The Influence of Social Capital on Information Diffusion in Twitter's Interest-Based Social Networks

D I S S E R T A T I O N
of the University of St.Gallen,
School of Management,
Economics, Law, Social Sciences
and International Affairs
to obtain the title of
Doctor of Philosophy in Management

submitted by

Thomas Plotkowiak

from

Germany

Approved on the application of

Prof. Dr. Katarina Stanoevska-Slabeva

and

Prof. Dr. Miriam Meckel

Dissertation no. 4231

Difo-Druck GmbH, Bamberg 2014

The University of St. Gallen, School of Management, Economics, Law, Social Sciences and International Affairs hereby consents to the printing of the present dissertation, without hereby expressing any opinion on the views herein expressed.

St. Gallen, October 21, 2013

The President:

Prof. Dr. Thomas Bieger

Abstract

Recent years have seen an increase in the use of interest-based social networks such as Pinterest, ChimeIn and Twitter. While classic online social networks like Facebook or LinkedIn offer a social utility that deepens social connectivity with our existing social graphs, in interest-based social networks people form social ties based on their interest with people that they don't already know. The information diffusion in such interest-based social networks has only been marginally researched. Twitter offers a promising natural laboratory to study such networks, because the social ties, interactions and exchanged information in this network are public. This offers the possibility to study the influence of social ties and thematic interest on information diffusion in such interest-based communities. In this work, the theory of social capital was employed to study this subject. The theory of social capital explains why individuals receive retweets in networks as a result of their social ties. The operationalization of the social capital concept for Twitter uses the publicly available data traces of Twitter users. Social ties express the structural and relational dimension of social capital, whereas the interest of a user expresses his cognitive dimension. By studying the retweet traces in 166 interest-based networks, this study finds that in Twitter the structural and cognitive dimension of social capital positively influence information diffusion. This influence has been studied on an individual and group level for bonding and bridging social capital resulting in four perspectives on the phenomenon. The study of the influence of individual bonding social capital on information diffusion shows that highly central individuals who share a group's interests receive more retweets from members of their own interest group. Individuals that build up a high individual bridging social capital by positioning themselves structurally between interest groups and who express a high number of different interests, receive a high number of retweets from members of other interest groups than their own. In the group bonding social capital perspective the work shows that the internal group structure in terms of clustering, density and short paths between the actors of the group positively influences the information exchange among group members. However, a high reciprocity of social ties does not positively influence information exchange between group members. Interest groups that are located in the center of the Twitter ecosystem and whose members have a high number of diverse interests possess a high group bridging social capital. This helps them obtain many retweets from other groups. Studying the interdependencies between bonding and bridging social capital showed that when groups primarily build up group bonding social capital, they decrease their chances of receiving retweets from other groups. However, groups that build a high bridging social capital do not harm their internal information exchange. At the individual level, the study showed that individuals with a high bonding social capital receive less retweets from people outside the group and that individuals with a high bridging social capital receive less retweets from people within the group (i.e., group members).

Zusammenfassung

In den letzten Jahren verbreiten sich interessensbasierte Netzwerke wie Pinterest, ChimeIn aber auch nach wie vor Twitter zunehmend. Sie unterscheiden sich im Gegensatz zu klassischen online sozialen Netzwerken wie Facebook oder LinkedIn dadurch, dass einander unbekannte Personen aufgrund ähnlicher Interessen Beziehungen untereinander aufbauen. Dadurch entstehen Interessengemeinschaften. Die Informationsdiffusion in solchen Gemeinschaften ist bisher weitgehend unerforscht. Twitter bietet in diesem Kontext ein großes Potenzial, da hier die sozialen Beziehungen, Interaktionen und die ausgetauschte Information öffentlich sind. Dies bietet die Möglichkeit den Einfluss von sozialen Beziehungen und thematischen Interesse auf die Informationsdiffusion in solchen Gemeinschaften zu untersuchen. Hierzu wurde die Theorie des sozialen Kapitals eingesetzt. Sie erklärt warum Personen in diesen Netzwerken aufgrund ihrer sozialer Beziehungen Retweets erhalten. Die Operationalisierung von Sozialkapital für Twitter wurde auf der Basis von öffentlich verfügbaren Nutzerdaten mit Hilfe der sozialen Netzwerkanalyse durchgeführt. Soziale Verbindungen stellten die strukturelle und relationale Dimension des Sozialkapitals dar, während das thematische Interesse des Nutzers die kognitive Dimension widerspiegelte. Durch die Untersuchung von 166 Interessensgruppen konnte gezeigt werden, dass sich vor allem die strukturelle und kognitive Dimension des Sozialkapitals positiv auf die Informationsdiffusion auswirkten. Dieser Einfluss wurde sowohl auf der individuellen als auch auf der Gruppenebene nachgewiesen und für beide Typen des sozialen Kapitals (bonding und bridging) aufgezeigt. Die Resultate beim individuellen bonding Sozialkapital zeigten, dass zentrale Akteure, die das gleiche Interesse der Gruppe teilen, auch mehr Retweets von Gruppenmitgliedern erhielten. Individuen, die ein hohes bridging Sozialkapital aufbauen d.h. sich strukturell zwischen Interessensgruppen positionieren und eine hohe Anzahl von Interessen aufzeigen, erhielten entsprechend mehr Retweets von Mitgliedern außerhalb ihrer Gruppe. Die Gruppensicht des bonding Sozialkapitals zeigt, dass ein hohes Clustering, eine hohe Dichte sowie kurze Wege im Interaktionsnetzwerk der Gruppenmitglieder den Austausch von Information innerhalb der Gruppe begünstigten, aber Reziprozität sozialer Beziehungen nicht förderlich war. Interessensgruppen, die ein hohes bridging Sozialkapital auf Gruppenebene aufbauen, indem sie sich im Zentrum der Twittersphäre verorten und aus Mitgliedern bestanden, welche eine hohe Anzahl an verschiedenen Interessen besitzen, konnten auf diese Weise viele Retweets von anderen Gruppen erhalten. In der Untersuchung der Interdependenzen zwischen bonding und bridging Sozialkapital konnte diese Studie zeigen, dass ein hohes bonding Sozialkapital von Gruppen einen negativen Einfluss auf den Erhalt von Retweets von anderen Gruppen hatte. Andererseits zeigte sich kein negativer Einfluss von hohem bridging Sozialkapital von Gruppen auf den internen Informationsaustausch. Auf der individuellen Ebene konnte gezeigt werden, dass ein hohes bonding Sozialkapital die Chancen auf Retweets von Mitgliedern anderer Gruppen reduziert und dass ein hohes bridging Sozialkapital einen negativen Einfluss auf den Erhalt von Retweets von Mitgliedern aus der eigenen Gruppe hat.

Acknowledgment

This dissertation would not have existed without the contribution of several people. First of all, my Ph.D. supervisor, Prof. Dr. Katarina Stanoevska-Slabeva helped me throughout the process with insightful comments and a strategic direction. Also, the discussions with my co-supervisor, Prof. Dr. Miriam Meckel, had a crucial role in setting the direction and objectives of my study.

I also received great help from colleagues who over time also became very good friends. I would like to thank Jana Ebermann, Friederike Hoffmann, Anne Suphan, Jana Baumgartner, Tobias Heinisch, Michael Etter, Thomas Wozniak and Roland Pfister.

Finally I want to thank my family members for the confidence and energy they invested in me. My parents did not only enable me to receive a good education, but also encouraged me to constantly explore the world. My younger sister Margarethe, who finished her thesis before me and showed me that persistence and hard work pays off. My biggest gratitude goes to Jana who stood by me sharing the joy and pain of conducting research and writing a Ph.D. thesis.

St. Gallen, December 2013

Thomas Plotkowiak

Contents

List of Figures

List of Tables

List of abbreviations

API	Application programming interface
AT	@-Interaction
CMC	Computer mediated communication
DV	Dependent variable
H	Hypothesis
ID	Identification
IRC	Internet relay chat
IV	Independent variable
IT	Information technology
FF	Friend and follower
K-S	Kolmogorov–Smirnov
LDA	Latent dirichlet allocation
MDS	Multi dimensional scaling
MMOG	Massively multiplayer online game
SQL	Structured query language
RT	Retweet
RQ	Research question
SMS	Short message service
SNA	Social network analysis
URL	Unique resource locator
VIF	Variance inflation factor
ZPRED	Standardized predicted values
ZRESID	Standardized residuals

1 Motivation

Social media has emerged as the most important source of information as well as the most important distribution channel of information (Newman, 2009; Shirky, 2011). Whereas in the beginning social media focused either on a specific type of content (e.g., videos on YouTube) or on the management of relationships (e.g., Facebook, LinkedIn), at present they are evolving into comprehensive ecosystems where users spend most of their time managing their contacts, getting connected, creating and sharing content, playing, working and coordinating common activities as well as getting and providing information. More than 75% of people[1] find news online or get it through email or social networking sites, and 65% of adult Internet users now say that they use a social networking site[2]. Already in 2009, a substantial part of the information and content exchanged on these sites related to daily news, political events and business topics (Analytics, 2009), showing that communicating information in social media has indeed matured beyond its infancy. This revolution has had major implications on how we share information, replacing traditional media with social media as the main channel of distribution. The next sections will lay out these dramatic shifts.

Social media vs. traditional media

Just a few years ago, the biggest barrier for individuals or corporations wanting to disperse information were the high costs of the technical infrastructure needed in order to reach a mass audience. Newspapers, radio and television dominated the dissemination of information (Maletzke, 1963). Today, this bottleneck has been completely removed thanks to widespread access to the Internet and the ease of use of social media. Online social networks have seen a true boom in the sharing and dissemination of information. Facebook has, at present, more than 1 billion[3] active users, who interact with over 1.2 billion objects such as pages, groups, events and community pages. More than 1 billion pieces of content (e.g., web links, news stories, blog posts, notes, photo albums, etc.) are being shared every day. Twitter (among other[4] notable microblogging services) is the second most used social media site and provides

[1] http://stateofthemedia.org/2012/mobile-devices-and-news-consumption-some-good-signs-for-journalism/what-facebook-and-twitter-mean-for-news/

[2] http://pewinternet.org/Commentary/2012/March/Pew-Internet-Social-Networking-full-detail.aspx

[3] http://www.facebook.com/press/info.php?statistics

[4] http://www.tumblr.com, http://www.jaiku.com or http://posterous.com

similarly impressive statistics: over 200 million[5] active Twitter users produce over 340 million[6] Tweets per day and the increasing tendency for mobile phone use has made Twitter the "SMS of the Internet" (Smith and Brenner, 2012). Its popularity is high not only among private users, but also among mass media channels, politicians, journalists or governmental institutions (Wu et al., 2011).It is also becoming widely popular in work or enterprise environments[7] (Zhao and Rosson, 2009). These developments are permanently changing the way in which information is produced, distributed and consumed (Newman, 2009).

Active sharing in a networked distribution process

The process of information dissemination has changed from passive consuming to active sharing: users of social media sites are creating vast networks of friends with whom they share information and messages on a yet unseen magnitude. People are no longer silently consuming media, but instead are interacting with it. This means that they are commenting and discussing information items while forwarding the most interesting ones to their friends. Public figures are increasingly creating networks and disseminating their messages directly to the public without using media outlets as mediators (Newman, 2009). Users themselves have developed important networked sources of information and news (Newman, 2009; Shirky, 2011) where "the witnesses are taking over the news" (Jarvis, 2008) and taking dissemination into their own hands. The new culture of sharing has profoundly changed the linear process of information diffusion, replacing centralized distribution channels with personal networks, where information is distributed directly from user to user. In Twitter, this has led to the creation of an ecosystem of users who work together as a highly networked organism. This, in turn, has led to a decidedly non-linear information diffusion process that is influenced by the network structure of individuals themselves, rather than by the effective reach of the medium.

Social networks vs. interest-based social networks

The emerging networks in social media are in some cases based on existing personal networks (e.g., Facebook or LinkedIn) but in other cases they are based on the users' interests. These "interest-based social networks" or "interest graphs"[8] have a markedly different focus and approach than social networks. Networks such as Pinterest[9], Thumb[10], Foodspotting[11],

[5] http://mashable.com/2012/12/18/twitter-200-million-active-users/
[6] http://blog.twitter.com/2012/03/twitter-turns-six.html
[7] http://www.socialcast.com, http://www.yammer.com or http://www.jivesoftware.com
[8] http://advertising.twitter.com/2012/08/interest-targeting-broaden-your-reach.html
[9] http://pinterest.com
[10] http://itunes.apple.com/us/app/thumb/id368595692?mt=8
[11] http://www.foodspotting.com

Fitocracy[12], Formspring[13] or, Chime.in, and especially Twitter, enable users to focus and organize first and foremost around their interest, whereas traditional social networks such as Facebook or LinkedIn focus on a user's personal relationships. Facebook offers a social utility that deepens social connectivity with our existing social graphs, while these new interest-based social networks enable users to express their interests in "new, engaging ways and offer authentic, high value connectivity with new people we don't already know"[14]. In 2011, Twitter itself branded its social network with the slogan: "Follow your interests". Researchers have found that this is one of the key motivations for Twitter use. Indeed, users engage with each other in Twitter because they like "sharing information with interesting people"[p.4], (Java et al., 2009), "raising the visibility of interesting things and gathering useful information"[p.1] (Zhao and Rosson, 2009) or subscribing to a "people-based RSS feed"[p.7] (Böhringer, 2009), and (Weng et al., 2010) noted that users follow each other based on shared interests and a following relationship is a strong indicator of similar interests among users.

Interest-based social networks might seem distinct from existing social networks, yet they are not a completely new phenomenon. Rather, they are simply a logical successor of previous definitions and conceptions of online communities and share the underpinnings of a) social relations between members and b) a common shared cognitive element (Etzioni and Etzioni, 1999) between members. Indeed, Etzioni and Etzioni (1999) defined online communities as being comprised of two attributes: "First it is a web of [...] relationships that encompasses a group of individuals - relationships that crisscross and reinforce one another rather than simply a chain of one-on-one relationships. Second a community requires a measure of commitment to a set of shared values, mores, meanings and a shared historical identity - in short a culture."[p.241]. Looking at communities of practice, Wasko and Faraj (2005) find the same ingredients noting that "communities of practice [are] self-organizing, open activity systems focused on a shared practice that exists primarily through computer-mediated communication"[p.37]. Using an anthropological perspective, Howard Rheingold has captured the same essence under the term of virtual communities (1993) a term which illustrates the social network component of the virtual society: "my seven year old daughter knows that her father congregates with a family of invisible friends who seem to gather in his computer. Sometimes he talks to them, even if nobody else can see them. And she knows that these invisible friends sometimes show up in the flesh, materializing from the next block or the other side of the world."[p.1]. The study of such virtual communities was later described by Berry Wellman as "the study of online social networks" (1997). Rheingold later mentioned that had he read Barry Wellman's work earlier, he would have called his book "Online Social Networks". One of the main characteristics of these emerging online social communities is mentioned by Pipa Norris as being that "online communities indeed bring together like-minded souls who share particular beliefs hobbies or interests"[p.37] (2002).

[12] http://www.fitocracy.com

[13] http://www.formspring.me/

[14] http://techcrunch.com/2012/02/18/beyond-facebook-the-rise-of-interest-based-social-networks/

1.1 Problem definition and research goals

For decades, the notion of community has been studied under a wide variety of names. With this in mind, it becomes clear that the study of Twitter's "interest-based social networks" are very much in line with the study of these communities. Apart from the long research tradition pertaining to online communities, the main reasons to study interest-based social networks today can be summarized into four main points. First, in Twitter the user's main motivation to join and interact with others in such network communities is grounded in sharing a particular interest (Java et al., 2009; Zhao and Rosson, 2009; Böhringer, 2009; Weng et al., 2010). This fact makes it particularly interesting to explore how the user's interest affects information diffusion in such networks. Second, whereas up until now various online communities have been analyzed as isolated research subjects (e.g., see above (Garton et al., 1997; Rheingold, 1993; Etzioni and Etzioni, 1999)), Twitter has created an ecosystem that allows multiple communities to simultaneously co-exist and interact with each other, making it possible for the first time to observe interactions between entire groups or communities. Third, in this ecosystem of a new form of loosely coupled communities emerges, where members do not explicitly have to join a community but gradually become a part of one or multiple such interest-based social networks. Fourth, the social ties in those online communities that have previously been hidden are now becoming apparent. This means that communities can be observed and analyzed from a network-based perspective.

In other words, people who share the same interests are now becoming aware of each other and are starting to organize themselves into numerous interest-based online communities (Norris, 2002). Researchers have shown that the internet facilitates new connections in that it provides people with an alternative way to connect with others who share their interests or relational goals (Ellison et al., 2011; Horrigan et al., 2001). In Twitter, a universe of multiple co-existing online communities of unseen size and diversity has started to emerge. As an example figure 1.1 shows how we can imagine interest-based social networks on Twitter: In this case three different interest groups have emerged around the sports interests of cycling, running and swimming. And while in this case the majority of users form online relations around only one interest, we also find users, who are interested in all three disciplines (e.g., triathletes) and users who exhibit an interest in two disciplines (e.g., duathlon athletes).

Thus whereas formerly people were partly unaware of others who shared the same interests, now entire networks of people who share the same interest are starting to crop up. And if ties in such networks are determined by the individuals' interest, then the structure of such interest networks might also be determining how information is diffusing in those networks. While a number of factors such as a user's demographics or behavior have been studied extensively (see review of factors determining information diffusion in chapter 4), little is known about the manner in which the structure of such interest-based social networks influences information diffusion in social media. The following citation is representative of the many efforts information systems researchers have made to demand a closer investigation of the information diffusion process: "it is [...] important to explain why individuals select to share

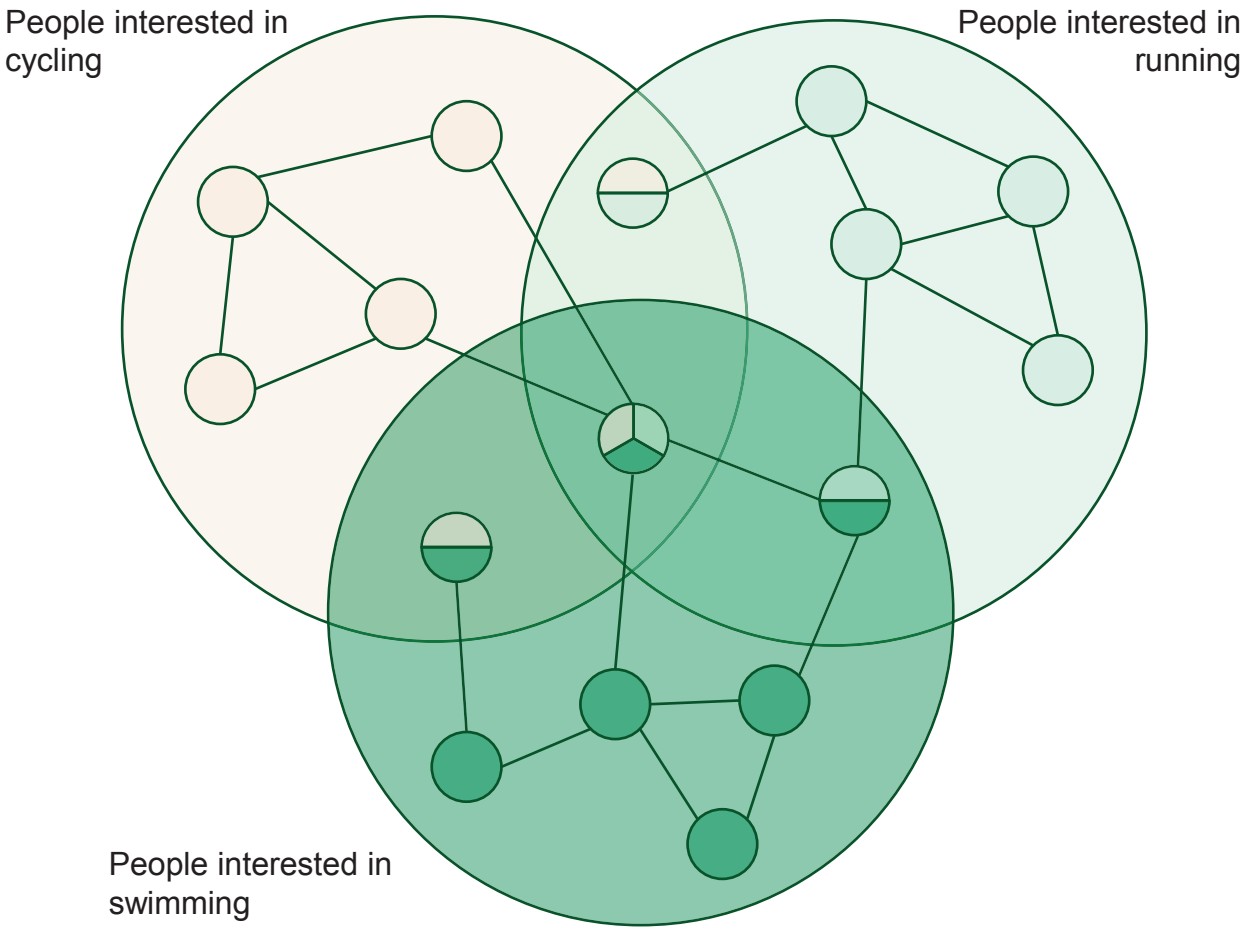


Figure 1.1: Example of interest-based networks for swimming, running and cycling. Color represents interest, edges social ties.

or not to share information and knowledge with other community members when they have a choice. Identifying the motivations underlying the knowledge sharing behavior in online communities would help both academics and practitioners gain insights into how to stimulate knowledge sharing in online communities"[p.1873] (Chiu et al., 2006). Given the magnitude and emerging popularity of such interest-based social networks (see above) and the urge to explore this process, we therefore have to study the underlying structure of the emerging interest-based networks and understand how this structure impacts the current information exchange process. Twitter provides us with the "promising natural laboratory"[p.1] (Bakshy et al., 2011) to do so.

