stumme, editors, *Proceedings of the Dagstuhl Seminar on Social Web Communities*, 2008.

[BGL10] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS*, pages 1–10. IEEE Computer Society, 2010.

[BH07] Dominik Benz and Andreas Hotho. Position paper: Ontology learning from folksonomies. In Alexander Hinneburg, editor, *Workshop Proceedings of Lernen - Wissensentdeckung - Adaptivität (LWA 2007)*, pages 109–112. Martin-Luther-Universität Halle-Wittenberg, September 2007. http://lwa07.informatik.uni-halle.de/kdml07/kdml07.htm.

[BHS10] Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In *Proceedings of the 2nd Web Science Conference (WebSci10)*, Raleigh, NC, USA, 2010.

[BKH+11] Dominik Benz, Christian Körner, Andreas Hotho, Gerd Stumme, and Markus Strohmaier. One tag to bind them all : Measuring term abstractness in social metadata. In Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Jeff Pan, and Pieter De Leenheer, editors, *Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011)*, Heraklion, Crete, May 2011.

[BMB11] Douglas Bates, Martin Maechler, and Ben Bolker. Package lme4. `http://cran.r-project.org/web/packages/lme4/lme4.pdf`, 2011.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[CBHS08] Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic analysis of tag similarity measures in collaborative tagging systems. In *Proceedings of the 3rd Workshop on Ontology Learning and Population (OLP3)*, pages 39–43, Patras, Greece, July 2008. ISBN 978-960-89282-6-8.

[CHBG10] Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, and P. Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. In William W. Cohen and Samuel Gosling, editors, *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010.* The AAAI Press, 2010.

[CMG09] Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 721–730, New York, NY, USA, 2009. ACM.

[Die05] Reinhard Diestel. *Graph Theory.* Springer-Verlag Heidelberg, 2005.

[DKM+06] Micah Dubinko, Ravi Kumar, Joseph Magnani, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Visualizing tags over time. In *Proceedings of the 15th international conference on World Wide Web(WWW '06)*, pages 193–202, New York, NY, USA, 2006. ACM.

[DuV10] Adam DuVander. Twitter reveals: 75day). `http://blog.programmableweb.com/2010/04/15/twitter-reveals-75-of-our-traffic-is-via-api-3-billion-calls-per-day/`, April 2010.

[EK10] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* Cambridge University Press, 2010.

[Gel08] Andrew Gelman. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15):2865–2873, July 2008.

[GH05] Scott A. Golder and Bernardo A. Huberman. The structure of collaborative tagging systems. *CoRR*, abs/cs/0508082, 2005. informal publication.

[GH06] Scott A. Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32:198–208, April 2006.

[GOJ$^+$02] Kwang-Il Goh, Eulsik Oh, Hawoong Jeong, Byungnam Kahng, and Doochul Kim. Classification of scale-free networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12583–12588, October 2002.

[Got09] John M. Gottman. *Time-Series Analysis: A Comprehensive Introduction for Social Scientists.* Cambridge University Press, 2009.

[Hay06] Andrew F Hayes. A primer on multilevel modeling. *Human Communication Research*, 32(4):385–410, 2006.

[HJSS06a] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In *The Semantic Web: Research and Applications, volume 4011 of LNAI*, pages 411–426. Springer, 2006.

[HJSS06b] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Trend detection in folksonomies. In Yannis S. Avrithis, Yiannis Kompatsiaris, Steffen Staab, and Noel E. O'Connor, editors, *Proc. First International Conference on Semantics And Digital Media Technology (SAMT)*, volume 4306 of *LNCS*, pages 56–70, Heidelberg, December 2006. Springer.

[HRS07] Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy, editors, *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 211–220. ACM, 2007.

[HRW08] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *CoRR*, abs/0812.1045, 2008.

[HST$^+$11] Denis Helic, Markus Strohmaier, Christoph Trattner, Markus Muhr, and Kristina Lerman. Pragmatic evaluation of folksonomies. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 417–426, New York, NY, USA, 2011. ACM.

[HTE10] Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, pages 173–178, New York, NY, USA, 2010. ACM.

[HTSA10] Denis Helic, Christoph Trattner, Markus Strohmaier, and Keith Andrews. On the navigability of social tagging systems. In Ahmed K. Elmagarmid and Divyakant Agrawal, editors, *Proceedings of the 2010 IEEE Second*

International Conference on Social Computing, SocialCom / IEEE International Conference on Privacy, Security, Risk and Trust, PASSAT 2010, Minneapolis, Minnesota, USA, August 20-22, 2010, pages 161–168. IEEE Computer Society, 2010.

[JC97] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, cmp-lg/9709008, 1997. informal publication.

[KBH$^+$10] Christian Körner, Dominik Benz, Andreas Hotho, Markus Strohmaier, and Gerd Stumme. Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th international conference on World wide web*, pages 521–530, New York, NY, USA, 2010. ACM.

[Kit10] Genshiro Kitagawa. *Introduction to Time Series Modeling (Chapman & Hall/CRC Monographs on Statistics & Applied Probability).* Chapman and Hall/CRC, 2010.

[KKGS10] Christian Körner, Roman Kern, Hans Peter Grahsl, and Markus Strohmaier. Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In *21st ACM SIGWEB Conference on Hypertext and Hypermedia (HT 2010)*, Toronto, Canada, June 2010. ACM.

[KKS10] Roman Kern, Christian Korner, and Markus Strohmaier. Exploring the influence of tagging motivation on tagging behavior. In *Proceedings of the 14th European conference on Research and advanced technology for digital libraries*, ECDL'10, pages 461–465, Berlin, Heidelberg, 2010. Springer-Verlag.