The study of such network structures and how they lead to certain goals has been known for decades (see historical development of the concept of social capital in chapter 2) under the theory of social capital. Social capital is a concept that seeks to explain how the structure of social networks can help individuals or groups obtain certain goals or goods: persons first build up social capital by embedding themselves in social networks, and then use this social capital to obtain the desired goal or good, which in real life could be help, money or services. Over the last decades, researchers (see online social capital in chapter 2) have agreed that social capital does not only apply to traditional real-life social networks but that it can also be created online, and that the goals and outcomes do not actually have to be tangible goods, but can also be virtual goods or services. By following these new developments, this study will explore how the social capital of Twitter helps individuals and groups obtain retweets from other people in their interest network, thus helping them better spread their information online. We will see that using the social capital concept to explain information diffusion has certain benefits in comparison to the competing concepts that have traditionally been used (see chapter 3). These benefits can be summarized as the following: a) researchers have already attempted to measure online social capital, b) social capital is compatible with models of interpersonal information diffusion (see chapter 3), b) social capital can be based on a networked structure of the population (see chapter 2), c) social capital allows researchers to investigate the role of groups and individuals in the diffusion process and finally d) it is operationalizable with high[15] quantities of network- and information-diffusion-data (see chapter 5). Therefore the overall research question of this thesis is:

Research Question: How does online social capital influence information diffusion in Twitter's interest-based social networks?

This question consists of three building blocks: The a) interest-based social networks, b) the social capital concept and c) its effect on the diffused information. Given the public nature of the Twitter network, the information diffused in social networks is easy to operationalize

[15] As part of the sampling process a total of 830 448 Twitter lists with 1.71 Mio. unique Twitter users were reviewed. The final raw data set consisted of over 16 thousand people, their 46 231 808 tweets, and a total of 613 404 616 retweets or so-called diffusion traces (see 6.4 in chapter 6).

through the retweets that people receive from other people. However, the social capital concept for Twitter needs to be operationalized first, and although social capital has been studied from many different perspectives (see perspectives on social capital chapter 2), and researchers have attempted to measure social capital in online environments such as Facebook (Ellison et al., 2007; Steinfield et al., 2008) the literature lacks a network-based operationalization for social networks such as Twitter (see section 3.4). This lack of operationalization stands in stark contrast with the research possibilities Twitter allows: Twitter not only facilitates the study of online social capital that people acquire, but also the study of information diffusion traces left behind by the users. Indeed, unlike other user-declared networks, Twitter is "expressly devoted to disseminating information in that users subscribe to broadcasts of other users"[p.3] (Gupta et al., 2011). The publicly visible ties between Twitter users not only allow the precise mapping of underlying social networks (including the measure of tie strength and direction between Twitter users), but also allows the tracking of each piece of information traveling through the network, without violating privacy concerns since all this information is public. Despite the abundance of behavioral traces that this network offers, Twitter lacks a social capital framework that uses these traces. Therefore, the first operational goal will be to provide a theoretical concept of network-based online social capital for Twitter. The second goal will be to operationalize this framework by providing indicators for the different dimensions of social capital on Twitter. The final goal will be to operationalize interest-based communities on Twitter (see figure 1.2) in which the social capital is exhibited.

This operationalization will allow to apply the reasoning on the multi-dimensional concept of (online) social capital to test hypotheses on its influence on information diffusion: the theory of social capital (see chapter 2) allows to break down the research question into four types of questions (see table 1.1 below). Answering these questions will allow us to observe the effects of social capital on a group network level and an individual network level (Borgatti et al., 1998): At the *individual* level this work will study the social capital of individuals, and at the *group* level, it will study the social capital of an entire group. Additionally, this thesis will also study different *types* of individual social capital and group social capital. This distinction also stems from the research on social capital that distinguishes two kinds of social capital: *bonding* or *bridging* social capital. Both types of social capital can potentially help the individual and group better spread their information. The basic premise of each of these four questions is that the realization of social capital has a positive influence on information diffusion. The resulting questions concerning the influence of social capital on information diffusion are presented in table 1.1, forming four quadrants.

The operationalization of the multi-dimensional concept of social capital (see chapter 4) will allow us to investigate *which* social capital factors in each quadrant lead to information diffusion: research question one will answer which factors of group bonding social capital of a group leads to more information diffusion among all members of this group. Research question two will investigate which factors of individual bonding social capital allow individuals to receive more retweets from members of their interest group. Research question three will determine which factors of group bridging social capital allow groups to obtain more retweets from other

	Bonding	Bridging
Group	**RQ1:** How does group bonding social capital influence the overall diffusion of tweets inside the group?	**RQ3:** How does group bridging social capital influence the diffusion of the group's tweets to other groups?
Individual	**RQ2:** How does a person's individual bonding social capital influence the diffusion of their tweets to other members of the group?	**RQ4:** How does a person's individual bridging social capital influence the diffusion of their tweets to members of other groups?

Table 1.1: Overview of the resulting research questions

groups. And finally, research question four will investigate which factors of individual bridging social capital allow individuals to receive more retweets from members from other interest groups. In order to highlight the relevance of these questions in practice, the next paragraph will provide examples of different domains in which they are important and show how these questions apply to these specific domains.

1.2 Problem relevance

Problem relevance in practice

Domains like politics, news, journalism and marketing are all affected by the dramatic shifts that social media brings (Kwak et al., 2010; Tumasjan et al., 2010; Jansen et al., 2009): In all of these domains within Twitter, interest-based social networks are emerging. May it be interest networks related to a political party (e.g., democrats vs. liberals), a certain crisis (e.g., Iranian uprisings, Syrian conflicts, Haiti earthquake), a certain brand or product (e.g., Apple or iPhone) or simply a certain profession or hobby (e.g., programming, knitting or swimming). People follow their interests in Twitter and so embed themselves in networks in accordance to their interest. Thus not only does the internal structure of those networks decide on how information is exchanged in Twitter, but also the network of interest-based networks decides where information is diffused in Twitter. Therefore the next paragraph shows how answering the research questions can contribute to the understanding of similar information diffusion processes that are of relevance in a number of different domains.

Politics in social media: Ensuring that political information reaches a wide non-fragmented audience on the Internet has been a major research issue (Sunstein, 2002) for decades. The main argument behind this is that the Internet tends to only attract participants with identical interests, thus prohibiting the heterogeneous dissemination of information. However, discussions of this topic have always been divergent: while some researchers (Ackland, 2005; Adamic and Glance, 2005) showed that homogeneous online networks of political parties were emerging, others argued that there was a significant lack of support for the clustering of ideologically-related blogs and websites (Cornfield et al., 2005). If we answer the four research questions of this thesis, we will be able to generate answers to relevant questions that are at the intersection of social media and politics. In order to do so, we can think of political parties (e.g., "Democrats") as the unit of analysis on the group level, and political candidates (e.g., "Barack Obama") as the unit of analysis on the individual level. Moreover, when candidates create networks with members of their own party, they are building up bonding social capital; when they connect to members of other parties, they are building up bridging social capital.

By answering the first research question (RQ1) we can analyze the process of political parties "flocking together"[p.417] (McPherson et al., 2001) and may discover out which factors of group bonding social capital political parties should build up in order to better spread information to their members. By answering RQ2 we might find out which factors of individual bonding social capital might help politicians to better disseminate their message to members of their own party, which is often referred to as "preaching to the converted"[p.1] (Norris, 2003). By answering RQ3 we might find out which factors of group bridging social capital help political groups better disseminate their information or ideas to other groups and thus dominate the political discourse. By analyzing the overall position of the political group in interest networks, we might also expect to find out what role politics actually plays in the Twitter ecosystem. Is it unimportant and in the periphery of the network, or is it at the very core of users' interests? Finally, by answering RQ4 we might find out which factors of individual bridging social capital might help politicians transfer information or ideas from one political party to the other, thus reaching members of the opposite party.

Journalism in social media: The dramatic impact that social media has had on the way news is distributed has also created a shift in journalistic practices, leading to a convergence (Singer, 2004b,a; Sunstein, 2009) of social media and classical journalism. The newly-emerging social media sources (e.g., supporters of the Iranian revolution, or simply tech enthusiasts tweeting live from the Apple keynote), journalists, news corporations, and their readers, have become entangled in a complicated networked medial ecosystem (Bowman, S. And Willis, 2003). Some of the most interesting questions regarding this new ecosystem deal with explaining how information actually flows in this network and finding ways for sources, journalists and readers to benefit from this new environment. We can consider interest groups (e.g., supporters of the Iranian revolution, or Apple fans, or a collection of journalists) as the unit of analysis on the group level, and individual Twitter accounts as the unit of analysis on the individual level. When people of a special interest group (e.g., Iranian revolution) flock together with others who have similar interests, they create bonding social capital, and when individuals

(e.g., journalists) become part of such interest groups, they build up bridging social capital since they are connecting readers and sources.

Answering RQ1 might help us determine which factors of group bonding social capital drives information sharing among group members. Why is it that in some cases such as the "Iranian social media revolution" information was circulating quickly among members, while in other cases it is not - the groups bonding social capital might reveal answers to this question. By answering RQ3, we might understand which factors of individual bonding social capital help members of such groups better disseminate their message to members of the interest group. Indeed, which factors determine that some tech evangelists[16] or Iranian protesters[17] are highly retweeted (Weng et al., 2010) by their peers while others are not? By answering RQ3, we might discover which factors of group bridging social capital help entire groups of journalists and news corporations better disseminate their information to other groups, in other words, their readers. Finally, by answering RQ4, we might find out which factors of individual bridging social capital help journalists transfer information from the social media sources to their readers. This is a process often referred to as gatekeeping.

Marketing in social media: Brands, marketers and advertisers are being continuously challenged to develop strategies and processes to effectively connect with consumers. For decades, consumers have been segmented according to their demographics or milieus[18] (Sinus Sociovision GmbH, 2010) which led to scatter data that could only be interpreted statistically. Only recently have advertising companies[19] discovered that online users actually have specific interests, which, in turn, has led to the emergence of social targeting[20]. Advertisers and brands have also discovered the high value of brand communities that form around their product. Researchers (Bauer and Grether, 2005) have found that these "communities that provide social capital are a key instrument to establish satisfaction, trust and commitment within a business relationship"[p.91]. Accordingly tie creation with users from the same brand community (group) sharing similar interests can be seen as creating bonding capital and tie creation with users from different brand communities can be seen as creating bridging social capital.

By answering RQ1, we might figure out which factors of group bonding social capital are necessary to facilitate information diffusion among members of brand communities. Such information is highly valuable for the community managers of brand communities since the more members are willing to share information, the more vital and valuable this community is for a brand (be it to obtain feedback on the product or to use this free diffusion in order to better promote the product). By answering RQ2, we might find out which factors of individual bonding social capital are necessary for commercial Twitter accounts to obtain more retweets from their interest group. Answering RQ3, would help determine which factors of group bridging social capital might help humanitarian organizations obtain retweets from other

[16] e.g., https://twitter.com/Scobleizer
[17] e.g., https://twitter.com/oxfordgirl
[18] http://www.tdwi.com
[19] e.g., http://www.adtelligence.de
[20] http://de.wikipedia.org/wiki/Social_Targeting

groups, and in doing so, promote awareness about an issue or raise donations for a good cause. Finally, by answering RQ4 we might discover which factors of individual bridging social capital help commercial Twitter accounts to obtain retweets from group members that are not their main target group.

Knowledge exchange in social media: The research questions can also provide interesting findings for new generations of Twitter-based knowledge exchange systems[21], which are increasingly transforming the way in which people communicate within corporations. We might expect that knowledge exchange in such systems mirrors the preexisting communication channels within corporations. Indeed, people meet in the hallways or around the water-cooler and tend to "assemble around a topic of common interest"[p.478] (Henri and Pudelko, 2003) that dominates the way information disseminates and improves organizational performance (Millen et al., 2002). Indeed, there is a long line of information systems studies (see chapter 2) that use the concepts of social capital to explain how these factors affect knowledge exchange in organizations (Wasko and Faraj, 2005). Here we can think of people as the individual unit of analysis and departments as the group unit of analysis. The social ties between people from the same department create bonding social capital, while the social ties between people from different departments create bridging social capital.

By answering RQ1, we can determine which factors of group bonding social capital lead to better information diffusion among members of the same department and thus to better team collaboration. By answering RQ2, we can show which factors of individual bonding social capital managers need to create to better spread information between the colleagues of their own department. By answering RQ3, we might be able to determine which factors of group bridging social capital can help certain departments (e.g., Management or HR) better spread their information to other departments of the company. Finally, by answering RQ4, we can show which factors of bridging social capital should be fostered among employees to help exchange information between different divisions.

These examples show that although the nature of the domains presented are very different, the analysis of emerging interest-based social networks has the potential to create fundamental insights into the interplay of social capital and information diffusion inside and between groups.

Problem relevance in academia

The research questions also contribute to existing academic research on social capital and information diffusion. In the current literature (see chapter 3), there are three main knowledge gaps to which this research can contribute: a) research on information diffusion inside and between communities in Twitter is lacking, b) there is no common framework uniting the fragmented literature on information diffusion in Twitter, and finally, c) there is a lack of knowledge concerning the influence of people's interests on information diffusion in Twitter.

[21] e.g., Yammer

Contribution to information diffusion inside and between communities: While there is a wide field of research (see chapter 3) that generally describes interpersonal information diffusion (Bass, 1969; Katz and Powell, 1955; Rogers, 1995; Valente, 1995) and that highlights the theoretical importance of groups (Rogers, 1995), norms (Wejnert, 2002) and individuals and their underlying networks (Bass, 1969; Gladwell, 2000; Katz and Powell, 1955; Lazarsfeld et al., 1965; Merton, 1957; Schenk, 1993; Valente, 1993; Weimann, 1994), there is a lack of empirical studies that investigate multiple groups simultaneously. Up until the present day the community level of such networks is not fully understood (Valente, 2010). Historically, the main problem in analyzing information diffusion at the level of the community was the enormously high effort required to collect the group network data — a problem that came with the questionnaire-based design (Scott, 2000): Collecting data for a single group took a matter of months or years. These collection problems went along with the lack of reliable information diffusion data (Huberman et al., 2009; Trusov et al., 2010) that allowed researchers only to infer diffusion from the actual underlying social structure. Finally an additional problem hindering the research of groups and diffusion processes was the segmentation of the research community into theorist approaches (Monge and Contractor, 2003), which were mentioning the need for such concepts, and data driven approaches, which tended to neglect much of the theoretical work (Kossinets et al., 2008).

Contribution to a theoretical framework of information diffusion in Twitter: Regarding the research on information diffusion on Twitter, numerous academic contributions have appeared (see chapter 3) highlighting a) the importance of underlying networks' structure (Bakshy et al., 2011; Galuba et al., 2010), b) the influence of individuals (Gayo-Avello, 2010) c) their behavior (Romero and Kleinberg, 2010) or the influence of the d) message (Cha et al., 2010; Suh et al., 2010) itself in this process. Although the results of such studies are enormously important, there seems to be no common framework or perspective unifying the different contributions in such a way that would help connect these isolated observations. The social capital framework for Twitter might help to provide a clearer view on the fragmented body of literature. The motivation for such an approach is rooted in the observation of the information systems research on knowledge exchange, which addresses very similar research to the ones on Twitter information diffusion, yet here researchers managed to integrate their work under the common perspective of social capital.

Contribution to investigation of users' interests on information diffusion: On one hand there is a wide body of knowledge (see chapter 3.4) explaining how social media users create networks. In a nutshell, the creation of ties can be based on a) real-life friendships or contacts (e.g., Facebook, LinkedIn) or b) a variety of network mechanisms (Golder and Yardi, 2010) (e.g., closing triads) or c) user's demographics (Takhteyev, Y, Gruzd, A, Wellman, 2010; Yardi and Boyd, 2010; Brzozowski and Romero, 2011; Conover et al., 2011; Weng et al., 2010; Wu et al., 2011). The different notions of homophily that describe tie creation among similar persons in social electronic networks has been partly explored (Choudhury et al., 2010; McPherson et al., 2001). Yet up to today there is only anecdotal research on how the user's personal interests influence information diffusion in such networks: Choudhury et al.

(2010; 2010) find that the choice of a homophilious attribute can impact the prediction of information diffusion, but also note that this homophily differs across topics and that more research is needed. Wu et al. (2011) have analyzed different user interests finding strong attention homophily but they only analyzed four relatively small interest-based networks. They also recommend the exploration of opinion leaders in a range of different topics. Romero et al. (2011) have analyzed differences in the information diffusion across topics, but focused explicitly on the adoption of hashtags using only 8 users' interests and did not cover information diffusion between different interest networks. Finally a number of authors (Counts, Scott, Pal, 2011; Hong and Davison, 2010; Ramage et al., 2010; Tunkelang, 2009; Weng et al., 2010; Zhao et al., 2011) focused on classifying users or tweets according to a set of given topics, but these works either focused on the accuracy of the classification or on the ranking of the users, but did not focus on the implications for the information diffusion process.

This work will seek to improve on the these three gaps in academic research, in three ways: First in a conceptual way by creating a network-based online social capital framework that works on a group and individual level (see chapter 4) and conceptualizes the structural, relational and cognitive dimensions of social capital by using the behavioral data traces of online social networks such as Twitter. Secondly providing network-based measures for this social capital framework and by highlighting which dimensions of social capital are driving information diffusion in Twitter. Thirdly by using the social capital framework to offer insights on the influence of users' interests on information diffusion on both a group and individual level.

1.3 Outlook

This thesis is divided into eight chapters and structured accordingly (see figure 1.2): Chapter 2 will review the literature on social capital and in particular online social capital in order to define the theoretical concepts of social capital that will be used for its operationalization for Twitter. In order to be able to generate meaningful hypotheses on the influence of social capital on information diffusion, chapter 3 will review different theories on information diffusion with a focus on existing results of information diffusion studies on Twitter in the background of social capital. In chapter 4 the operationalization of social capital for Twitter will be developed, by introducing a framework of network-based online social capital for Twitter. Combining insights from the literature review of information diffusion in chapter 3 and the social capital framework for Twitter in chapter 4, the hypothesis and measures in chapter 5 will break down the research questions into four classes of hypotheses regarding their influence on information diffusion in Twitter. For each hypothesis it will also introduce the measures of social capital. Chapter 6 will explain how the underlying data of interest-based communities was collected, by describing the process of determining a representative set users' interests on Twitter and then showing how the sampling of users' interest networks has been performed. Chapter 7 will present and discuss the empirical results and thereby provide answers to the hypotheses

on the influence of social capital on information diffusion. Finally chapter 8 will present the conclusions of this work.

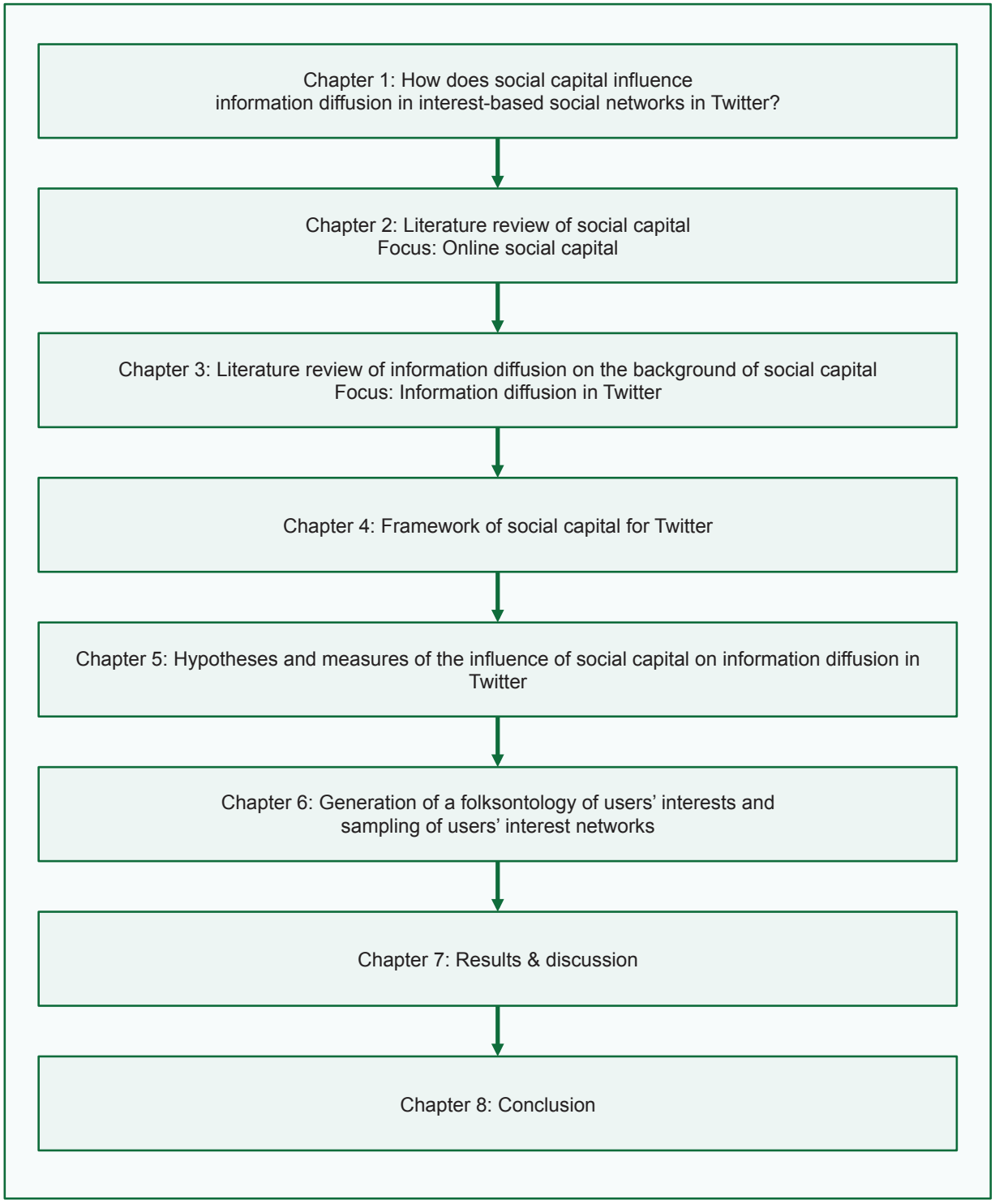


Figure 1.2: A flow chart overview of the thesis outline.

2 Literature review of social capital

This literature review is structured according to the guidelines of Webster & Watson (2002) and Levy & Ellis (2006) and has two goals, which are covered in two consecutive chapters: the first goal is to review the existing fundamental theories on social capital and online social capital in order to build up the theoretical foundation necessary to the creation of an online social capital framework for Twitter. The second goal is to review diffusion theories and existing studies on information diffusion in Twitter, in terms of their contribution to the structural, relational and cognitive dimensions of social capital. This will provide the theoretical basis for the creation of hypotheses concerning the influence of social capital on information diffusion in Twitter.

The framework used in both following chapters follows a funnel approach (Webster and Watson, 2002): the first part of each review will provide an overview of the field and the second part consists of a systematic literature review that focuses on the literature that is most relevant to the research question. A schematic overview of each chapter is depicted in table 2.1 and table 2.2. The systematic literature review is divided into three major stages: "1) inputs (literature gathering and screening), 2) processing, and 3) outputs"[p.1] (Levy and Ellis, 2006). In other words, each systematic literature review will first present the keywords that were used and the databases that were searched, and then will describe the insights gained from the remaining relevant studies. Each systematic review will be concluded when "new articles only introduce familiar arguments, methodologies, findings, authors and studies"[p.192] (Levy and Ellis, 2006) and "new concepts in the article set cannot be found"[p.16] (Webster and Watson, 2002). Finally, each chapter will provide a conclusion.

Literature review on social capital

The goal of chapter 2 is to review the various dimensions and perspectives of social capital theory and find studies that help operationalize this concept for online social networks such as Twitter by using literature from leading peer-reviewed journals. This literature will be retrieved using a multi-keyword search (the keyword search details can be found in table 2.4 in section 2.2.3) on ProQuest[1]. The risk of missing relevant material is small since the social capital theory and construct is very stable (Robey et al., 2000) in the literature. The review will follow a funnel or narrow-down approach (see table 2.1): indeed, the broad theory of social capital will be narrowed down in such a way that the chapter will provide an overview of the concept

[1] `http://search.proquest.com`, which is the leading literature database according to (Levy and Ellis, 2006) and lists the majority of IS world's top 50 journals

and then focus more specifically on studies dealing with online and virtual social capital. This means that the social capital concept will first be analyzed in terms of the various prevalent perspectives and dimensions found in the literature. The review will then focus on the structural and cultural perspectives of social capital, and its measurement will be discussed. Following this, the review will focus on studies that provide insights into online or virtual social capital and their measurement. The final part of this review will highlight studies that measured online social capital using only data-driven network-based measures.

Following the funnel approach, the conclusion of this review will show that a) there is a strong theoretical background for the structural perspective of social capital, b) that there are several existing models that capture the multidimensionality of this concept, c) that these models have been used successfully online, and, finally, d) that the operationalization of such models for Twitter does not yet exist.

Social Capital Theory		
Perspectives	Structural View	Cultural View
	Bridging & Individual	Bonding & Group
Main Authors	Burt, Lin, Borgatti, Nahapiet	Hanifan, Putnam, Coleman, Bourdieu
Focus	Online Social Capital (134) Virtual Social Capital (54) Twitter Social Capital (2)	
Main Authors	William, Wellison, Smith, Ellison	
Conclusion & Goals		

Table 2.1: Overview of the funnel view on literature review of social capital. Sources of systematic literature review in brackets. Grey shaded area shows the focused literature.

Literature review on diffusion

The goal of chapter 3 is to review the main diffusion theories and the existing studies on information diffusion in Twitter, in terms of their contribution to the structural, relational and cognitive dimensions of social capital. These theories will be reviewed in terms of the insights they provide on information diffusion in networked systems and also in terms of their potential to help form hypotheses about information diffusion.

Following the funnel approach (see table 2.2), a systematic literature review will be conducted on the existing research relevant to information diffusion on Twitter. This section will provide an in-depth review of aspects explaining information diffusion and knowledge exchange in Twitter. The review will also be performed by using a multi-keyword search (Keyword details can be found in table 3.2 in section 3.4) on ProQuest. The systematic review process is especially applicable to the research of information systems such as Twitter because the literature on this topic is diverse and interdisciplinary. The remaining relevant literature will be categorized according to its contribution to the structural, relational and cognitive perspectives of social capital.

Information Diffusion			
Dimensions	Cognitive	Structural	Relational
	Influence of group norms	Opinion leaders and Two-Step flow	Strength of ties
Main Authors	Rogers, Mentzel, Bass	Lazarsfeld, Merton, Schenk, Valente, Weimann	Granovetter, Valente
Focus	Information Twitter (283) Diffusion in Twitter (21) Knowledge Exchange Twitter (3)		
Main Authors	Kwak, Romero, Choudhury	Galuba, Wu, Bakshy	Yang, Cha, Hong
Conclusion & Goals			

Table 2.2: Overview of the funnel view on literature review of information diffusion. Sources of systematic literature review in brackets. Grey shaded area shows the focused literature.

The conclusion of this review will show that there is a) a strong theoretical background on information diffusion which can be used to create hypotheses concerning the influence of social capital on information diffusion. It will also show that b) information diffusion inside and across homophilious groups has not yet been studied extensively. It will also highlight that c) the numerous attempts to explain information diffusion on Twitter do not follow a common theoretical framework. Finally, the review will show that there is d) a lack of information diffusion observations on a group level and that e) despite a myriad of academic studies on Twitter, there are only three studies that explore interest-based diffusion in Twitter.

2.1 Definition of social capital

Ever since the start of research into social capital, social scientists, political scientists and economists have been interested in the value of social relations, that is the so-called social capital. Social capital is becoming a core concept in business, political science, public health and sociology (Moore et al., 2005; Williams, 2006). Although the term social capital today still faces criticism (Huber, 2009), the popularity of the social capital concept is undaunted, due to the fact that it promises to explain outcomes based on social structure alone (Kadushin, 2004). The various resulting perspectives and definitions in the discipline explain why there are so many controversial discussions: in the literature, we find a variety of different definitions and measurements for the term "social capital" (Adler and Kwon, 2002). Empirical results often come to contradictory conclusions (Franzen and Pointner, 2007) and social capital is often conceived of as both a cause and an effect (Resnick et al., 2000; Williams, 2006). Therefore, the goal of the following literature review is to first provide a general overview of social capital (according to seminal reviews of social capital by Lin (2002), Adler and Kwon (2002), Borgatti (2003) and Williams (2006)) and reveal the multiple dimensions of this concept and its measures. The main section of this chapter will then focus on the studies related to the measurement of social capital, especially in the form of online and virtual social capital. The conclusion will highlight the resulting findings.

Historically, social capital has often been defined according to a structural (Burt, 1995; Lin, 2002) and cultural dimension (Bourdieu, 1986; Coleman, 1988; Putnam, 1995). Whereas the structural dimension includes the notion of networks between people, the cultural dimension is related to the values, norms and attitudes of people. To obtain a better understanding of the definitions, a brief review of the literature will be provided in order to highlight the various ways social capital has been defined historically[2].

A widely accepted definition of social capital according to Lin (2002) describes how the social resources that are available to an individual in his network influence his success at reaching his goals[3]. Social capital usually consists of three basic ingredients: (1) resources that are bound in the network, (2) access to these resources through the network and (3) the appropriate use of these resources, according to a formalized definition of social capital by Lin (2002). Social capital can be seen as an analogue to human or financial capital, only with the distinction that here individuals are fostering relations and connections instead of other goods. In comparison to financial or human resources, social resources are goods such as social status, favors, collaborations or access to information. The access to these resources depends on if the individual has a connection to a person that owns or offers this resource, and on where this person is located in his individual network. In a communication network in which information is the basic resource, social capital explains how the social network or its structure allows or

[2]For an in depth historical review of the original works on social capital, the review Social Capital Theory - Towards a methodological foundation by Julia Häuberer (2010) is recommended

[3]for a current list of definitions of social capital, see `http://www.socialcapitalresearch.com/definition.html`

denies access to the informational resources in this network (Burt, 1995; Garton et al., 1997). It also explains how people can strategically use their position in order to push information forward to others within the network.

2.1.1 The cultural view of social capital

The earliest (Borgatti et al., 1998) mentions of social capital go all the way back to Judson Hanifan (1920) and appear later on in the work of Jane Jacobs (1961) and Ulf Hannerz (1971), who use this term in a more contemporary way. However, the term "social capital" grew in popularity with the works of Bourdieu (1986), who defines social capital as an aggregate of resources which are "linked to the possession of a durable network of more or less institutionalized relationships of mutual acquaintance and recognition"[p.248]. These resources can be used by actors for their own purposes, transforming social capital in other forms of capital. Coleman (1988) later defines social capital as "a variety of entities with two elements in common: they all consist of some aspect of social structure, and they facilitate certain actions of actors within the structure"[p.302]. Thus, according to Coleman's definition, social capital is anything that enables individual or collective action, generated by networks of relationships, reciprocity, trust and social norms. Putnam (1995) later develops his concept of social capital following the tradition of Coleman and Bourdieu. He defines social capital as networks of engagement, trust and reciprocity and highlights its social outcomes. Putnam focused early on the importance relationships that had so far been neglected by Bourdieu and Coleman's early definition of social capital. Building on Granovetter's research (1973) on tie-strength, Putnam distinguishes between two types of social capital: social capital that is created in closed (bonding) structures, and bridging social capital that emerges in open structures, when people are connected to people outside of their own groups and enable the flow of information between their group and the outside group. This form of social capital is closely related to information diffusion, specifically the diffusion of non-redundant information across networks. Individuals connected by weak ties to people outside the individual's network are more likely to provide new information such as a job opening in another company. Bonding social capital can be found in close-knit clubs (e.g., yacht clubs or tennis clubs) and primarily serves the group by reinforcing its homogeneous nature. Bonding social capital relates to the benefits associated with one's closest friends, including the social, emotional and tangible support they provide, and therefore favors the diffusion of information within the group. Studies adopting a cultural perspective on social capital mainly investigate the shared cultural norms of groups such as families, communities, immigrants or even gangs and focus on measuring the effect of this cultural social capital on outcomes such as academic achievement, education or volunteering. Social capital in these areas is often confused with concepts of social classes, i.e., the poor vs. the rich, or the elite vs. the people. Finally, a stream of research on cultural social capital also highlights "the dark side of social capital" (Adler and Kwon, 2002; Gargiulo and Benassi, 1997), which describes how social bonds can be an obstacle to pursuit economic interests. Here, social capital has also been proven responsible for negative outcomes such as

susceptibility to infection (Cohen et al., 1997), weak teamwork performance (Hansen, 2002), or over-embeddedness leading to inertia and provincialism (Gargiulo and Benassi, 1997).

2.1.2 The structural view of social capital

Portes (1998) is among the first researchers to present a structural, network-oriented perspective of social capital, highlighting social networks as a core concept of social capital. Nan Lin (2002; 1999) further conceptualizes social capital as a structural concept and focuses on the structure of social ties between individuals. Lin's concept represents a formalized theory of social capital including axioms and theorems. He defines social capital as an "investment in social relations with expected returns in the marketplace"[p.19]. In order to produce profits, individuals mutually network and interact. The social ties between people facilitate the flow of information, and some social ties carry more valued information than others. This means that in an imperfect market situation social ties can provide opportunities and choices to individuals that are otherwise unavailable. Social ties also serve as an individual's social credentials: they reflect the individual's accessibility to resources through social networks and relations. Moreover, according to Lin, social ties are a reinforcement of an individual's membership to a social group.

Ronald Burt conceptualizes social capital in the structural theory of action (1995; 2000): a person's goal is to maximize their utility while having their own set of resources at hand. The position of an actor in social structure determines the calculation of his utility and models his interest. Burt's view is similar to Lin's: he believes that the social capital of an actor is expressed by his relationships. Burt describes relationships by their range, density and multiplexity: the range measures the diversity of the actor's contacts, the density measures the connectedness of the network and the multiplexity measures the variety of connections that a given actor has. Actors can use their relationships to gather specific resources such as information. Accordingly, actors build relations where useful parts of information arrive, so that they can benefit early on from access to information and quickly forward it to others. Burt shows that social capital can be generated in two different ways: it can be generated from a *brokerage position* in the network, where the individual becomes a *broker* spanning "structural holes" between unconnected actors in the network, or from a *bonding position*, where the individual becomes an *opinion leader* in the group. For brokers, the resulting competitive advantage of spanning structural holes is timely access to information. Access to information is crucial because it does not flow evenly throughout the network. Timing is relevant because the earlier one receives information, the better the competitive advantage. For opinion leaders, the competitive advantage results from the resulting referrals: these help opinion leaders become individuals whose "conversations make innovations contagious"[p.11] (Burt, 1999). The referrals received disperse information about the individual himself, thus proving his authenticity and credibility.

Following Burt's tradition, other researchers (Adler and Kwon, 2002) have further explored the

effects of social capital on business outcomes such as (product) innovation, entrepreneurship, pioneering, company performance and on individual outcomes such as promotions (Brass, 1995; Burt, 1995) or finding a job. Another area of social capital research can be found in the organizational and medical environment (Mehra et al., 2001), where the outcomes of social capital on child-, adult- and senior-health related issues (Berkman and Syme, 1979) such as mental health, well-being or HIV risks are measured. The third main area is rooted in organizational research, where the influence of social capital on the implementation of new strategies, organizational learning or support of new policies is explored (Ahuja, 2000).

2.1.3 Multi-dimensional view of social capital

Rather than consider social capital from the dualist standpoint, with the cultural dimension opposing the structural dimension, Nahapiet and Ghoshal (1998) introduce a widely popular (Wasko and Faraj, 2005; Chiu et al., 2006; Lin, 2011) three-dimensional model (see figure 2.1) of social capital. It contains a structural dimension (network ties, network configurations, and organization), a cognitive dimension (shared codes, language, culture and shared understanding), and a relational dimension (trust, norms, obligations, identification). Whereas the structural and relational dimension correlate with theoretical models of the structural view of social capital, the cognitive dimension is closely related to the ideas of cultural social capital. Relational capital involves social actors "trusting other actors within the group and being willing to reciprocate favors or other social resources in the community."[p.3] (Lu and Yang, 2011). Trust is a "set of specific beliefs related to benevolence integrity and reliability with respect to another party"[p.3] (Nahapiet and Ghoshal, 1998; Lu and Yang, 2011). Cognitive capital facilitates a common understanding of collective goals and proper ways of acting within a social system. Studying the degree of social capital in the model requires a) the analysis of the existing social networks, the corresponding ties (structural analysis); b) the analysis of the existing shared language, frames of meaning, and stories (a cognitive analysis); and c) an analysis of the existing level of trust and reciprocity in a relational analysis.

2.2 The measurement of social capital

Though we know that social capital can be intuitively felt in a given relationship, quantitatively measuring it has proven to be a complicated task since social capital is not a clearly defined concept. The last two decades have shown that social capital is characterized by a multitude of concepts and definitions resulting in a multidimensional concept. The most comprehensive attempt at providing an overview of the existing definitions can be found in the seminal review of Adler and Kwon (2002). Due to the multitude of definitions, there are an equivalent number of methods and metrics that seek to measure social capital (Narayan and Cassidy, 2001). Despite the increase in studies on social capital, the community has not been successful at

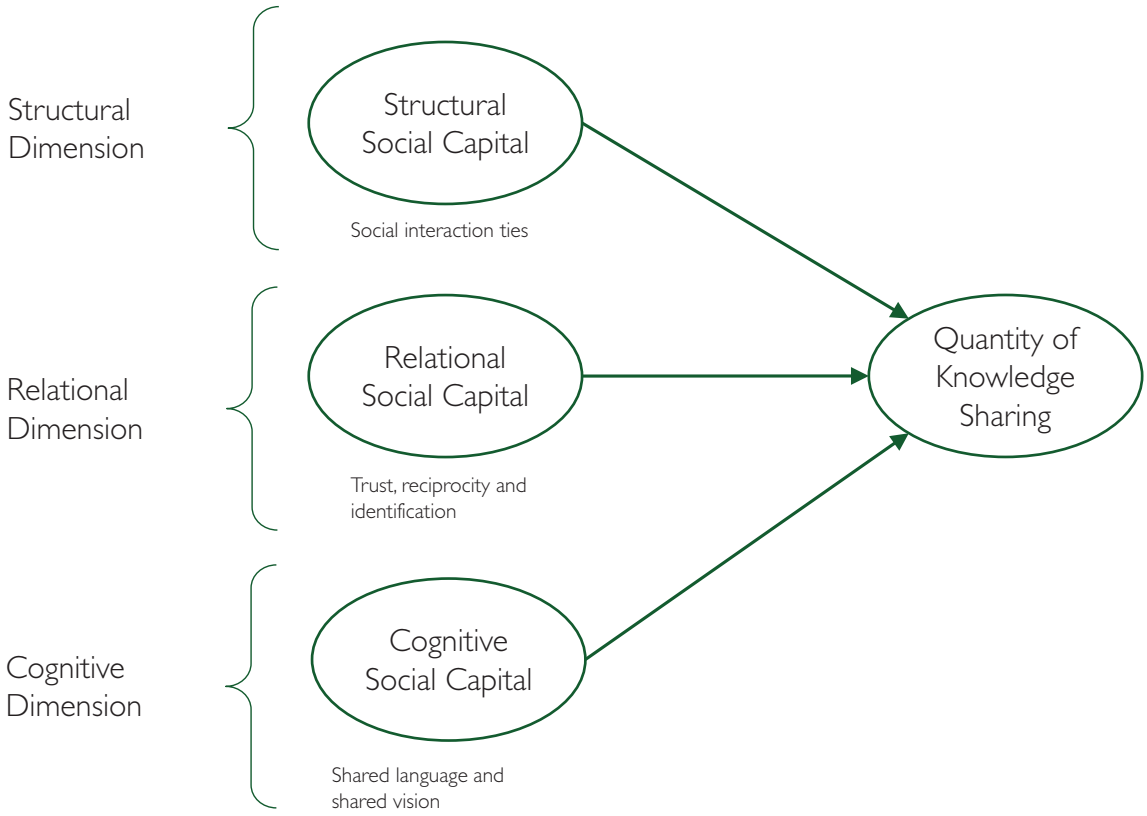


Figure 2.1: Representation of the dimensions of Nahapiet's model of social capital.

defining a single widely accepted method of measurement. In the last twenty years, a multitude of instruments and indicators have been proposed that often resulted from the nature of the underlying data and were not created for the single purpose of studying social capital (Gaag, 2005). On one hand, this has led to interesting research results, but on the other hand it hinders the comparability of the statements made and the methods used (Flap, 2004). Early on in the discussion on social capital, Coleman (1990) doubted if social capital should even be measured. Other authors have also discussed the problem of its measurement: on one hand there is a considerable gap between the concept and its measurement (Paxton, 1999), on the other hand there is also a need to drive forward the reliability and validity of existing methods (Narayan and Cassidy, 2001). Generally, the biggest flaw of the social concept is the lack of consensus on how social capital could be measured in a standardized way (Fukuyama, 2001).

In order to find potential measures for online social capital in Twitter, one might decide to review all social capital measures ever produced, only to find that the measurement of social capital depends on the research subject and the nature of the underlying data (see above). Therefore, the most promising attempt is to first review the dimensions across which social capital can be measured, and then focus on the existing studies that describe the measurement of online social capital according to these perspectives. The following sections will focus specifically on this.

2.2.1 Individual vs. group social capital

Because social capital can be found at different hierarchical levels such as the individual level, the collective level, the micro-, meso- and macro-level (Bhandari and Yasunobu, 2009), in empirical research it can also be analyzed according to these different levels. At the micro-level, relations between individuals, households or neighborhoods are analyzed. The meso-level contains municipalities, institutions and organizations. The macro-level deals with entire regions and nations. This multi-level view comes with a cost: theorists debate whether social capital is a private good, meaning that individuals invest in the formation of relationships so they can access the resources of others, or a public good, such that everybody belonging to a social group with social capital may enjoy its benefits. Putnam's work (1995) and the original works of Hanifan (1920) or Woolcock (2001) describe social capital a quality of groups or societies. More recent studies describe social capital as "group-level social interaction, group-level social trust cues, and group-level social shared codes and language."[p.1964] (Huang and Lin, 2011). In contrast, researchers such as Burt (1995) and Lin (2002) describe social capital in terms of an individual. Here, an actor's position in his social network determines his opportunities and constraints. Burt (1999) describes individual social capital as the individual benefits that result from an individual's strategically-selected bonding and bridging network position. Similarly, Coleman (1988) focuses on individual social capital and describes it as the sum of resources that are available to an individual. Luckily, despite the debate if social capital happens at a group level or at an individual level, both levels can be united under a

network perspective (Borgatti et al., 1998). Borgatti argues that "those two usages [of social capital] primarily reflect two different levels of analysis: the individual and the group (or social system)."[p.1]. This means that both levels of analysis are simultaneously applicable in empirical research.