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA, 2010. ACM.

[KNT06] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, New York, NY, USA, 2006. ACM.

[Mas10] Mashable. Twitter surpasses 145 million registered users. `http://mashable.com/2010/09/03/twitter-registered-users-2/`, September 2010.

[MC03] Peter R. Monge and Noshir Contractor. *Theories of Communication Networks.* Oxford University Press, USA, 2003.

[MCM+09] Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme. Evaluating similarity measures for emergent semantics of social tagging. In Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl, editors, *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 641–650. ACM, 2009.

[ME87] P. R. Monge and E. M. Eisenberg. Emergent communication networks. In F. M. Jablin, L. L. Putnam, K. H. Roberts, and L. W. Porter, editors, *Handbook of organizational communication*, pages 304–342. Sage, Newbury Park, CA, 1987.

[Mes07] Chris Messina. Groups for twitter; or a proposal for twitter tag channels. `http://factoryjoe.com/blog/2007/08/25/groups-for-twitter-or-a-proposal-for-twitter-tag-channels/`, 2007.

[MHHD07] Knud Möller, Tom Heath, Siegfried Handschuh, and John Domingue. Recipes for semantic web dog food: the eswc and iswc metadata projects. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ISWC'07/ASWC'07, pages 802–815, Berlin, Heidelberg, 2007. Springer-Verlag.

[Mil67] Stanley Milgram. The small-world problem. *Psychology Today*, 1(1):61–67, 1967.

[MJK08] Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci. *Introduction to Time Series Analysis and Forecasting (Wiley Series in Probability and Statistics)*. Wiley, 2008.

[MM06] George Macgregor and Emma McCulloch. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, 2006.

[MSLC01] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.*, 27(1):415–444, 2001.

[New03] M. E. J. Newman. The structure and function of complex networks. *SIAM REVIEW*, 45:167–256, 2003.

[Pac04] Martin Packer. Overview of threats to the validity of research findings. `http://www.mathcs.duq.edu/~packer/Courses/Psy624/Validity.html`, 2004.

[PM91] Z. Pan and J. M. McLeod. Multilevel analysis in mass communication research. *Communication Research*, 18:140–173, 1991.

[Row11] Matthew Rowe. Forecasting audience increase on youtube. In *User Profile Data on the Social Semantic Web Workshop, Extended Semantic Web Conference*, 2011.

[RP91] L. D. Ritchie and V. Price. Of matters micro and macro: Special issue for communication research. *Communication Research*, 18:133–139, 1991.

[RPD98] John O. Rawlings, Sastry G. Pantula, and David A. Dickey. *Applied Regression Analysis: A Research Tool (Springer Texts in Statistics)*. Springer, 1998.

[Sch06] P. Schmitz. Inducing ontology from flickr tags. In *Proceedings of the 15th International Conference on World Wide Web (WWW)*, Edinburgh, UK, 2006.

[Sco00] John P. Scott. *Social Network Analysis: A Handbook*. SAGE Publications, January 2000.

[Sha48] Claude E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27:379–423, July 1948.

[SHPC10] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In Ahmed K. Elmagarmid and Divyakant Agrawal, editors, *Proceedings of the 2010 IEEE Second International Conference on Social Computing, SocialCom / IEEE International Conference on Privacy, Security, Risk and Trust, PASSAT 2010, Minneapolis, Minnesota, USA, August 20-22, 2010*, pages 177–184. IEEE Computer Society, 2010.

[SKC10] D. A. Shamma, L. Kennedy, and E. F. Churchill. Tweetgeist: Can the Twitter Timeline Reveal the Structure of Broadcast Events? *CSCW Horizons*, 2010.

[SKK10] Markus Strohmaier, Christian Körner, and Roman Kern. Why do users tag? detecting users' motivation for tagging in social tagging systems. In *International AAAI Conference on Weblogs and Social Media (ICWSM2010)*, Washington, DC, USA, May 2010.

[SRH04] Anders Skrondal and Sophia Rabe-Hesketh. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman and Hall/CRC, 2004.

[SS06] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics)*. Springer, 2006.

[TCM+08] Emma Tonkin, Edward M. Corrado, Heather Lea Moulaison, Margaret E. I. Kipp, Andrea Resmini, Heather Pfeiffer, and Qiping Zhang. Collaborative and social tagging networks. In *Ariadne*, January 2008.

[TM69] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

[TRM11] Jaime Teevan, Daniel Ramage, and Merredith R. Morris. #TwitterSearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 35–44, New York, NY, USA, 2011. ACM.

[Twi11a] Twitter. Twitter frequently asked questions. `http://support.twitter.com/articles/13920-frequently-asked-questions`, 2011.

[Twi11b] Twitter. What are @replies and mentions? `http://support.twitter.com/entries/14023-what-are-replies-and-mentions`, 2011.

[Twi11c] Twitter. What is retweet? (rt). `http://support.twitter.com/articles/77606-what-is-retweet-rt`, 2011.

[WF94] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press, 1994.

[WG10] Shenghui Wang and Paul Groth. Measuring the dynamic bi-directional influence between content and social networks. In Peter Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Pan, Ian Horrocks, and Birte Glimm, editors, *The Semantic Web ISWC 2010*, volume 6496 of *Lecture Notes in Computer Science*, pages 814–829. Springer Berlin / Heidelberg, 2010.

[WLJH10] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In Brian D. Davison, Torsten Suel, Nick Craswell, and Bing Liu, editors, *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*, pages 261–270. ACM, 2010.

[WS98] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[WS10] C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proc. of the Semantic Search 2010 Workshop (SemSearch2010)*, april 2010.