2.2.2 Bonding (internal) vs. bridging (external) social capital

The second big discussion on social capital deals with its type or source: much of this discussion on bonding and bridging social capital has already been captured in the views of Coleman (1988) and Burt (1999), which have highlighted different sources of social capital. While in Coleman's view social capital is mostly based on the assumption that it results from internal group closure, in Burt's view it results from exactly the opposite group mechanic, namely from structural holes. According to Borgatti and Adler and Kwon, these two types of social capital are actually not that different when one takes into account the different levels of analysis. That is the group and the individual level. In fact, the majority of studies has studied individuals and their embedding in the group (individual bonding social capital) or the individual's position between groups (individual bridging social capital) (Adler and Kwon, 2002). Yet whenever groups have been the research subject, the group "has been implicitly conceived as a universe, nothing outside the group is considered."[p.3] (Borgatti et al., 1998). Adler and Kwon (2002) describe the same process in their review of theoretical views on social capital stating that "external ties at a given level of analysis become internal ties at the higher levels of analysis and conversely, internal ties become external at the lower levels"[p.35]. According to Borgatti (1998), one consequently has to accept the duality of bonding and bridging social capital and see groups as embedded actors in their own social environments. The logical conclusion is to combine the individual vs. group dimension with the inside (bonding) vs. outside (bridging) dimension. This can be demonstrated in an example regarding a work department: when looking internally at the group level, one would analyze the working relationships among the members of the department, but when looking externally at the group level (beyond this department), one would analyze the relationships the department has with other departments inside the company. Thus, in terms of a network, an internal view looks at the group's relationships within the group, while an external view looks at the structure of the group's relationships to outsiders. When we additionally consider the duality of individuals and groups, this leads to a fourfold classification as shown in Table 2.3 that has been suggested by Borgatti (1998), which is a cornerstone of many definitions of social capital and highly used in organizational research (Oh et al., 2006) on social capital.

Figure 2.3 distinguishes between groups and individuals and an internal focus, which is the focus inside the group and an external focus is the focus outside of the group. Quadrant A describes the collective good idea of social capital as described by Bourdieu or Putnam (Bourdieu, 1986; Putnam, 1995). Here the social capital does not belong to an individual, but rather to the whole group. Quadrant B describes the bonding social capital of an individual.

		Type of Focus	
		Internal	External
Type of Actor	Group	A	C
	Individual	B	D

Table 2.3: Overview of the different perspectives on social capital

In Borgatti's original interpretation, quadrant B should be left empty, since there is no way to analyze the individual's internal ties. This would correspond to analyzing the inner workings of an individual, such as the networks formed in his brain. In this definition, quadrant B will be changed accordingly in order to be able to express all four dimensions of social capital using one matrix instead of two[4]. Therefore, quadrant B will be used to describe the individual's internal group focus. Using this logic, the social capital that can be harvested in quadrant B depends on the individuals' ties with the internal group. This idea translates to the popular concepts postulated in the works of Coleman (1988) and Nahapiet (1998), where individual bonding social capital stems from relations of this individual to group members. Quadrant C describes the potential social capital that a whole group can acquire by maintaining ties with other groups. This view has mainly been explored in organizational research of social capital, where groups, organizations, teams or corporations are the research subject[5], and researchers (Oh et al., 2006; Huber, 2009) analyze how teams, companies or departments create networks with each other and how these networks are beneficial to the group's success. Finally, quadrant D corresponds to the individual social capital a person acquires by maintaining external ties outside of the group. This view has also been predominantly described by the structural hole theory of Ronald Burt (1995).

2.2.3 Online vs. offline social capital

Offline measurement

Measuring cultural social capital in the "offline world", as in the traditional works of Putnam (1995) or Coleman (1988), has already introduced a variety of different measurements of social capital (Adler and Kwon, 2002). These have mostly been dependent on the researchers' background, tradition and research focus. In the cultural view of social capital, Onyx and

[4] In Borgatti's original matrix we start with a highly "zoomed-in" matrix where we assume that a person is a "group" of individual nerve cells. This means that in order to study how a person bonds with other individuals, one would have to study how the group of nerve cells creates group bridging social capital (quadrant C) in this matrix. The other three types of social capital are then expressed in the "zoomed-out" matrix where the individual is an actual person and the group is an actual group of persons.

[5] Here most data sets neglects the group's inner workings (Everett and Borgatti, 1999) since only aggregated data is available on such a level.

Bullen provide (2000) an overview of such measures of social capital in their empirical study of Australian adults, and Stone (2006) provides a more current review of survey-based cultural social capital measurement systems. In the structural view of social capital, classic network-based operationalizations can be found in the works of Lin (1999) and Burt (1999). Using a dyadic and triadic perspective, Täube (2004) has additionally provided a set of network measures for Burt's brokerage roles, also building on Gould and Fernandez' (1989) typology of broker roles. Generally, the multidimensionality of social capital has been widely reflected in the literature in terms of the way it is measured, and researchers are often advised to match their method with their focus of research and their underlying dataset (Onyx and Bullen, 2000). What all structural offline social capital measures have in common is that they use a number of generators with various forms in order to construct the underlying networks such as the name generator[6] (Franzen and Pointner, 2007), the position generator[7] (Lin and Dumin, 1986; Gaag, 2005), the resource generator[8] (2005). When direct generators cannot be used (e.g., on a global level), a number of proxies are used instead: for example, trust, volunteerism, public engagement and participation in local communities are often used to determine social capital on a global level for communities, groups and countries.

The presented offline measures of social capital highlight the different perspectives on the types of different entities that can be sampled and measured in order to obtain a measure of social capital. Additionally, they show that depending on the nature of the underlying data, different approaches might be used. Finally, all of the generators highlight the structural network perspective, which plays an important role in the operationalization of social capital, since it is expressed as the network configuration of individuals. Yet those measures also reveal, that by using proxies of social capital, they create their main drawbacks, since they make data collection very cumbersome and complicate the reconstruction of the actual social network from which the social capital is stemmed. The next section will now focus on the measurement of online social capital and show how offline measures of social capital have been transferred

[6] The name generator is mostly used in questionnaires and interviews in which the interviewees are asked to name individuals with whom they talk to about the important issues in their lives or who they call for advice. These people are allowed to name an arbitrary number of individuals. The resulting networks can be analyzed according to their size, density or multiplexity.

[7] The position generator is concerned with the actual resources in the network, rather than the connections between persons. Individuals are asked: "do you know anyone who is a lawyer, a doctor, a manager..?", thus indicating their instrumental use of their network. This generator compensates the flaws of the name generator, and thus helps determine how individuals are able to access different social positions.

[8] The resource generator is focused on highlighting the resources in the network (e.g., "Do you know anyone who can repair a car, a bike, etc."). People receive a list of resources and indicate if they know any individuals that could make these available to them. Van der Gaag and Snijders (2005) note that social capital can either offer access to potential resources that are provided by other network members, or it can describe the actual use of the resources obtained through other network members. By measuring access, there is the problem that only a fragment of the existing social capital is captured, because many network members offer the same resource but often only one resource is needed for individual success. Because of the complexity correlated with the measurement of resource use, Van der Gaag and Snijders decided that it should be measured through access to that resource. The resources are selected in such a way that they represent the needs of an average person in a modern society.

	Online AND social capital	Virtual AND social capital	Twitter AND social capital
Online AND social capital	134	22	0
Virtual AND social capital	-	54	2
Twitter AND social capital	-	-	2

Table 2.4: Overview of the systematic online social capital search process. Each cell shows the intersection of number of studies found for the combination of keywords.

to the online world with all of their advantages and disadvantages. The next section will show what consequences this new environment meant for the measurement of social capital.

Online social capital

To find both studies and measures that are related to online social capital, a systematic literature review of Webster's work (2002) et al. has been performed (see table 2.4). The systematic research on ProQuest for the keywords "social capital" and the term "online" revealed **134** peer reviewed articles, the combination with the keyword "virtual" revealed additional **54** articles, yet the combination with the keyword "Twitter" revealed only **2** articles. This set of articles has been reviewed by studying both the abstracts and the full versions of these articles, and then by undertaking a backward search. A large number of studies dealt with social capital, but showed up in the results because the questionnaires had been hosted "online". These papers are clearly not related to online social capital and have thus been removed from relevant body of knowledge in the review. From the remaining studies only such articles have been included in the review which either a) offer a measurement of online social capital or b) propose conceptual frameworks or c) offer insights on its effects on diffusion of information in online networks. The results of this remaining body of knowledge has been summarized in this section under the term online social capital, which deals with the traditional approach and discussion of online social capital research. The part of relevant literature that deals with network-based measurements of social capital online has been summarized in section 2.2.3 which is called network-based online social capital.

In 2000, Putnam formulated the theory that more technologization leads to more isolation, resulting in a loss of social capital (2000). So when researchers started to measure online social capital, a large number of them focused on the question whether computer-mediated communication (CMC) enhances, diminishes or supplements social capital in the offline world (Wellman et al., 2001). Although Putnam and others were rather skeptical of the benefits

of Internet-based interaction, studies have shown that virtual activity in online communities fosters online social capital (Wellman et al., 2001). Even before the social media revolution, researchers started to note "that new forms of social capital and relationship building will occur in online social network sites. Bridging social capital might be augmented by such sites, which support loose social ties, allowing users to create and maintain larger, diffuse networks of relationships from which they could potentially draw resources"[p.1146] (Resnick et al., 2000). In sum, evidence has shown many different ways (Ellison et al., 2006) in which the Internet builds traditional forms of social capital and online social capital, "including new relationships, access to and co-creation of practical information and theoretical understandings, and networks of friendship, purposive community, and political organizations."[p.406] (Katz et al., 2001). Recently, a number of researchers asked the question if social capital could be transported from the online to the offline world (Skoric and Kwan, 2011). Ye (Ye et al., 2012) suggested that social capital could be transferred from the real world to the virtual one and impact the level of activities that the user undertakes with other users in the online world. Some researchers (Valenzuela et al., 2009, 2008) even found that for young adults and early adopters in particular, these two forms of social capital become more and more indistinguishable.

Survey-based online social capital

Among the first articles that actually focused on measuring online social capital, Williams (2006) was the first to look for survey-based measures of bonding and bridging online social capital. He was motivated by Putnam's original article, where Putnam writes that he has "found no reliable, comprehensive, nationwide measures of social capital that neatly distinguish 'bridgingness' and 'bondingness"[p23-24] (2000). Williams created a set of 10-item scales for online bonding and online bridging social capital. Bonding online social capital is measured by questions like "there are several people online I trust to help solve my problems", while bridging social capital is measured using questions like "interacting with people online makes me interested in things that happen outside of my town". Building upon Williams work, Ellison has (2006; 2007; 2009; 2011) been consistently working on providing better questionnaire-based metrics of bridging and bonding online social capital. By adding a temporal dimension, Ellison has also provided one of the first longitudinal analyzes (Steinfield et al., 2008) of social capital in Facebook, showing that Facebook usage leads to more bridging social capital. In a more recent paper, they also found that Facebook intensity had no influence on bonding social capital (Vitak et al., 2011). Young et al. (2011) came to different conclusions: they found that on Facebook, social ties are strengthened, social networks are expanded and an environment built on trust and reciprocity is created. They note that social networks have the potential to increase users' social capital through consistent exposure to the activities of members of their network and that users socialize with a larger and more diverse group of acquaintances, thus extending potential social capital. They also find that continued research is needed to evaluate the quality of social capital built online. In order to distinguish between bonding (internal) and bridging (external) social capital in online communities, Pipa Norris Norris (2002) provided a framework

allowing for a more nuanced view. The framework differentiates between social and ideological homogeneity and heterogeneity. Socially homo/heterogeneous groups are measured according to demographic variables such as age, gender or class. Ideologically homo/heterogeneous groups are divided according to their perceived similarity of ideas, interests or values. Norris finds that "online communities bring together like-minded souls who share particular beliefs, hobbies or interests"[p.37], resulting in the creation of bonding social capital. Regarding bridging social capital, she only finds rare evidence of groups that are socially and ideologically heterogeneous. She concludes that communication on the Internet is driven through the need for similarity. Norris notes that this happens mostly between friends or ideologically homogenous actors. Finally, there are also a number of studies (Zhong, 2011; Kobayashi et al., 2007; Huvila et al., 2010; Mathwick et al., 2008) that have evaluated the concept of social capital in the online gaming environment (MMOG, Second Life, PlayStation 3 communities etc.). Here it has been found that online gaming positively influences gamers' online bonding social capital and online bridging social capital.

Many of the articles reviewed that come from the information science domain use the social capital concept to explore knowledge diffusion or information diffusion between members in online and virtual communities. All of these studies used Nahapiet and Ghoshal's three-dimensional model of social capital (1998) as a common reference. Chiu et al. (2006) proposed an instantiation of Nahapiet's model of social capital to investigate the motivations behind knowledge sharing in virtual communities. Their study found that certain facets of social capital such as social interaction ties, trust, norm of reciprocity, identification, shared vision and shared language do positively influence the way individuals share knowledge in virtual communities. Yoon (2011) also built on Nahapiet's model of social capital and found that social interaction ties, trust, norm of reciprocity, identification and shared goals positively influence members' knowledge-sharing intentions and knowledge quality in virtual communities. Davenport (2011) also used the social capital lens in order to highlight how different social capital formation processes contribute to member identification. The results highlight how some dimensions of social capital augment each other and affect identification through the four conditions that influence social capital development: time, interdependence, interaction and closure. Lu et al. (2011) also used Nahapiet's model of online social capital to explain information exchange in virtual communities. Their results showed that Nahapiet's different levels of social capital might interfere with each other in such a way that structural capital significantly affects cognitive capital, but doesn't negatively affect relational capital. Bauer et al. (2005) analyzed the online community of a virtual liquor brand and found that social capital was a key instrument to establishing satisfaction, trust and commitment between the vendor and the members, and that it often facilitated information diffusion within the community. Xiao et al. (2012) analyzed a Chinese virtual community, and used Nahapiet's model of social capital to measure its effects on knowledge exchange. Their model takes into account the mediating factors of perceived trust and outcome expectations and also shows its positive influence on knowledge exchange. Finally, Lin et al. (2011) use Nahapiet's three dimensions of social capital and find that they simultaneously positively influence knowledge sharing and team

performance in online virtual communities.

The final group of studies reviewed focus on the effects of online social capital on virtual team performance. The respondents' perceptions of team social capital were measured via online surveys. Schenkel et al. (2009) explored the role of various forms of social capital on the performance of entrepreneurial teams in a virtual context. Using Nahapiet's model, the respondents' perceptions of social capital were measured in the form of relational capital and cognitive capital. Schenkel et al. found that all of these forms of social capital were significantly positive predictors of team efficacy. Entrepreneurial orientation was also found to significantly increase team efficacy. Baruch et al. (2012) found that knowledge sharing is indirectly positively influenced by team politics and social capital via the mediation of cooperation and competition, while team performance is indirectly positively affected by team politics and social capital via the mediation of cooperation, team emotional intelligence and team competence.

From survey-based online social capital to network-based online social capital

The last section has shown that the research on online social capital is beneficial in studying the effects of social capital on knowledge and information exchange and in creating models of social capital. Yet despite the new researched contexts of online and virtual social capital, the so far presented researchers have used traditional research tools. This is not so surprising as it was a natural step to use already existing offline measures of social capital in an online context. This means sampling from a population of people and measuring their social capital by using questionnaires which rely on the old offline social capital generators. Yet this also means that those generators also introduce the same problems and issues in the online context, such as network construction or scaling problems. This is something that can be avoided, as the online world allows us to capture the ties and interactions of individuals in a highly detailed, fast and unobscured way. We only have to study the data traces that each individual leaves behind in the online world and then reconstruct the social ties and interactions from them. Additionally by making use of the available interactions that can take place on each of the researched online networks we can adopt the social capital measurement to each individual online environment. Since it makes a great difference if people for example can exchange goods or ideas on an online network or not, or if they can form reciprocal relations or not, or if they can interact with each other or not, it becomes important to adopt the social capital framework to the underlying online environment. Although people use those online networks in slightly different ways, what remains the same is that by carefully interpreting those networks and the contained resources one is able to measure the social capital from the data traces alone. Therefore the next section is reviewing contributions that are paving the way towards a new way of measuring online social capital, which is determined directly from the available data traces people leave behind.

Network-based online social capital

The explosion of social media networks has led to the creation of huge amounts of data. New methods that collect and analyze this data have been created: they are often referred to as webometrics (Thelwall, 2010). These webometric studies all try to directly measure online social capital by using the traces people leave behind online. These traces are combined into networks of online users that are then used to conceptualize and measure social capital. Such a new way of using networks and network metrics is endorsed by established researchers (Lu and Yang, 2011) of online social capital who claim that "further research should attempt to explore other indicators [of social capital] or multiple measures for structural capital, such as degree centrality or betweenness centrality"[p.536]. Stegbauer and Rausch (2006), social scientists that focus on the empirical analysis of online communities, also advocate this method. Their structuralist approach is grounded, theoretically and methodologically, in a (social) network analysis of communities, where so-called "non-reactive" data becomes the basis for analysis. "Non-reactive data" refers to the traces people leave behind online, which do not have to be collected by questionnaires. Vergeer at al. (2011) advocate that "to study online social capital, traditional and new means of data collection and analysis can be used." They also note that these "new means" are actually not that new: before the emergence of social networks and metrics, such attempts were often referred to as hyper-network analyses, and were usually used to analyze bloggers rather than social media users. The following review of studies will present the various operationalizations of network-based online social capital and the results they have produced.

One of the most prominent studies in this field is the study of online communities by Wasko et al. (2005): it proposes an online social capital measurement that, among other traditional metrics, measures individual structural social capital by using the actors centrality in the network. Wasko et al. further evaluated the effects of online social capital on knowledge contribution and found that network centrality, measured by the number of ties to other members, had a positive influence on knowledge diffusion. Hsiao et al. (2011) also used a network-based social capital operationalization and found that a player's network centrality in an online gaming community affects his/her attitude and continuance intention toward a massively multi-player online game (MMOG).

Two studies investigated the influence of structural holes in online communities: Burt (2010) investigated if actors in virtual worlds such as Second Life were more likely to span virtual structural holes or pursue group closures. He used the rich network data available in virtual worlds to measure social capital in an online context. He found that network brokers were more likely to provide social infrastructure that makes the virtual world valuable and attractive. Groups founded by such brokers were more likely to survive and attract a greater number of members. Pitt et al. (2010) constructed small actor-networks (100 actors or less) using the links between company websites, and identified the most prominent actors in these networks. Then, using the structural holes analysis, they showed that the entrepreneurial opportunities surrounding these actors can be unveiled.

A number of studies have measured network-based online social capital at the group level: Gaines et al. (2009) took small samples of the Facebook network and measured online social capital by analyzing the clustering coefficient of random groups. They found evidence that members cluster according to their political views. Yet they also concluded: "the Facebook network makes it somewhat cumbersome as a platform for social scientific research, due to privacy concerns"[p.220]. Rafaeli et al. (2004) investigated social capital in a virtual community by creating a social communication network of activities in authenticated discussion forums and then measuring the density of the network. They showed that by building up group bonding social capital users can easier be motivated to join the group conversation. Finally, Brooks et al. (2011) offered three structural measures of students' social capital using their online social network data. They define general social capital by the size of the students' network; bridging social capital by the number of student clusters; and bonding social capital by the average degree of each cluster.

Only one study provides a tie-centric measure of social capital (Xiao et al., 2012) and finds that the two dimensions of online interaction, frequency and centrality, can positively influence knowledge exchange reciprocity in the virtual community.

Regarding an integrated model of network-based social capital online, the only relevant study is the work of Marc Smith (2011) who provides his own social capital framework, which can serve as the basis for the framework of social capital for Twitter. Smith bases his framework on the notion of implicit affinity networks. In such networks links are implicit in the patterns of natural affinities among individuals. He provides an effective mathematical formulation of social capital based on implicit and explicit connections. He measures this online social capital for three different online communities and shows that it is growing throughout the time of observation. His framework consists of three main components of online social capital:

- Relationship strength as measured either by the explicit or implicit relation.
- Resources as categorized by their duplicability, transferability, exhaustibility, returnability, quantifiability and durability.
- Interactions which affect the relationships. They are categorized by whether the interaction involves the exchange of resources and how the interaction is perceived (positive, negative or neutral)

Using agent based simulations he proves the usability of his framework and demonstrates that the majority of social capital concepts are well quantifiably transferable into the online domain. Instantiations of these concepts for Twitter based on parts of his framework (relationship strength and interactions) will be discussed in detail in chapter 4 and serve as the basis for the framework of social capital for Twitter.

2.3 Conclusion and goals

This literature review has led to a number of conclusions and goals. First, this chapter detailed the different perspectives on the concept of multidimensional social capital. Second, it showed that the measurement of social capital must be made according to these multiple dimensions and that the measures should be carefully selected to match the researched subject. Third, the systematic review of online social capital studies has shown that despite the fragmentation of this body of knowledge, Nahapiet's (1998) three-dimensional model of social capital is widely used as a frame of reference to conduct research on online social capital. Fourth, social capital research has proven itself as a rigid framework that looks at relations between individuals and groups from a very analytic perspective. Finally, the review has shown that the concept of social capital is useful to predict a multitude of outcomes or benefits that result from social networks. This is what makes social capital such a useful concept - social capital describes multiple dimensions of network structures and their influence to obtain certain goals.

The systematic research on online social capital has found very strong evidence for Lin's original prediction "that an increasing number of individuals are engaged in this new form of social networks and social relations, and there is little doubt that a significant part of the activities involve the creation and use of social capital. Access to free sources of information, data and other individuals create social capital at unprecedented pace and ever-extending networks"[p.46] (1999). The review has shown that the significance of online social capital research is stronger than ever, as "we are witnessing a revolutionary rise of social capital as represented by cyber-networks. In fact, we are witnessing a new era where social capital will soon supersede personal capital in significance and effect"[p.45] (1999). Moreover, Lin's assertion that "much work is urgently needed to understand how cyber-networks build and segment social capital"[p.47] (1999) is still valid today. While the reviewed information system studies are a big step forward in explaining the effects of social capital on information diffusion, researchers (Reagans and McEvily, 2003) continue to argue that "future research should continue to probe how knowledge transfer is affected by behaviors induced by network structure and consider how these effects interact with other factors influencing the knowledge transfer process"[p.264]. Indeed, the literature review has revealed several serious gaps in the research on social capital. First, most studies of online social capital have focused on individual and bonding social capital (quadrant B in table 2.3). The other three quadrants of online social capital have been less explored. Second, most of the research on online and virtual social capital view social capital as a dependent variable. Because of this, most studies have tried to explain which factors lead to the creation or decline of social capital. This is a direct result of Putnam's work, which posited that Internet use would diminish social capital. Putnam's study motivated myriads of researchers to analyze this hypothesis: in the end, they found contradicting evidence. Third, the body of literature that studies online social capital in Twitter is almost nonexistent (see table 2.4), even though Twitter offers one of the biggest collections of virtual communities in the world. Finally the effects of social capital in Twitter on information diffusion have not been explored yet.

The most serious gap in the literature is related to the measurement of online social capital. Given the abundance of data, we urgently need new online-data-based metrics that are able to measure social capital directly in the online world. The review has shown that there is a strong tradition of using network metrics to measure social capital: already in 1998, Portes (1998) presented a network-oriented perspective of social capital, indicating that social networks are a core concept of social capital and that social capital determines the social exchange process and results. Much seminal work on social capital, such as the works of Burt (1995) and Lin (1999), have been based on this premise. The most prominent models of social capital such as the widely-used model of Nahapiet (1998) have embraced the concepts of network structure, relational embedding and cognitive norms. However, even though there is an abundance of available network data, researchers still use questionnaire-based methods for online research. This has lead to recent suggestions to replace the survey-based method of capturing social capital for network based measures, also called webometrics or sociometrics (Thelwall, 2010). Indeed, it is a huge waste to neglect the great amount of behavioral and structural data traces online and still continue to use survey-based methods. As the literature review shows, some work has been done, but there is still much to do in this area. First of all, there is no consensus on a single unifying framework for online social capital. Second of all, the existing studies only scratch the tip of the iceberg in regards to the available network metrics and behavioral traces. Third of all, only one study tries to conceptualize a network-based social capital measure for the Twitter network.

This brings us to the three research objectives that result directly from this review and create an academic basis for this contribution. First, the existing literature on information diffusion, has to be reviewed, in order to better understand the process of how network structure affects information diffusion or knowledge exchange. This review will have to be performed in respect to the structural, relational and cognitive dimensions of Nahapiet's model of social capital. Additionally a systematic literature review will have to show what studies and insights are already available on information diffusion in Twitter. Second, since the current literature on social capital does not offer a concept of online social capital for Twitter, the second objective of this study is to formulate this concept. It will have to adapt Nahapiet's model of online social capital to the context of Twitter. This concept will also have to take into account all of the structural and behavioral dimensions of this virtual community. Finally the third objective of this study is to use this social capital concept in order to provide a network-based operationalization of social capital for Twitter. The operationalization will need to measure social capital directly using the available traces (Rausch and Stegbauer, 2006) that people leave behind in Twitter. This will allow us to postulate hypotheses on its effects on information diffusion in Twitter.

3 Literature review of social capital and information diffusion

The outcome of the literature of social capital in chapter 2 has shown that there is a strong need to measure online social capital in a network based way. If one wants to conceptualize online social capital in a network based manner and form hypotheses on how it affects information diffusion, this also means that the available literature on how networks influence information exchange has to be reviewed. Therefore, the first part of this literature reviews how network structure affects information diffusion or knowledge exchange and helps to organize these insights with the existing perspectives and dimensions of social capital. This means that the first goal of this chapter is to review and map diffusion theories onto the network-based model of social capital. The second goal of this chapter is to review how much is already known about information diffusion processes in Twitter. Therefore, the second part of this chapter consists of a systematic literature review on information diffusion on Twitter. A keyword search will be performed to find all material relevant to information diffusion, which will then be organized according to Nahapiet's framework, that is in terms of structural, relational and cognitive social capital. The insights from this literature review will highlight research gaps, which will then be addressed by the social capital framework for Twitter and the resulting research questions.

Social capital and information diffusion

According to Valente (2010), social capital is a phenomenon that is rarely associated with information diffusion: indeed, only a few common bodies of literature describe both phenomena (see also Burt (1999)). Astonishingly though, the fundamentals of the social capital concept are based on the assumption that the most basic resource transmitted through interpersonal networks actually *is* information or knowledge. In the last chapter, numerous studies (Wasko and Faraj, 2005; Chiu et al., 2006; Lin, 2011) especially from the information science have shown how the theory of social capital can be used to explain information diffusion in virtual communities. On one hand, these studies have highlighted how the research on online social capital provides a significant extent of evidence how information is likely to be distributed in networks and how different types of information are likely to accrue to individuals and groups. On the other hand, these studies have mostly lacked the actual network perspective. The additional analysis of network-based information diffusion theories in regard to social capital will therefore not only help to create more theoretically founded hypotheses on the

Diffusion Theory	Concept	Social Capital
Diffusion of Innovation	Interpersonal Diffusion & Group Norms	Cognitive Social Capital
Diffusion of Information	Two/Multi-Step-Flow, Opinion Leaders	Structural Social Capital
	Strength of weak ties	Relational Social Capital

Table 3.1: Overview of the concepts from diffusion theories and their connection to social capital

influence of social capital on information diffusion but also contribute to the social capital framework for Twitter.

Networked information diffusion processes have been studied in communication studies, social studies, social psychology, public administration, marketing and epidemiology. The analysis of information diffusion has been of interest to researchers from various domains ranging from social sciences, epidemiology, disease propagation, physics and economics (Newman, 2002b; Watts and Peretti, 2007; Zachary, 1977). Different research traditions have therefore come to different conclusions and insights whose review is important, because it contributes to our understanding of the social diffusion process. The following chapters will briefly introduce the insights made from the research on diffusion of innovations and information and connect them to the social capital concept (see table 3.1). By finding parallels between theories explaining information diffusion and the matching social capital concepts these insights will help to generate meaningful hypotheses on the influence of social capital on information diffusion.

3.1 Cognitive dimension of social capital

The main theories explaining the influence of the cognitive dimension of social capital on information diffusion are closely related to theories on interpersonal diffusion of innovation. This section will briefly lay out this theory and conclusively connect the findings to cognitive social capital.

Interpersonal diffusion of innovation

The main concepts of diffusion of innovation are founded on the basic observation that new ideas and information spreads through interpersonal contacts and are based on interpersonal communication (Bass, 1969; Katz and Powell, 1955; Rogers, 1995; Valente, 1995). The first groundwork to diffusion of innovation research has been postulated as early as 1943 by Ryan & Gross (1943). They were among the first to provide evidence, that the diffusion of hybrid-grains

in farmer communities was not based on the commercial channels (for example traders that were important sources of knowledge about this new grain). Instead it were rather neighbors and colleagues that influenced each other and thus decided if the innovation was accepted by an individual or not. Their research started a long lasting tradition of diffusion of innovation research. Amongst the most quoted contributions to the research on diffusion of innovation is the work by Rogers (1995) (see (Abrahamson and Rosenkopf, 1997; Bass, 1969; Mahajan and Peterson, 1985; Ryan and Gross, 1943; Valente, 1993; Young, 2005)). Diffusion in his terms is "the process by which an innovation is communicated through certain channels over time among the members of a social system"[p.35] (1995). His observations explain the spread of new ideas and practices within and throughout communities. Ideas or innovation can originate from outside the community and can then be carried into a community or to a person, through interpersonal relationships.

Rogers' model consists of four main elements: The characteristics of the innovation, different types of adapting individuals, their individual adoption process, and their final decision of adoption or its rejection. Rogers defines innovation itself by its relative advantage[1], compatibility[2], complexity[3], triability[4] and observability[5]. All of those factors influence the adoption of the innovation. The process of individually accepting an innovation is generally understood under the term adoption. When adapting innovation persons usually go through a psychological decision process (Rogers, 1995) in which they first learn about the existence of an innovation, then build up knowledge on this innovation, then develop an attitude towards the innovation and finally as a result adopt or reject an innovation. Based on their individual acceptance of the innovation in time Rogers defines individuals as early adopters (first 16%), early majority (17%-50%), late majority (51% -84%) and laggards (85%-100%). On a system wide level this leads to a shape of a S-curve of the diffusion process in the population.

Influence of group norms

Rogers also highlights the normative group effects between the individual and his social environment in the diffusion process. The integration of persons in groups or communities through the interaction and communication process forces each individual to examine group norms. Norms allow establish a certain desired behavioral pattern for the group members and so decide about the accepted behavior in the group (1995). Rogers defines three different possibilities of behavior for a group member: the conformity with the group, the following of rules upon being demanded (compliance) and the total obedience. With a high level of uncertainty (respectively with a high cognitive costs towards the group conformity) members

[1] Perceived improvement through the innovation
[2] Perceived compatibility with the existing standard
[3] Perceived complexity or difficulty in regards to understanding and dealing with the innovation
[4] Perceived possibility of easy trial
[5] Perceived degree of observability of the results of the innovation

can adopt opinions or behaviors of the group, because those are trusted more than its own ones[6]. The normative influences can also lead to the adoption of opinions or behaviors despite the existence of uncertainty among group members, in order to gain acceptance in the group or to avoid denial or dis-affirmation. Such compliance was also defined by Menzel and Katz (1955) as a "network-based decision [...] with a pressure towards conformity"[p.306] (Wejnert, 2002). Finally when the internalization of norms has already taken place and the person has shown its identification with the group, behaviors and opinions are usually obediently resumed (Rogers and Kincaid, 1981). In this case it is important to note that certain norms, the conformity of group members and their obedience can also be barriers for the diffusion innovations: The message does not diffuse through certain networks, because its content is frowned upon or even forbidden such as in certain religious communities (Rogers, 1995). The influence of the group is thus able to change cognitive behavioral aspects like decisions, opinions or the reasoning of persons, which has a significant influence on the forwarding of information in groups (Berth, 1993; Lapinski and Rimal, 2005).

The theory of diffusion of innovation provides fundamental insights about the when, why and how individuals adopt a certain innovation and how it influences the diffusion process. It introduces the concept that the diffusion of innovations is based on an interpersonal communication and provides definitions of the different phases of adaption that individuals go through. It draws conclusions on the resulting shape of the diffusion curve in the population and covers normative group concepts that provide answers how the individual diffusion process can be influenced by the normative forces inside a group. While those insights together paint a coherent picture of how diffusion of information or innovation can occur in social structure, the theory of diffusion of innovation almost completely omits an operationalization of the envisioned networked structure of individuals. Yet forming connections and passing information along those connections is a fundamental process of information diffusion online. Additionally when regarding the influence of group norms it lacks to further conceptualize how groups and their norms are enforced in a networked structure. Although the insights on norms of groups and their influence on information diffusion might certainly be valid in an online context their definition is problematic, since people belong to multiple social circles and have the possibility to freely change their groups without sanctions. adopting an innovation and passing along an information have certain similarities, they also differ in some crucial points such as the adaption process or simply the measure of innovation for a given information.

Conclusion for social capital

The diffusion process that Rogers describes, touches a lot of the concepts described in Nahapiet's work under the term of cognitive social capital both on a group- and individual-level.

[6]Menzel and Katz (1955) have observed this behavior in the adoption of prescriptions of drugs among physicians and note that "there are a variety of ways in which such simultaneous decisions may be reached: perhaps a decision, once reached by one member of a clique, is easily accepted by his associates who trust his judgment"[p.349] (1955).

In his work farmer communities create the cognitive frame, while interpersonal contacts create the social structural dimension of social capital. Practices that spread within communities are a manifestation of diffusion influenced by group bonding social capital, while bridging social capital influences diffusion when Rogers observes that these practices must also spread beyond the community. Although Roger's focus lies on describing how the adaption process fits into the diffusion process, he highlights that the bonding and bridging mechanisms of social capital have a strong influence on the diffusion of innovation. This stream of research highlights how certain cognitive group norms are created and how in turn these cognitive norms can enhance or harm the diffusion process. Transferring these insights to virtual online communities, this stream of research provides an explanation why stronger group norms, and thus a stronger cognitive dimension of group social capital might lead to more information diffusion inside the community. The parallels between the influence of group bonding and bridging social capital and group norms are especially interesting: One finds evidence that depending on the compatibility of message and group norm the diffusion process might be hindered or accelerated. It can generally be assumed that incoming bridging information from other groups will diffuse less in the target group.

3.2 Structural dimension of social capital

The most prominent theories covering the influence of structural dimension of social capital on information diffusion are related to media centric theories explaining the interpersonal diffusion of information. This section will briefly lay out the main concepts and explain what they mean for the influence of structural social capital.

Diffusion of information

Traditional media theories on dissemination information

Theories on the diffusion of information were mainly formed in the media theoretic domain, where theories tended to focus on mass-communication as a form of a "one-way message transmissions to a large undifferentiated and anonymous audience"[p.35] (Maletzke, 1963) or on interpersonal communication as a "message exchange between two or more individuals"[p.3] (Walther, 1996). Correspondingly, debates among media theorists have tended to revolve around the relative importance of these two modes of communication. The early concept of mass communication was founded on the assumption of an "hypodermic needle" (Lazarsfeld et al., 1965), assuming that information that flows from mass media to the audience influences it directly. This theory painted a diffusion process where professional observers and communicators (agencies and media) access, select and filter, produce, edit news and then

directly distribute them via the media to members of society and the interpersonal message exchange between individuals was mostly neglected.

Two- and Multi-Step-Flow

The two-step flow of communication model was therefore introduced to augment the assumption that mass media had a direct effect on its consumers. The theory of the Two-Step-Flow was one of the first and most influential theories describing the information flow from a mass media through opinion leaders to a wider range of individuals also called opinion followers. The theory describes the flow of information from a mass medium through so-called opinion leaders to his or hers opinion followers: "Ideas often flow from radio and print to opinion leaders and from these to the less active sections of the population"[p.64] (Lazarsfeld et al., 1965). Media effects thus do not flow directly from the medium to the recipient, but also through interactions between the recipients (Katz and Powell, 1955; Lazarsfeld et al., 1965). Merton (1957) also examined the role of opinion leaders[7] in local communities. This starting hypothesis was then extended in a series of follow-up studies (Berelson et al., 1954; Coleman and Katz, 1966; Coleman et al., 1957) encompassing the flow of information between the opinion leaders themselves. The Multi-Step-Flow was then introduced, where opinion leaders influence each other.

Opinion leaders

The opinion leader observation, deals with persons that play special role of an "opinion leader" and can influence decisions and information flows in the group. Persons with such a role have been subject to a myriad of studies (Bass, 1969; Gladwell, 2000; Katz and Powell, 1955; Lazarsfeld et al., 1965; Merton, 1957; Schenk, 1993; Valente, 1993; Weimann, 1994) in different disciplines. In marketing for example the idea that a small group of influential persons, can lead to a dilution of the adoption rate (Chan and Misra, 1990; Coulter et al., 2002; Van Den Bulte and Joshi, 2007; Roch, 2008; Vernette, 2004) is still discussed. But also among the sociologists (Arndt, 1967) and the media sociology (Gitlin, 1978) the theoretic concept has widely been adopted for a long time. Since the development of this approach through Katz and Lazarsfeld over 3900 studies have been published which have been analyzing influential opinion leaders (Gladwell, 2000; Goldenberg et al., 2009; Berry, 1995) and their personal influence factors (Weimann, 1994).

Those studies reveal a number of specifics on the characteristics of opinion leaders and their role in the process of diffusion of information and innovation: People that are perceived as

[7] He, too, found structures in which the same individuals were repeatedly asked for advice. Unlike Lazarsfeld, he calls the opinion leaders "influentials", but otherwise they have similar characteristics. As Duncan Watts points out (2007) the differences here are thus primarily of terminology not of content.

opinion leaders, filter, interpret and disseminate information purposefully in their target group. Opinion leaders are highly appreciated in the group and they are good at overlooking their social environment because they are connected to many persons. Opinion leaders reflect the norms of a community, rather than introducing new norms into a community, because a deviation from the accepted group behavior would put their privileged position in the network at risk (Rogers, 1995). Regarding the diffusion process it is crucial, which connections such opinion leaders maintain among each other and how often other persons perceive those opinion leaders as important in discussions or as consulting partners (Coleman, 1988). Persons that have multiple times been nominated as important consulting partners (Rogers and Kincaid, 1981), have been shown to act as opinion leaders in the group and influenced the observed diffusion process.

Rogers states (1995) that "when opinion leaders are compared with their followers, they are more exposed to all forms of external communication and thus are more cosmopolite, have somewhat higher social status and are more innovative (although the exact degree of innovativeness depends, in part on the systems norms)."[p.27] When regarding the time line of the adoption process (1995) opinion leaders are among the early, yet not the earliest adapters of new ideas and practices. Depending on compatibility with the particular innovation, the opinion leaders can perform the "early adapter" role (Becker, 1970). Yet normally the earliest and most innovative adapters are found in the periphery of the group. They initiate innovation because they are different. It is only with a high compatibility with the group norms where the opinion leader tends to be a very early adapter. If opinion leaders once adopt an idea, they can help to accelerate the diffusion of that innovation because they help to carry innovation to the rest of the group. The number of connections between adapters and non-adapters is expected to rise rapidly if opinion leaders participate in the diffusion process. Therefore opinion leaders help to shift the adoption from the periphery of the group into the center of the network and so accelerate the innovation or information diffusion process.

Regarding the media centric theories on diffusion of information we see that mass communication has experienced an expansion through social media and in the opposite direction interpersonal communication has gained new momentum through the same. Yet theories on the core of media theories describing diffusion of information still focuses on the flow of information between institutions and an unconnected audience. In a time where the audience is more connected than ever to the extent that media institutions are indistinguishable from their audience that poses a severe limitation of this theory. Accordingly this results in the shortcomings of the operationalization of the Multi/Two-Step-Flow theory, that highlights the role of intermediaries between groups but it does not capture the inherently networked structure of those groups. The resulting research on the role of opinion leaders emphasizes the importance of certain individuals in the diffusion process, that are able to accelerate this diffusion, yet lack to provide details on for their operationalization. Their importance in the system is often accredited to their attributes and not to their relations.

Conclusion for social capital

In regard to online social capital, the perspectives on diffusion of information translate directly to the structural component of social capital, manifested in the notion of brokers (Burt, 1987; Merton, 1957) and opinion leaders (Bass, 1969; Gladwell, 2000; Katz and Powell, 1955; Lazarsfeld et al., 1965; Merton, 1957; Schenk, 1993; Valente, 1993; Weimann, 1994). While brokers achieve individual bridging social capital by brokering between two different communities opinion leaders achieve social capital by embedding themselves in the group structure. Thus brokers are among the first to obtain and forward new information and opinion leaders are the first that are able to accelerate the diffusion of information inside the group. Regarding the diffusion process, the theory of Two-Step-Flow provides us with additional evidence, that such brokers play a crucial role in the diffusion process between different communities. In regard to online social capital the opinion leader theory translates into the structural view of social capital, where certain well embedded individuals with high individual bonding social capital can have an influence on information diffusion inside the group. Members that hold a strong structural position in the group will be able to better diffuse their own information among less influential members.

3.3 Relational dimension of social capital

Strength of weak ties

Mark Granovetter's theory of "the strength of weak ties" (1978) highlights the influence of the type connections between individuals on the information process. The initial assessment of his theory is the notion, that society is divided into small homogenous groups that are rather closed to the outside. The collection of such groups forms the fabric of the structure of society. Granovetter notes that members of those groups are connected by different ties, which he classifies, based on their strength, into weak and strong ties. New information like for example relevant information on a new job position might reach the group only through weak ties that the group forms to other groups. Those ties then have a bridging function, by connecting the individual homogenous subgroups and so allowing members of the group to access social capital resources outside of the group. Granovetter finds that while weak ties are represented more often, they have a smaller interaction frequency. Weak ties often serve a specific function like e.g., maintenance of connections to work colleagues, or other institutional frameworks. Because of the instrumental character of the weak ties, they are emotionally perceived as rather neutral. Weak ties have obvious influence on the diffusion because their presence creates bridges that connect different segments of a network. The main argument for the strength of weak ties is that they are responsible for the information diffusion and not for the adoption of behavior. Although Granovetter notices that there is no correlation based on

the strength of the tie and the willingness for adoption, weak ties though can be very effective at the communication of information (Burt, 1995; Valente Thomas W, Foreman, 1998).

Conclusion for social capital

In regard to social capital the strength of weak ties directly mentions the relational dimension of social capital. Granovetter puts a strong emphasis on the tie itself and points out that it is the relational dimension that matters. In regard to social capital it has also been observed that bonding ties between members of the same group are stronger than bridging ties between members of different groups. Granovetter observes that through exactly these weak ties that connect different groups novel information reaches the group. This concept has been exactly captured in the notion of brokers, who strategically take advantage of such ties and try to span as many structural holes (Burt, 1995) as possible. This allows them to capture novel information quickly and so creates an advantage to other group members. On the other hand in regard to social capital inside the group the observations of strong ties, have been mentioned by Coleman and others who emphasized that strong relations inside the group are beneficial for the individuals. This duality between internal and external information diffusion and the according tie strengths makes the exploration of relational social capital very interesting.

3.4 Information diffusion studies on Twitter

After connecting the existing literature on information diffusion to social capital concepts the next section will focus on reviewing studies that are covering information diffusion and knowledge exchange in the Twitter network. The insights from this review will show which dimensions of social capital have been explored in Twitter and will help to operationalize the social capital framework for Twitter. The systematic literature review has been performed according to the guidelines of Webster et al. (2002) where a keyword search (see table 3.2) has been performed on ProQuest[8]. The systematic research on the keywords "Twitter" and "information" revealed **283** articles, the combination with the keyword "diffusion" revealed additional **21** articles while the combination with the keyword "knowledge exchange" revealed only **3** articles. The result set of articles has been reviewed, by both studying the abstracts and full versions of these articles, followed by a backward search. Due to the novelty of Twitter as a research subject - in contrast to the systematic search on social capital - here additionally non-peer reviewed scholar journals, dissertations and conference papers have been considered. In general a large number of the found studies deals with the Twitter network and its different applications but only showed up in the results because the keyword "information" is rather general and part of many abstracts. It is yet not possible to completely exclude this keyword from the search, since it is the most commonly used keyword in combination with diffusion studies in Twitter. From the remaining relevant studies only such articles have been reviewed in detail which either a) describe an actual diffusion process in the Twitter network or b) propose conceptual frameworks for it or c) offer operationalizations of the different theoretical social capital models discussed before. The results of the review are structured twofold: Firstly this section will provide an overview of the existing studies according to their size, method of data collection (see table 3.3). Secondly and most importantly the relevant studies and their results will be categorized and summarized under Nahapiet's structural, relational and cognitive dimensions of social capital, which will be performed in sections 3.4.1, 3.4.2 and 3.4.3.

Different types of diffusion

While it is seems obvious that the research of diffusion of information on Twitter needs to observe some sort of "entity" that diffuses through the network, there are more than one perspectives in literature on what this observable entity might be:

- **Tweets and Retweets:** The majority of studies (see table 3.3) study the diffusion of tweets and their respective retweets throughout the network. Galuba et al. (2010) distinguish between so-called "F-cascades", where the flow of information is constrained to the follower graph and "RT-cascades" for which the follower graph is disregarded, and the cascade is computed from the content of the tweets on who-credits-whom. Although

[8] `http://search.proquest.com` which is proposed by Levy and Ellis (2006) as the leading literature database listing the majority of IS World's top 50 ranked journals

	Information AND Twitter	Diffusion AND Twitter	Knowledge exchange AND Twitter
Information AND Twitter	283	6	2
Diffusion AND Twitter	-	21	0
Knowledge exchange AND Twitter	-	-	3

Table 3.2: Overview of the systematic Twitter information diffusion search process. Each cell shows the intersection of the number of studies found for the combination of keywords.

this seems to be a minor[9] distinction their work indicates that 33% of the retweets that they observe credit users that the retweeting users do not follow.

- **URLs:** A smaller proportion of studies (see table 3.3) focuses only on observing the diffusion of URLs throughout the network. While this leaves out a great proportion of tweets it allows researchers to collect bigger cascades, and to find retweets that have been produced by persons not following the original author. To measure how often a URL has been retweeted can be either measured by (1) how often a URL can be found in the corpus of collected URLs or (2) by the number of clicks a URL has received.
- **Hashtags:** Researchers (Romero et al., 2011; Kwak et al., 2010; Ratkiewicz et al., 2010) have also analyzed the spread of hashtags in the Twitter network. Here the adoption of the usage of a certain hashtag is in the focus of the analysis.

Due to the fact that often the research questions of the information diffusion studies on Twitter vary widely, the next section will first provide a tabular overview (see Table 3.3) of the reviewed studies and report their sample sizes, the method of collection, their definition of social ties between users and the analyzed entity. The table 3.3 shows that the studies differed on sample sizes where the smallest sample contained 7 thousand users and the biggest 41 Mio. users. It also shows that usually not all tweets of the collected users have been collected, but the main sampling method was simply to capture the Twitter stream of the gardenhose, which results in capturing only few tweets per user. For big data samples the information diffusion was simply studied by scanning the corpus for re-occurrence of the according link, while the rigor of actually the retrieving the retweets of a certain tweet was applied rarely.

[9] A major implication of this result is not limiting the retweet cascades in the research design of this work (see chapter 6.4).

Author	Title of paper	# Users	# Tweets	# Links / Type	Method of data collection	Diffusion of
Ratkiewicz et al. (2010)	Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams	-	305M with focus on 600k	-	1.5 month sample from Twitter stream	hashtags
Lerman and Ghosh (2010)	Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Networks	6.2M focus on 137k users[10]	398k[11]	3.9M / follower connections	Collection 398 memes from Tweetmeme[12]	retweets
Ediger et al. (2010)	Massive Social network analysis: Mining Twitter for Social Good	H1N1: 46k Atlflood: 2k Stream: 735k	H1N1: 3.5k Altflood: 279 Stream: 171k tweets	H1N1: 36k Altflood: 3k Stream: 1M	Collection of the Twitter stream	retweets
Choudhury et al. (2010)	"Birds of a Feather": Does User Homophily Impact Information Diffusion in Social Media?	465k	25M	836k / follower connections / @mentions / retweets	Crawl using a snowballing technique with a seed set of 500 users	URLs, hashtags, retweets
Hong et al. (2011)	Predicting Popular Messages in Twitter	2.5M	10M	-	-	retweets
Weng et al. (2010)	TwitterRank: Finding Topic-sensitive Influential Twitterers	7k	1M	50k follower connections	Crawl using a snowballing technique with a seed set of 1000 Singapore users.	-
Kwak et al. (2010)	What is Twitter, a Social Network or a News Media?	41M	106M	1.47B / follower connections	Entire crawl of the Twittersphere	retweets, hashtags
Romero and Kleinberg (2010)	Influence and Passivity in Social Media	450k	22M	1M / links based on same URL mentions	300h sample of Twitter stream	bit.ly-URLs
Cha et al. (2010)	Measuring User Influence in Twitter: The Million Follower Fallacy	55M with focus on only 6M users	1775M	1963M / follower connections	Entire crawl of the Twittersphere	retweets
Van Liere (2010)	How Far Does a Tweet Travel? Information Brokers in the Twitterverse	10k	13k	-	12h sample from Twitter stream	retweets

[10] Users were described as "active users " or users that retweeted a story at least once

[11] 389 memes x 1000 retweets

[12] For each meme up to 1000 retweets were collected. The Twitter API was used to download profile information (including friends and followers) for each retweet in the data set.

Yang and Counts (2009)	Predicting the Speed, Scale and Range of Information Diffusion in Twitter	3.2M	22M	dynamic network-based on @ -mentions	One month sample from Twitter stream	topics
Hansen et al. (2011)	Good Friends, Bad News Affect and Virality in Twitter	-	COP15: 207k / RAND: 348k	-	[random] collection of tweets from the Twitter stream	retweets
Suh et al. (2010)	Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network	-	74M with focus on 10k	-	Collection of tweets from the Twitter stream	retweets
Galuba et al. (2010)	Predicting Information Cascades in Microblogs	2.7M	15M	218M / follower connections	300h sample from the Twitter stream	URLs
Romero et al. (2011)	Difference in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags and Complex Contagion on Twitter	60M with focus on 8.5M users	3000M	50M @ mentions	Entire crawl of the Twitterverse	hashtags
Zaman et al. (2010)	Predicting Information Spreading in Twitter	7.3M	102M	50M / retweet connections	9 day sample of Twitter stream	retweets
Wu et al. (2011)	Who Says What to Whom on Twitter	524M with focus on 20k users	5000M	dynamic follower and retweet networks	223 day sample tweets from Twitter firehose + snowball sample of Twitter lists	URLs
Bakshy et al. (2011)	Identifying "Influencers" on Twitter	1.6M seed users and 56M total users	1.03B with focus on 87M with URLs	1.7B / follower connections (S2)	2 month sample from Twitter stream + snowball crawl of 1.6M seed users (S1)	bit.ly-URLs
Gayo-Avello et al. (2010)	Deretibus socialibus et legibus moment	see Gayo-Avello (2010)	see Gayo-Avello (2010)	see Gayo-Avello (2010)	see Gayo-Avello (2010)	-
Gayo-Avello (2010)	Nepotistic Relationships in Twitter and their Impact on Rank Prestige Algorithms	5M with focus on 1.8M users (S1)	27M	134M follower connections of (S1)	214 day sample from Twitter stream	-
Lee et al. (2010)	Finding Influentials Based on the Temporal Order of Information Adaption in Twitter	41M	223M	1.47B / follower connections	Entire Twitter crawl	retweets

Table 3.3: Overview of the reviewed information diffusion studies on Twitter used abbreviations: - = no data provided, k = thousand, M = million, B = Billion

General description of information diffusion in Twitter

Generally these studies reveal a lot of descriptive findings on information diffusion on Twitter, which shall be introduced now. The framework presented by Yang et al. (2009) serves as good reference to review the descriptive analysis of information cascades in Twitter. It describes information diffusion of tweets by three major properties: scale, speed and range. Scale is refers to the number of affected instances or users, speed to the duration until the first diffusion takes place, and range refers to the length or depth of the diffusion cascade.

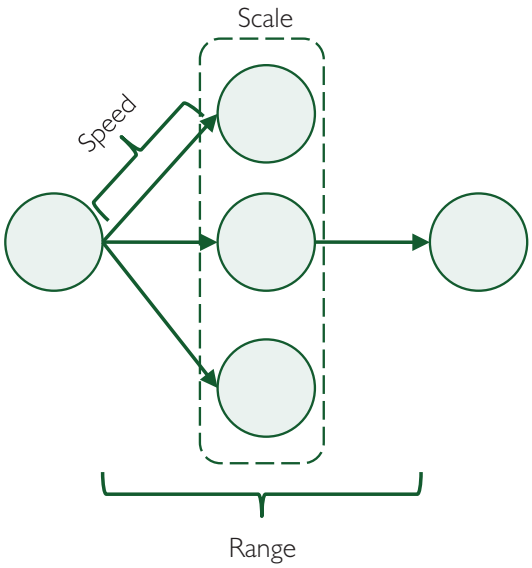


Figure 3.1: Three measures of information diffusion in Twitter according to Yang et al. (2009)

Scale

Scale is the most studied of the aforementioned dimensions and refers to how many Twitter users retweet a certain tweet. There are a number of descriptive results reporting on the scale and distribution of retweets in Twitter: Galuba et al. (2010) report that most tweets do not lead to any retweets and that the distribution of the retweets scale follows a power law. This observation is also confirmed by Bakshy et al. (2011) who report that most tweets are not retweeted and the distribution of retweets follows a power law. Zaman et al. (2010) also find that in their sample 99.8% of tweets were not retweeted. Hong et al. (2011) find that a large number of messages do not receive any significant retweets (less than 10) and that very few

messages receive more than 10000 retweets. Lerman et al. (2010) point out that the distribution of the number of times that particular tweets are retweeted shows a typical long-tail distribution where a high number of tweets are retweeted rarely and only a few tweets are retweeted more than 3000 times.

Speed

In a temporal analysis of retweets[13] Kwak et al. (2010) note that half of all retweets occur within one hour, 75% in under a day and only 10% of a month later. Galuba et al. (2010) also report that the diffusion delay between a tweet and a retweet has a median of 50 minutes. Lerman & Ghosh (2010) claim that it takes a day for a retweet cascade to saturate and that the number of retweets grows smoothly until it saturates. Lee et al. (2010) find that in comparison to aggregation sites like Digg[14], stories on Twitter show a much more slower rate of diffusion, though they also display a constant and organic rate of growth. They also note that information diffusion mostly happens in the early period, since the cumulative number of potential readers increases quickly in the first 5 days and then slows down over time. They note that writers contributing later to the topic might thus have less influence than earlier ones. Yang et al. (2009) point out that a topic might have different diffusion speeds at different time stages of its life cycle. They observe that the speed of a retweet varies over time versus being linear. They also note that the author's rate of being mentioned by other people and the time of day when a tweet is posted are important predictors on how fast a tweet on a topic will be retweeted. Similar to the observations of Lee et al., they note that for many cases in their data, earlier tweets contributing to a topic were more effective in producing retweets.

Range

Kwak et al. (2010) find that the retweet range (also called the retweet cascade/tree height) is of height one in 95.8% of the analyzed cases. In all of their analyzed retweet trees, 97.6% have a range of less than 6 and no tree had a range bigger than 11. Bakshy et al. (2011) also report that the range of cascades has a power-law distribution and most cascades have the range of one. Yang et al. (2009) report that more than half of the tweets failed to produce an offspring and less than 30% of tweets they examined had a range of two. Only 5% of tweets achieved retweet trees with a depth of five or more. Galuba et al. (2010) found that across all the cascades the maximum distance-to-root falls off exponentially, and the average distance falls off even faster. One hypothesis for the short range of retweets is that when a user receives an interesting URL along a path longer than 1, that user is very likely to start following the original source of the

[13] Regarding the temporal analysis of *hashtags* they found that 73% of trending hashtag topics have a single active period, which is shorter than a week. Most of such periods are only one day long and only 7% of periods are longer than 10 days.

[14] Digg is a social news website which has the functionality of promoting a story `http://digg.com/`

URL and thus shortening the potential future path to one hop. Yet when Lerman et al. (2007) compared the spread of Digg stories with those in the Twitter social network they found that retweets in Twitter generally penetrated deeper into the network than Digg stories.

After highlighting the different types of entities that are diffused through the Twitter network, and showing that their diffusion process can be examined among the dimensions of *scale, range and speed* the next sections will now focus on the review of studies that cover the diffusion of retweets as an entity, and mostly observe the scale of the resulting diffusion. In order to make the most use of these contributions they will be categorized according the structural, relational and cognitive dimensions of social capital of Nahapiet's model: Research outcomes dealing with the structural position of the actor in the network will be described under the structural dimension of information diffusion. Research outcomes dealing with the user's activities (e.g., number of @replies exchanged, number of retweets received and sent) are reviewed as part of the relational dimension. Finally research dealing with the actual content of tweets (including message factors like sentiment, number of hashtags, links etc.) and general notions of homophily will be reviewed as part of the cognitive dimension of social capital.

3.4.1 Structural factors affecting information diffusion

Friends & Followers

A variety of studies mention the influence of structural centrality in the Twitter network on how often a person's tweets are retweeted. On one hand some authors report its weak influence: Romero et al. (2010) found that the number of followers is a weak predictor of the retweets of a certain tweet (which they measure by the number of clicks on the URL in the tweet). They also computed the PageRank of those who tweet a certain URL and found that it does not correlate well with the number of clicks the URL will get. Cha et al. (2010) also reported that users who have a high indegree (i.e. lots of followers) are not necessarily successful in terms of spawning retweets. On the other hand there is also a number of studies that report that the structural centrality in a network is a good predictor of retweets: Hong et al. (2011) tried to predict the number of retweets that a user receives and found that among the analyzed factors the retweet probability is mostly influenced the number of followers a user has. Suh et al. (2010) found that the number of followees and followers are strongly predictive of retweet probability. They hypnotized that the larger the audience, the more likely the tweet gets retweeted and accordingly that the more sources the user follows the more interesting the user's tweets are and hence they will be retweeted more. Bakshy et al. (2011) question those observations since in their study they found that the number of friends or followers is not a predictive feature for large retweet cascades in their model. They point out that although there is a tendency that large cascades tend to be driven by previously successful individuals with many followers, the model fit in their data is relatively poor due the extreme scarcity of such cascades. This means that most individuals with these attributes on average are not particularly successful

either. Although Galuba et al. (2010) found that URLs tweeted by the highly connected users reach a large audience and are likely to be retweeted by their followers they challenge question of causality: A user's URLs might be retweeted more often because they have many followers, but they have might many followers because what they tweet tends to be interesting and viral.

Opinion Leaders and Influentials

Structurally directly related to the number of followers and friends is the analysis of higher level network effects, which is also known as a measurement of influence. A question raised in this setting is whether or not the extent of diffusion can be maximized by seeding a piece of information or a new product with certain key individuals, whose connectivity or position in the network allows them to generate a large "multiplier effect" by triggering a cascade of retweets (Kempe, 2003; Watts, 2002). The prospect of identifying such individuals called "super-spreaders" (Watts, 2002) in the epidemiological literature and "influentials" (Kozinets et al., 2010) by marketers and "opinion leaders"(Katz and Powell, 1955) in the media theoretical field has aroused great interest in recent years. Unsurprisingly a significant body of literature on Twitter and information diffusion is trying to answer the question how to determine the most influential Twitter users. Ranking users in Twitter has a strong tradition and there are a number of commercial services that compute "ad hoc" scores (e.g., Klout[15] score, PeerIndex[16] score), which provide a glimpse into the influence or authority of a given Twitter user. Yet the actual details for such commercial scores are often undisclosed. In contrast to the undisclosed commercial services, the below presented studies allow to review which structural network metrics (and algorithms based on these metrics) are currently used in academia to identify such opinion leaders or influentials in Twitter. These studies also reveal which types of Twitter users or interest domains are considered as influential and so dominate the information diffusion on Twitter. An in depth survey of the ongoing academic discussion of what influential means in Twitter can be found in the work of Gayo-Avello et al. (2010) where he offers a comprehensive overview of feasible algorithms for ranking[17] users in social networks.

- Ediger et al. (2010) present a very simple influence score by simply computing the indegree on a crawl of the whole Twitter network. They note that in their analysis they found only few high degree ("broadcast") actors to which Twitter users tend to attach to. Examining some of the top actors in their Twitter dataset they find that the high indegree group is dominated mainly by major media outlets and government organizations. Their interpretation is that Twitter's user network structure is defined by news dissemination and users track topics of interest from major sources and occasionally retweet that information. Information flows one way: From the broadcast hub out to the users.

[15] http://klout.com/

[16] http://www.peerindex.com/

[17] He also examines their vulnerabilities to linking malpractice in such networks and suggests an objective criterion against which to compare such algorithms.

- Kwak et al. (2010) were among the first to comparatively identify influentials on Twitter by ranking users by the number of their followers, their retweets and the PageRank metric. They found that rankings created according to followers and PageRank are highly similar but a ranking by retweets gives a different ordering. When ranking users by retweets they mostly found mainstream new media accounts such as the breaking news wire, ESPN sports news, the Huffington post or NPR news. They indicate that the popularity of such news outlets can be a result quality of the material of these media outlets, yet leave this question empirically unanswered.
- Another highly cited measure is the TunkRank[18] which was proposed by Daniel Tunkelang (2009) in 2009. The TunkRank is based on three assumptions: Each user has a given influence which is a numerical estimator of the number of people who will read that user's tweets. The attention that a user pays to his followees is equally distributed and if a user reads a tweet from his followees he will retweet it with a constant probability.
- Cha et al. (2010) have measured the user's influence by comparing three different measures: Number of followers, retweets and mentions. While the most followed users were news sources (e.g., CNN, New York Times), politicians (e.g., Barack Obama), athletes (e.g., Shaquille O'Neal) as well as celebrities like actors, writers musicians and models (e.g., Ashton Kutcher, Britney spears), the most retweeted users were content aggregation services (e.g., Mashable, Twittertips or Tweet-Meeme), businessmen (e.g., Guy Kawasaki) and news sites (e.g., The NY times, The onion). The most mentioned users were mostly celebrities. They found a strong correlation in their retweet influence and mention influence scores. This means that users who get mentioned often also get retweeted often and vice versa. Indegree however was not related to the other measures.
- Weng et al. (2010) proposed to compute a topical Page Rank which they called the TwitterRank. It differs from a normal PageRank in that way that the random surfer performs a topic-specific random walk, i.e. the transition probability from one Twitterer to another is topic specific. This allows to measure a Twitterers influence by taking into account the topical similarity among Twitter users as well as the link structure. They found that news channels dedicated to a certain topic (e.g., "singaporenews") are among the influential users in their corresponding topic group. However, they also report that the most influential users can hold significant influence over a variety of topics by counting only the retweets and mentions a user spawned on the given topic. Additionally, they point out that influence is not gained spontaneously but through an ongoing effort such as limiting tweets to a single topic. As an example they point to users who talked about only one news topic (i.e. for Iran: iranbaan, oxfordgirl, TM_outbreak) and then suddenly became popular over the course of the event.
- Lee et al. (2010) proposed a novel method to find influentials by considering both the link structure and the temporal order of information adoption in Twitter. Their temporal

[18] Highly influential people according to the TunkRank can be found online under `http://tunkrank.com/score/top`

approach considers the aspect that for example web pages, can only link to already existing web pages and thus old information weighs rather more than new information. They note that when comparing the number of followers and the number of effective readers of each user (which are users that are newly exposed to the topic of a tweet a user writes) they found that for 80% of the users only 20% of their followers turn out to be effective readers. Similar to other researchers they come to the same conclusion that having many followers does not always make a user influential. According to their influence score the most influential users are news media like CNN breaking news, the NY times or E! online.

- Romero et al. (2010) introduced an influence rank (IP-influence) based on the assumption that in order for individuals to become influential they must obtain attention and overcome user passivity. They created a relative influence score and a passivity score for every user which reinforce each other. The passivity of a user is a measure of how difficult it is for the other users to influence him. The influence of a user depends both on the quantity and the quality of the audience he influences. They found that the group of users with the most IP influence in the network is dominated by news series from politics, technology and Social Media (e.g., Mashable, Google, The Onion). These users post many links which are forwarded by other users, causing their influence to be high. Users that are followed by many but have a relatively low influence were people that are very popular (e.g., Britney Spears, Ashton Kutcher) and have the attention of millions of people but are not able to spread their message very far. In most cases their messages are consumed by their followers but not considered important enough to forward to others. The most influential people with few followers (less than 100.000) were local politicians and political cartoonists, who created content that was of importance for a local region.
- Gayo et al. (2010) provide a new method of defining influence based on a physical metaphor of velocity (velocity score). In comparison to the existing methods they propose a method where the user graph can be greatly disregarded and only user mentions can be used to compute an accurate and dynamic of influence score for Twitter users. They reported that in comparison to the most of the commonly applied scores (such as the number of followers, PageRank, TunkRank, or Influence-Passivity) their algorithm exhibited the highest correlation with web site visits. They come to the same conclusion that influential users are often news organizations or users that are dominating certain events and topics.

Despite the slightly different approaches all studies surprisingly come to the same conclusion that news media corporations, government organizations, celebrities or social media accounts seem to be highly influential. A question that is left unanswered is how these accounts fit into the bigger picture of the Twitter network structure. Can those actors all be found in the center of attention, or do they occupy different regions of the Twitter ecosystem.

Comparing the rigor and insights of the study of opinion leaders and influentials on Twitter with the existing literature on opinion leaders in the media centric field, it can be subsumed

that Twitter studies reduce opinion leadership to a structuralist approach relying on network measures. Regarding the applied methods, it can be subsumed that the majority of the structural measures presented, rely to a high degree on the measure of network centrality such as eigenvector centrality or the PageRank (Page et al., 1998) centrality. Twitter studies on opinion leadership refrain from using questionnaire based methods which seems only plausible given the large and accessible corpora. On one hand this structuralist approach is able to produce insights at a much larger scale, on the other hand this scale is traded against precision and qualitative insights on opinion leadership. Additionally comparing those insights to the existing research on opinion leaders and social capital one finds that the discussion on how opinion leaders reflect norms of a community is underrepresented or almost non-existing. Although the presented studies confirm the observation that opinion leaders accelerate diffusion in a social group, often there is a lacking discussion on how that opinion leadership might only be limited to a topical domain.

Brokers and Two-Step-Flow of Information

The structural model of the Two-Step-Flow of communication is only covered in the study of Wu et al. (2011) where they studied the production, flow and consumption of information. They found that in their dataset roughly 50% of tweets consumed are generated by only 20.000 "elite" (highly listed) users, where the media produces the most information but celebrities are the most followed. They found that users classified as "media" easily "out-produce" all other categories: While ordinary users originate on average only about 6 URLs per week in their data set, media users produced 1000 URLs on average. Although they found that media outlets are by far the most active users on Twitter, only about 15% of tweets received by ordinary users are received directly from the media. To investigate the Two-Step-Flow of information among users they sampled 1M random "ordinary" users and found that 46% of URLs that they received was via non-media friends. This indicates that content from media outlets reaches the masses via an intermediary rather than directly. Van Liere (2010) is the only work among the reviewed literature who identifies different information diffusion patterns not only depending on the structural centrality of a user but also on local and global information brokerage. Yet his contribution is in a preliminary state and only suggests that information diffusion might be driven by information brokerage patterns.

3.4.2 Relational factors affecting information diffusion

Studies on information diffusion in Twitter have also explored how direct interaction or tie strength between users affect information diffusion in the network. The studies below provide a state of the art on how relational factors influencing information diffusion in this dimension.

Amount of tweeting

The first and most simple form of undirected relational interaction is the simple act of tweeting. Here numerous studies found that simply writing tweets is not correlated to becoming retweeted: Suh et al. (2010) found that among other features the amount of past tweets or the account age and the number of favorites are not significant in predicting retweets. Thus broadcasting more often to ones audience in Twitter does not necessarily lead to greater engagement. Bakshy et al. (2011) also confirmed that the number of tweets is not a predictive feature. However, other researchers (Yang and Counts, 2009) found that the activity level of a person in terms of number of tweets produced can be a significant predictor for retweets. There are numerous studies that have provided descriptive analyzes on the amount of interaction that takes place in Twitter: Harvard Business School researchers[19] analyzed a random sample of 300.000 Twitter accounts and found that the top 10% of prolific Twitter users accounted for over 90% of tweets. Sysomos[20] analyzed 11Mio. Twitter users and found that most people post only once per day. Galuba et al. (2010) note that how frequently users tweet, is distributed via a power law, with the majority of users tweeting less than 10 times (or never) and only a few users producing more than 10.000 tweets per week. Kwak et al. (2010) located a correlation between the number of followers a user has and how often he or she tweets. In fact, the majority of users who have fewer than 10 followers have never tweeted. This means that there is a high variability in the relational component that can be expressed between persons, suggesting that most relations on Twitter can be expected to be weak ties.

Amount and reciprocity of interactions

While some researchers such as Suh et al. (2010) found that user mentions have a negative association with retweeting, the majority of findings report that being mentioned has a strong influence on being retweeted: Yang et al. (2009) found that the rate with which a user is mentioned is a strong predictor of retweets. They show that the amount of how often a person is mentioned is the best predictor for more retweets and longer diffusion chains across topics and that this feature accounts for the majority of the variance. Cha et al. (2010) found that the correlation of user mentions and retweets is generally high (above 0.5), meaning that popular users who are good at spawning mentions from others might be able to trigger retweets. They report that this ability holds even for a wide range of topics for highly mentioned users. The reciprocity of the network was also evaluated as a factor of retweets in the network: Yet on one hand Kwak et al. found a low level of reciprocity between users, where 77.9% of user pairs were connected in an one way direction, and only 22.1% if users had a reciprocal relationship. Cha et al. (2010) also reported a low reciprocity of 10% in their dataset. On the other hand Weng et al. (2010) claimed that they found a much higher level of reciprocity when they

[19] Bill Heil and Mikolaj Piskorski, "New Twitter Research: Men Follow Men and Nobody Tweet" (Harvard Business Blog 01.06.2009)

[20] An in-depth look inside the Twitter world (June 2009) `http://sysomos.com/insideTwitter`

computed the correlation between the number of friends and the number of followers for each Twitter user in their dataset. They reported that 72.4% of the users follow more than 80% of their followers and 80.5% of the users are followed by 80% of the users they are following.

Amount of re-tweeting

Finally there are mixed findings on the influence of the number of received retweets as a predictor of being retweeted: Romero et al. (2010) found that the number of retweets (the number of times a user has been credited in a retweet) is a poor predictor of the maximum number of clicks a certain link will receive in a new tweet. Contrary to this finding Bakshy et al. (2011) computed an individual-level influence as the average size of all cascades for which that users was a seed. They report that among a variety[21] of features that they included in their model they found that the past performance (in form of received retweets) provided the most informative set of features. Hong et al. (2011) also report that the probability that a message will be retweeted depends whether the message has been retweeted before. They contemplate that highly retweeted messages unlike non-retweeted messages receive tens of thousands of additional viewers, while normal non-retweeted messages only have a small audience (i.e. only the followers of the author). Zaman et al. (2010) report that the most of the predictive power in their model resulted from how often a certain Twitter user has been retweeted before or retweeted others before. Bakshy et al. (2011) note that individuals who have been retweeted in the past and who have many followers are indeed more likely to be retweeted in the future; however this intuition is correct only on average[22], since in the observed data large retweets cascades happen so seldom that it is actually impossible to predict them with a relevant accuracy.

3.4.3 Cognitive factors affecting information diffusion

The cognitive dimension of how a shared language or a shared vision influences information diffusion on Twitter has also been explored. Researchers have studied either the influence of message content or the homophile tendencies of users on information diffusion.

Message Content

A variety of researchers have analyzed the influence of "structural" message content of tweets on their information diffusion. By structural message content we understand all structural

[21] (1) number of followers, (2) number of friends (3) number of tweets (4) account age (5) past influence

[22] They suggest that marketers rather than attempting to identify exceptional individuals, they should instead adopt "portfolio-strategies" which target many potential influentials at once and therefore rely only on average performance.

elements of a message such as URLs, hashtags, acronyms, or @signs. Yang et al. (2009) note that tweets that contained mentions and URLs were are good predictors of longer retweet chains. Suh et al. (2010) also analyzed if certain message features are able to predict the number of retweets: The message features they used were: number of URLs in a tweet, number of hashtags in a tweet, number of user names specified in a tweet. They found that hashtags and URLs have a significant positive effect on retweet probability. They also noticed that if the URL contained specific news media domains (e.g., twitter.com, mashable.com, nytimes.com) this leads to higher retweetability of tweets. A similar observation is made by Wu et al. (2011) who found that different types of content exhibit very different life spans: In particular, media-originated-URLs are disproportionately represented among short-lived URLs while those originated by bloggers tend to be overrepresented among long-lived URLs. Finally they found that the longest-lived URLs are dominated by content such as videos and music which are continually being rediscovered by Twitter users and appear to persist indefinitely. Kwak et al. (2011) found that 80% of their analyzed hashtags fall into the category of headline news which have a time span of less than a day. They also found that long-lasting topics with an increasing number of tweets do not always bring in new users into the discussion: For e.g., #iranelection there is only a set of core members generating many tweets over a long time period. Regarding the message content, Cha et al. (2010) found in their study that retweets are driven by a high overlap of the tweet contents with the interests of the community where tweets limited to a single topic had a higher chance of receiving retweets. Hansen et al. (2011) have investigated the influence of message sentiment on the retweet probability of a tweet. They found that negative sentiment enhances the probability that a tweets gets retweeted in the news segment, but not in the non-news segment. They concluded that in short "if you want to be cited: sweet talk your friends or serve news to the public"[p.2].

On the other hand there are also researchers that found that the message content has no effect on the retweet probability: Zaman et al. (2010) found that including any form of actual message contents of the tweet seems to lower prediction accuracy in their predictive model. Bakshy et al. (2011) measured the influence of content among low, mid and highly[23] retweeted URLs. On one hand found indications that content that is rated[24] as more interesting and content that elicits more positive feelings tends to generate larger cascades on average. They also found that shareable media tend to spread more URLs associated with news sites, while some types of content (e.g., "lifestyle") spread more than others. On the other hand when they included all of this information in a their regression model which previously only included structural features they found that none of the content features were informative relative to the seed user features, nor was the model fit improved by the addition of the content features. Their elaborated study indicates that indeed there is no "silver bullet"[p.3] making content sticky or

[23] They bin retweeted URLS in 10 logarithmic bins (in terms of cascade size) and where each bin contains 100 URLs

[24] Using Amazons Mechanical Turk they asked users to go to the web page of the URL and classify it in one of 10 categories and asked them how relevant the site was on a scale from 0 to 100. Additionally they asked users how interestingly the site was, how interesting the average person would find it and how positively it made them feel. Finally they asked them if they would share the URL using social networks.

viral. Their explanation for this observation is that most explanations of success tend to focus only on observed successes which invariably represent a small and biased sample of the total population of events.

Influence of homophily on information diffusion

Choudhury et al. (2010) have investigated the impact of user homophily on information diffusion in online media. They consider four different attributes of homophily associated with the users (1) same location of users, (2) same information roles[25] of users (3) same content creation roles[26] of users and (4) same activity behavior of users i.e. the distribution of a particular social action over a certain time period. They found that homophily indeed impacts the diffusion process; however the particular attribute that can best explain the actual diffusion characteristics often depends upon (1) the metric[27] used to quantify diffusion and (2) the topic under consideration: The location based homophily among users seems to predict well the reach and speed of retweets, because users' local social neighborhoods are often clustered around commonality in their location. They find that the diffusion on the topic of politics takes place extensively over the location attributes of users. While the attribute "information roles" performs poorly for topics like politics it works extremely well for topics such as technology, where users connect with similar responsive users.

To test for topic specific homophily Wu et al. (2011) classified Twitter users into "elite" and "ordinary" users based on how often they are mentioned on thematic specific Twitter lists. Highly listed users were then categorized into four categories of media, celebrities, organizations and bloggers. They also found significant homophily within categories by computing the volume of tweets exchanged between elite categories. Celebrities overwhelmingly pay attention (in terms of how many URLs are received from this category from other categories) to other celebrities, media actors pay attention to other media actors and so on. When analyzing how much information that originates from each category is retweeted by other categories, they found that retweeting is also strongly homophile among elite categories. However bloggers are disproportionately responsible for retweeting URLs originated by all categories. They find that this result reflects the characterization of bloggers as recyclers and filters of information.

Romero et al. (2011) modeled the spread of information by simulating retweet cascades on the real network of users. Analyzing different hashtag categories such as celebrities, games, idioms/memes, movies, music, politics, sports and technology they found that e.g., celebrities are significantly less successful in generating big cascades, while political and idiom/meme are significantly better at generating big cascades. In the set of the 500 most-mentioned hashtags

[25] Consisting of three categories of roles: "generators", "mediators" and "receptors".

[26] They distinguish between "meformers" (users who primarily post content relating to self) and "informers" (users posting content about external happenings) (see usage types of Twitter above)

[27] They introduce different metrics, which describe how well homophily describes the diffusion along categories such as user-based (volume, number of seeds), topology-based (reach, spread) and time (rate).

they observed that the average number of people that have already mentioned a topic and the chance that a given user mentions the topic too is highest when two, three or four users in their network have mentioned it. After that point the chances for a user to mention this topic fall off dramatically. Thus if a user does not adopt an idiom after a small number of exposures the chance they do so falls off quickly. The researchers also found that for political hashtags the persistence (number of exposures needed) to adopt a topic has a significantly larger value than the average. For other hashtags which they call Twitter idioms (such as #followfriday, #musicmonday ...) they found that the persistence is unusually low. On the other hand only relatively few people used political and games hashtags, but each one of them used them many times, making them the most mentioned categories.

3.5 Conclusion and goals

This literature review has led to three major conclusions and has helped define the four goals of this thesis. These will now be highlighted.

The first conclusion is that diffusion theories supplement well the social capital theory: they provide a more nuanced view of the dimensions of social capital that drive information diffusion and knowledge exchange. The diffusion theories reviewed in this thesis show that there is an urgent need to combine the social capital theory with diffusion theories. Indeed, as Widen-Wulff puts it: "The relation between information [sharing] behavior and social capital has seldom been realized [...], the majority of the studies have concentrated on individual information [sharing] behavior. It is obvious that the social capital dimensions can provide a useable framework explaining how group resources which are available in individual social settings, can contribute to our understanding of knowledge sharing mechanisms"[p.452] (Widen-Wulff, 2004). The valuable insights from the diffusion theories explain why the three dimensions of social capital provide an excellent ground for the formulation of meaningful theoretically grounded hypotheses on information diffusion. These hypotheses will be presented in chapter 5.

The second conclusion of this study stems from the systematic review of literature on information diffusion in the Twitter network. The review of descriptive studies that analyze the retweeting behavior in Twitter revealed a number of insights into the information diffusion process in general. Indeed, the descriptive studies provided a more exact picture of the exchange process: they showed that information diffusion in Twitter can be described as a small series of temporally-limited diffusion cascades, rather than as a "wild forest fire"[p.1] (Bakshy et al., 2011). However, even these cascades are very rare: most retweets are not retweeted at all (Bakshy et al., 2011; Galuba et al., 2010), yet if they are, it is usually under less than one hour (Lee et al., 2010; Lerman, 2007). After one hour, the chances of a tweet being retweeted fall dramatically (Yang and Counts, 2009). The resulting retweet cascades are in most cases compromised of the direct contacts of an author (Kwak et al., 2010; Lerman, 2007).

The third conclusion is that the reviewed studies greatly contributed to the understanding of the positive influence of the three dimensions of social capital on information diffusion in Twitter: first, at the structural level, the review revealed that different structural network metrics in Twitter are able to provide good measures of individual bonding social capital and can predict the positive influence of bonding social capital on information diffusion (Tunkelang, 2009; Weng et al., 2010; Romero and Kleinberg, 2010; Gayo-Avello, 2010). Several studies have shown that the number of interactions an author maintains in the network (Cha et al., 2010; Yang and Counts, 2009; Zaman et al., 2010), or the number of retweets the author has previously received (Romero and Kleinberg, 2010; Zaman et al., 2010), are valuable measures of tie strength and capture well the relational dimension of social capital at an individual level. Finally, the content of the tweet (Cha et al., 2010; Wu et al., 2011) and the homophily of the network (Choudhury et al., 2010; Wu et al., 2011) can adequately measure the cognitive dimension of social capital.

The literature review has revealed a number of research gaps that will be addressed in this work.

Framework: There is a very fragmented body of literature concerning information diffusion in Twitter. Although the literature contains very relevant contributions to the study of information diffusion on Twitter, it does not possess a unified framework that could show how the various contributions fit together. Although the volume and speed of research on Twitter is enormous, a majority of the contributions simply neglect the existing theory on information diffusion and, instead, provide "data-driven" approaches. Numerous papers focus on providing new influence measures for Twitter users, yet only a minority of studies connect their work to the existing discussion on the impact of influentials, opinion leaders or user roles on information diffusion. Only one study (Wu et al., 2011) explores theories on the effect of brokers on information diffusion. A multitude of papers explore the influence of only one particular variable on information sharing behavior and neglect the influence of other variables. The literature review also revealed that numerous studies find contradicting evidence in terms of the factors that influence information diffusion.

This gap in the literature reveals the need for a *multidimensional framework*. Indeed, the study of knowledge exchange in virtual communities needs such a framework because "knowledge sharing is a multidimensional activity"[p.452] (Widen-Wulff, 2004). Social capital can provide a multidimensional framework and is capable of integrating the numerous observations and challenges that surround the diffusion process. It serves the strong need "to develop a platform [or framework] for the different contexts of knowledge sharing mechanisms, different social capital dimensions, different measures and information behaviors"[p.455]. Indeed, it is no surprise that the numerous studies on online social capital in the information systems domain (see section 2.2.3) have used the social capital framework to explore information exchange in virtual communities. However, it is surprising that for virtual communities such as Twitter, that (a) provide an abundance of unobtrusive data, (b) allow measurements that are not influenced by social desirability, (c) provide highly accurate relational data, and (d) thus offer many

network-based metrics, there is no concept or operationalization of a network-based online social capital framework to explain information diffusion. A multidimensional social capital concept is not only useful at providing the theoretical grounding that was needed to organize the reviewed observations. It will also provide increased insights into which dimensions of social capital actually influence information diffusion in Twitter. Therefore, the next chapter will focus on the conceptualization and operationalization of such a framework.

Group level view: The review of Twitter literature on information diffusion reveals that the majority of research deals with information diffusion at an individual level. The lack of concrete observations of information diffusion at the group level stands in stark contrast to the amount of theoretical studies on information diffusion at the group level. This gap has recently been highlighted by researchers (Iyengar et al., 2010) who claim that new data is needed to track the diffusion of an idea, innovation or information through multiple communities (within their respective network structures). Using the network theory of social capital will help address both the individual and group levels. Indeed, at the individual level, this work will investigate the interactions between networks members, considering them to be the basic building blocks of community. At the group level, it will also study the group network where community emerges and is perceived to exist. Finally at the group level it will also study the information diffusion between groups.

Nuanced view on tie concepts: The review has shown, that there is a very fragmented reasoning about different social tie concepts affecting information diffusion: In the majority of diffusion research we find either a) a lack of networked views on the diffusion process or b) a wide mix of tie concepts in the literature review on Twitter. While most studies use the simple friends- and follower network to define social ties, only a few works actually employed the rigor of defining ties and their strengths (Huberman et al., 2009). Most works have been treating connections between individuals in a relatively categorized and binarized manner and the discussion on what is transmitted over such connections has only been marginal so far. Reputable network researchers (Valente, 2010) postulate that more nuanced view in this area is needed. Such a definition of ties will indeed be a cornerstone in the operationalization of network-based online social capital in the next chapter.

Cognitive view: The final gap in literature deals with the limited insights into how shared cognitive references between individuals can lead to information diffusion. While most works have studied the influence of the message structure or the influence of homophily on information diffusion, the majority of contributions have simply neglected the influence of user interests on information diffusion. Only two studies have provided a glimpse into how this factor might influence information diffusion. This question seems particularly important in the context of a social network whose slogan is “follow your interests”. The study of the cognitive social capital in interest-based communities will therefore shed light on this area of research.

In order to fill these research gaps, a social capital framework for Twitter will developed in the following chapter. It will contain individual and group perspectives on information diffusion. It will also explicitly conceptualize the structural, relational and cognitive dimensions of social

capital for each of these perspectives.

4 A social capital framework for Twitter

The goal of this chapter is to create a network-based online social capital framework for Twitter. This framework must fulfill three conditions, which will be used to structure this chapter: first, the social capital framework must make good use of the available data traces that Twitter has to offer, as it will not be based on a questionnaire design. The first part of this chapter will thus present the different data traces that are accessible in the network and interpret their meaning in terms of their usability for the framework. The second part of this chapter will focus on describing how data traces in Twitter can be described in a network-based perspective. The network based perspective will not only motivate how the social capital dimensions can be expressed as networks, but it will also show how the cognitive dimension of social capital is the basis for all group concepts and explain how bonding and bridging social capital can be distinguished. Third, the review of the social capital theory has shown that the data traces in Twitter must incorporate the cognitive, relational and structural dimensions of social capital, in order to represent Nahapiet et al.'s (1998) well-known social capital model. The main part of this chapter will therefore describe how to use the available data traces to express the cognitive, relational and structural dimensions of social capital.

4.1 Types of data traces in Twitter

Online social networks have not always been easy to study due to the private nature, inaccessibility, volatility and largely heterogeneous character of these networks (Boyd and Ellison, 2008). However, with the introduction of new popular online social networking sites such as Flickr, YouTube, Facebook or Twitter, things have changed. Although all of these networks seem to be promising candidates for the study of online social networks, they differ in terms of their accessibility and the richness of their available data. Networks like YouTube or Flickr are not particularly well-suited for the analysis of online social capital because they are content-oriented, and there is only little communication, interaction and diffusion of information between the users. However, networks such as Facebook and Twitter are more interesting since users interact with each other and since information is forwarded and shared between users. Yet, of these two networks, only Twitter allows researchers to publicly study the diffusion process of information. While in Facebook most communication between members is meant to be private and cannot be accessed by third party members, communication between Twitter users is public in nature and studying it does not violate users' privacy. A number of entities or data traces on Twitter can be collected, studied and combined to form a

framework of network-based online social capital. All of the entities described in this section are data traces that are freely accessible on Twitter either through the website or through a dedicated application programming interface (API). From a researcher's perspective, these entities provide an important data foundation that can be used to create a framework of online social capital.

Twitter users

Twitter is one of the few social network websites that allows Internet surfers or researchers to access their data without signing up. The majority of other similar websites (e.g., Facebook) cannot be accessed without prior registration. However, to participate, prospective Twitter users have to register with the site under a given user name. Generally, Twitter users provide some personal information (e.g., their birthday, place of residence, interests, etc.) that is then saved under the user's profile, which can be publicly accessed unless the user has chosen to make this information private. A study from the Pew Internet Project (Smith and Brenner, 2012) from 2012 reports that 26% of people aged 18-29 years and younger use Twitter, and that 15% of online adults use Twitter. After analyzing a set of 1.8 million users, Gayo-Avello (2010) found that the average Twitter user is between 15 and 29 years old and that users in this age range account for about 70% of the Twitter user base. He reports that 59% of users are men and 41% are women, though he claims that the male predominance in Twitter has "its days numbered because females surpass male users among the youngsters (10-19 year olds)"[p.40]. Java et al. (2009) found that the Twitter network is mostly popular in the USA, Europe and Asia (predominantly in Japan), with the most popular cities being Tokyo, New York and San Francisco. They noted that the probability of friendship between two users is inversely proportionate to their geographic proximity. In other words, there is a tendency toward geographical homophily, such that the majority of links are between, for example, Asia and Asia or North America and North America.

Links between followers and followees

Since Twitter is a social network, its users are linked together through networks of ties. However, in comparison to other social network sites, Twitter users can link to any other user without that user's approval (sites such as Facebook require the direct approval of the source and the target before a link is created). In general, the links between Twitter users are visible to those visiting a particular profile (as with "friends" on Facebook), but privacy settings allow users to hide their contacts if they wish to do so. Moreover, in contrast to other social networking sites, Twitter distinguishes the direction of the link. Users can search for people they are following (followees) and then be notified if those people have written a new message. However, a user who is followed by other users (followers) does not necessarily have to reciprocate. This makes Twitter a social network with directional relationships. An

overview of the different types of possible links is provided in figure 4.1. Since these types of connections are the baseline for all social relations in Twitter, they will be reviewed in detail in the social capital framework section 4.2 below.

Groups or lists

A variety of social networking sites allow users to create and join special interest groups (e.g., LinkedIn, Facebook). Users can post messages to groups and upload shared content. Groups usually differ based on whether they are moderated or unmoderated, and whether they are public or private. Some groups can be managed by multiple admins (as on Facebook), while others can be managed by a single person who decides who can be part of the group or list. While Twitter does not have an explicit notion of groups, every user is allowed to create 20 lists containing up to 500 users. Lists can follow users which is the equivalent of adding a user as a member to a list or be followed by other users (for example, user A is following list D in figure 4.1). Users that follow a certain list receive updates from all the individuals that are listed on such a list. Lists help Twitter users create meaningful collections of users that are discussing the same topic, that live in the same region, or have the same hobbies or interests[1]. Twitter lists have been successfully used to determine users' interests (Kim et al., 2010; Wu et al., 2011; Weng et al., 2010), to detect communities (Greene et al., 2012; Wu et al., 2011; Garcia-Silva et al., 2012) and to improve faceted search (Yamaguchi et al., 2011).

Messages or tweets

On social networking sites, users usually communicate through an internal messaging system. While certain messages can be made private and are only visible to the recipient (e.g., "direct messages" on Twitter and "private messages" on Facebook), Twitter users communicate most often through public messages that are distributed to all of their followers. These messages are called tweets. This broadcasting mechanism is a central way of disseminating information on Twitter. It is comparable to sending out emails to one's entire email contact list. While some sites (e.g., Facebook) lets users narrow down the list of recipients that will see the status message, on Twitter the message is always sent out to all of the user's followers. Furthermore, Twitter messages are limited to 140 characters (comparable to text messages on a mobile phone) in order to make their consumption easier. If a user composes a tweet from a mobile phone, Twitter also provides the geographical information of the user (if the user has permitted the disclosure of this information). This feature is similar to the geographical information provided on other networks (e.g., on Facebook, users can "check" in to predefined places, indicating their actual presence at a location).

[1] As an example, this action corresponds to users on Amazon creating a "best horror books" list and adding various books of the same genre into this list.

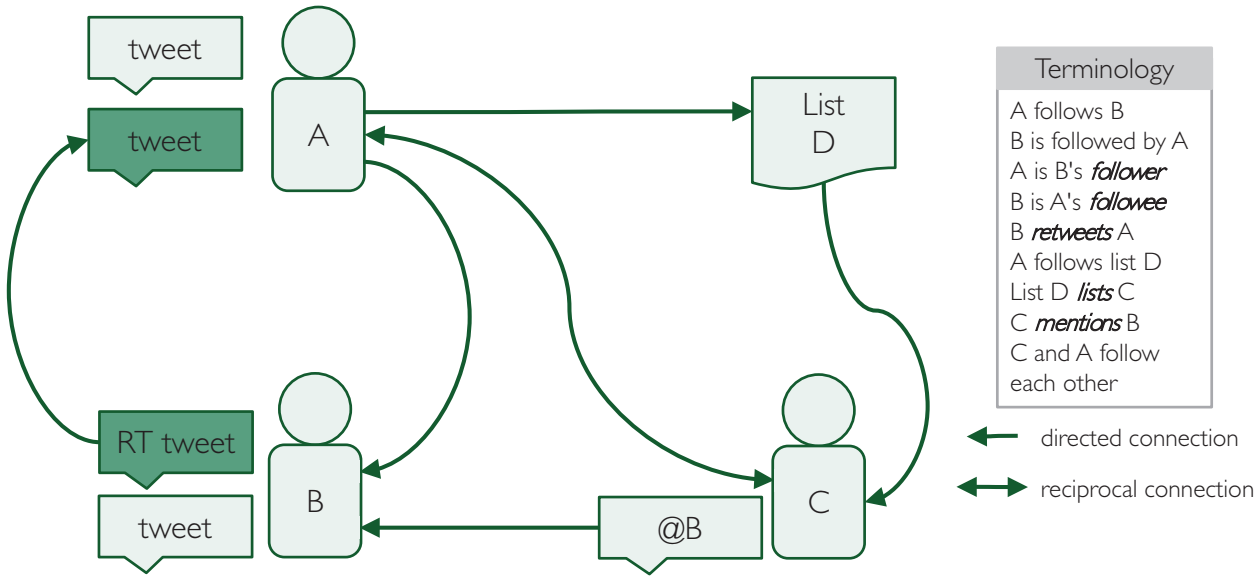


Figure 4.1: An overview of Twitter's specific features.

While people can always interact with Twitter directly through the website, as of 2012, a myriad[2] of third-party applications allow users to interact with the network indirectly, using different devices and applications. Early on, Twitter provided an application programming interface (API) which allowed developers and researchers to access all of the various entities (users, tweets and lists) in the network: this allowed Twitter to become an extensive one-of-a-kind online ecosystem. Krishnamurthy et al. (2008) note that 80% of tweets come directly from the web interface, but 20% come from applications using the Twitter API. Tweets may also contain hyperlinks to web pages and thereby reference content outside of Twitter, such as newspaper articles and blog entries. Banerjee et al. (2009) analyzed 21 million tweets from the public Twitter timeline and found that 80% of content-filled tweets are used to express the users' interests in real time through micro blogs.

Tweets can not only differ in their content, origin (mobile phone, web), private nature (public vs. private) but also in their addressability: Twitter users can either compose public direct messages or indirect messages. Direct public messages contain an "@" sign (Honey and Herring, 2009) and the name of the recipient. They are used when the user wishes to indicate that a message is addressed to a specific person. Indirect messages, on the other hand, are used when the message is addressed to a potentially unlimited Twitter audience. Users can also forward messages written by someone else to their followers by retweeting somebody's tweet. The forwarded message is thus called a retweet (or RT, for short). In a retweet the characters 'RT' and the name of the author are added to the original message in order to credit the source. Retweets play the major role in Twitter since they are massively employed to pass along information. Both types of messages are an important means of communication and

[2]See `http://en.wikipedia.org/wiki/List_of_Twitter_services_and_applications`

interaction and will be reviewed in detail below.

Hashtags and trending topics

Twitter not only allows communication within a user's own network, it also allows users to automatically create thematic areas where a reference to a given subject is made possible by using the hashtag ("#" sign). Hashtags are successful at creating micro-networks that are focused on a specific theme (Zhao and Rosson, 2009), thus creating a shared resource based on spontaneous reflections. Nevertheless, Boyd et al. (2010) found that only 5% of tweets contain a hashtag, although 22% of tweets contain a URL. Huang et al. (2010) note that hashtags on Twitter are about "filtering and directing content, where emergent topics for which a tag is created, are used widely for a few days and then disappear"[p.8]. This supports the studies of Java et al. (2009) and Zhao et al. (2009), who note that people also use Twitter to break news. Twitter automatically displays the most popular current topics in a sidebar on the website. These "trending topics" lists are individually linked to the current set of tweets related to that topic. While tweets without hashtags are also displayed in trending topic lists, the act of tagging a tweet increases the likelihood of its being displayed in such a list.

4.2 A social capital framework for Twitter

In order to provide an operationalization of social capital in Twitter, the goal of this study is to offer a network based perspective on social capital. In order to provide such a perspective the different tie concepts on Twitter have to be defined, since social capital results from social ties alone. There is fundamental research in the field of computer mediated communication that suggest different tie concepts (see table 4.1) for Twitter.

Latent, weak and strong ties

Haythornthwaite (2002) is one of the first researchers to distinguish three types of ties in online social networks: latent ties, weak ties and strong ties. Latent ties correspond to communication channels that are technically open but not yet activated. For example, two individuals who belong to the same email network or the same online social network such as Twitter have latent ties. Weak ties exist once individuals begin to communicate through any of their shared channels of communication (e.g., they follow each other on Twitter). Strong ties eventually arise as individuals expand their use of existing channels and create more media communication to maintain their interactions (e.g., Twitter users retweet or mention each other).

Implicit and explicit ties

Matthew Smith (2010) translated the concepts of latent, weak and strong ties into only two types of ties, namely implicit and explicit ties. If a communication medium (in our case Twitter) is regarded as a type of affinity among individuals, then latent ties correspond to implicit ties. Indeed, Smith defines implicit ties as "an affinity that connects individuals together based on loosely defined affinities or inherent similarities such as similar hobbies or shared interests. [...] Implicit ties are weighted by the amount of similarity estimated between individuals"[p.71]. Implicit ties are not directed; hence when a person A has something in common with a person B, person B also has something in common with person A. Hence on Twitter implicit ties can express all the inherent similarities between Twitter users. In contrast to latent ties, the weak and strong ties defined by Haythornthwaite correspond to Smith's explicit ties, which are defined as ties that "link one individual to another based on some purposive action (e.g., sending an email, visiting) or a well -defined relationship (e.g., being a friend of, collaborating with). Individuals thus linked are aware of the explicit connections among them"[p.85]. In the case of Twitter, explicit ties thus include all public and visible interactions and connections that people maintain with each other (e.g., by following each other, at-replying, sending direct messages), whereas implicit ties represent only a measure of similarity between people. Compared with implicit ties, which are only partly visible to the user, explicit ties are clearly visible. Moreover, implicit ties are inherent to the users' attributes, behaviors and nature, whereas explicit ties are based on purposeful action between Twitter users.

4.2.1 Network flow model

The network flow model (Borgatti and Halgin, 2011) (see figure 4.2) offers a third operationalization of ties on Twitter. It generally relies on the notion that individuals in networks can form multiple ties with each other. This concept is commonly known as multiplexity (Scott, 2000) in the network science field. In Twitter, social relations between users can be defined as multiplex. Indeed, users can either (a) follow each other, (b) interact with each other, or (c) retweet each other. By looking at these three types of ties we find that they manifest themselves in three different networks. The network flow model explains how and why these three types of ties are related to each other. Borgatti describes the model as two kinds of phenomena called "the backcloth phenomena" and "the traffic phenomena" in the original work of Atkin (1974; 1977). The backcloth phenomena describe the infrastructure needed to let the traffic flow. The traffic phenomena describe the actual flow of information through the network.

Borgatti's work shows that the data traces that we find in Twitter can directly be translated to the phenomena that he describes in four categories. The first of these categories is "the similarities category [which] refers to physical proximity, co-membership in social categories and sharing of behaviors, attitudes, interests and beliefs. Generally, we do not see these items as social ties, but we do often see them as increasing the probabilities of certain relations and

dyadic events."[p.8]. This definition corresponds to the notion of similarities between Twitter users. Without any similarity between users there is a very low tendency that social ties will be formed between them. Thus the similarity backcloth captures the intuition that homophily in terms of user attributes leads to all subsequent explicit social ties between them. Borgatti defines the next category of phenomena of social relations as "the classic kinds of social ties that are ubiquitous [(which is the case of friend and follower ties in Twitter)] and serve either as role-based or cognitive/affective ties. Role-based includes kinships and role-relations such as boss of, teacher of and friend of [(In Twitter, 'follower of')]. They can easily be non-symmetric [(which is the case of friend and follower ties)]."[p.8] It thus appears that these types of ties correspond exactly to the friend and follower ties in Twitter. Borgatti describes the interactions category as "discrete and separate events that may occur frequently but then stop, such as talking with, fighting with, or having lunch with"[p.9] someone. This category corresponds to the interactional ties in Twitter. Indeed, they share the exact same behavioral traits: people do intentionally mention each other in Tweets, but also might stop doing so for certain reasons. When one looks at a relationship between two users in Twitter, their interactional connection might be active at that time or it might not. Finally, Borgatti describes the flows category as "things such as resources, information and diseases that move from node to node. They may transfer entities (being only at one place at a time) and duplicate them (as in information)."[p.9] In Twitter, this category directly translates as retweet ties between individuals. These ties are automatically created when information is transferred or, in this case, duplicated, from one actor to another.

4.2.2 Attention, interaction and diffusion ties

Following the distinction of ties in the network flow model it becomes apparent that users in Twitter maintain a number of multiplex ties with each other (see table 4.1). While similarity between users is a perquisite for social ties, the friend and follower ties and interactional ties are the ties that can be used to operationalize the harvested social capital on Twitter. Diffusion ties are able to measure the outcome of the social capital, which is the information diffusion in form of retweets.

Similarities: The considerations on different tie concepts suggest that the perquisite for the existence of social ties is a notion of similarity. Similarity has also been described by other researchers as latent ties or implicit ties. As there are a number of different data traces that can express similarity between Twitter users - and in this work we are explicitly interested in the similarity of interest - the potential data traces will be discussed in the next section.

Friend and follower (FF) ties: The resulting network of friend and follower ties represents the attention that users pay towards each other or in other words the possibility of information consumption, where people only obtain information through the people that they follow. Such friend and follower networks have also been described as "attention-information" networks (Golder and Yardi, 2010) in which attention is exchanged for information. Following the

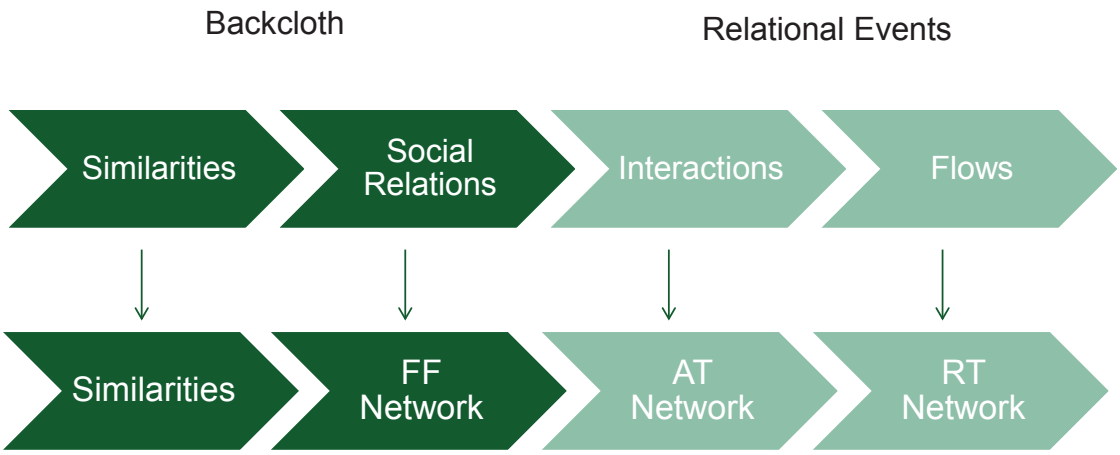


Figure 4.2: Network flow model with four types of dyadic phenomena according to Borgatti (2011). The top of the figure shows the original network flow model, the bottom the author's instantiation for the Twitter network.

distinction made for ties in the works above, FF ties can be considered as weak (due to their low maintenance), explicit (since they are publicly visible), social ties that people maintain in Twitter. These ties can be used to operationalize the structural social capital in Twitter.

Interaction (AT) ties: The exchange of direct tweets between people is considered a public conversation, a dialogue or, more generally, an interaction. People on Twitter usually interact with each other by placing the @ sign (AT) in front of the user they are publicly[3] addressing. The interactions between users, according to the definitions above are considered as strong (due to their interaction character), explicit (publicly visible), interaction ties that people maintain in Twitter. These ties serve as a source of structural and relational social capital in Twitter.

Diffusion (RT) ties: The last set of ties are retweet or information diffusion ties. Information diffusion or retweet ties are considered to be explicit, strong and to represent the actual flow of information in Twitter as a result of retweets (RT). While in a real world setting social capital explains why people receive certain goods or outcomes - for example a higher salary because of their social network - here we seek to explain why people receive more retweets from other users. This means that these ties can be used as the outcome that the social capital framework is seeking to explain.

The considerations of Haythornthwaite, Smith and Borgatti's network flow model show that a cornerstone of the network-based social capital concept of Twitter depends on the different definitions of ties. This means that we can think of implicit, latent ties as the similarity

[3] e.g., "@plotti Do you know of any social network analysis books?"

Ties capture	Social Capital			Outcome
Type of Ties in Twitter	Similarities between Twitter users	Friend & follower ties	Interaction ties	Diffusion ties
Types acc. to Borgatti	Similarity	Social relation	Interaction	Flow
Types acc. to Haythornth.	Latent	Weak	Strong	
Types acc. to Smith	Implicit	Explicit		
Valued	Yes	No	Yes	Yes
Directed	No	Yes	Yes	Yes
Visible	No	Yes	Yes	Yes

Table 4.1: The differentiation of different tie concepts for Twitter, adapted by author for the social capital framework.

perquisite for social ties. The subsequent ties (FF, AT and RT ties) can be classified according to their strength (weak, strong), visibility (implicit, explicit) or role (social ties, interactions, flows) in the information diffusion process. Especially the network flow model suggests that ties are not only an integral part of the Twitter entities but they do correspond to the theoretical reasoning about information diffusion in multiplex networks and allow us to capture the three dimensions of social capital. Following the considerations of the network flow model this suggests a very important implication for the network-based operationalization of social capital: The cognitive, relational and structural social capital dimensions are also able to explain information diffusion from a network perspective. The next sections will now follow the four stages of the network flow model and introduce data traces (Rausch and Stegbauer, 2006) that will describe the cognitive (similarities between users), structural (social ties), relational (interaction ties) social capital dimensions for Twitter and finally describe the outcome variable (diffusion ties) in the form of received retweets.

4.2.3 The cognitive dimension of social capital

This section will show which similarities between individuals have been analyzed for Twitter and so show which dimensions of "a shared language or vision" (Nahapiet and Ghoshal, 1998) in form of the cognitive dimension of social capital have been investigated. The investigation similarity between Twitter users has often been described as finding tendencies for homophily (McPherson et al., 2001). Homophily is the tendency to become friends with people who are similar to ourselves.

Numerous studies in Twitter have analyzed implicit ties or homophily: Weng et al. (2010) analyzed the thematic interests of Twitter users, which were defined by the content of the users' tweets. They found that Twitter users interested in the same topics followed each other more frequently than those who had different interests. They also found that users with shared interests had more reciprocal relationships (strong explicit ties). Kwak et al. (2010) analyzed geographic homophily and found that users like to connect with users in their proximity: indeed, for users with 50 or fewer friends the mean time difference between their locations was only 1 hour. This observation was confirmed by Takhteyev et al. (2010), who found that the distance between users, be it linguistic or geographical, has an effect on Twitter ties, despite "the substantial ease with which long-range-ties"[p.79] can be formed. They found that 39% of users connect within the same metropolitan area, and that ties between users more than 5000 km away from each other are underrepresented. In addition, Yardi et al. (2010) found that geographically local topics result in denser user networks in Twitter and that people who are central in Twitter networks tend to be geographically more central in the physical world. Finally, Bakshy et al. (2011) found that users listed in the same keyword categories such as "bloggers", "government", "celebrities" and "media people" tend to pay greater attention to members of their own interest groups. Kwak et al. (2010) found structural equivalence homophily based on the degree correlation of a node's degree against those of its neighbors.

This is equivalent to saying that high-degree users are more likely to connect with high-degree users than low-degree users. Golder et al. (2010) found that beyond structural effects the main reasons why Twitter users form ties are shared interests or organizations and activities, where sharing many interests with another person is viewed as a kind of similarity. The above studies show that the notion of similarity is well-observed in Twitter and that a number of different attributes can be used to describe it. These attributes are:

- Similarity of tweet content (Weng et al., 2010): measured by word usage, hashtag usage, sources cited, etc.
- Similarity of friends (Kwak et al., 2010; Golder and Yardi, 2010): individuals who follow similar people are perceived as similar.
- Similarity of followers (Kwak et al., 2010; Golder and Yardi, 2010): individuals with similar followers are perceived as similar.
- Similarity in structural equivalence (Kwak et al., 2010): individuals with a similar degree are perceived as similar.
- Similarity in geographical localization (Yardi and Boyd, 2010; Takhteyev, Y, Gruzd, A, Wellman, 2010): individuals living in the same region are perceived as similar.
- Similarity of hobbies or interests (Bakshy et al., 2011): individuals who are tagged or listed in the same category in Twitter lists are perceived as similar.

This paragraph showed, that depending on the definition of similarity, a different *aspect* of social capital's cognitive dimension can be observed. The next paragraph will show which Twitter data traces can be used to capture the similarity of the aspect of *interests* between users.

Using lists to determine cognitive social capital

Given the variety of data traces in Twitter, there are a number of options to determine a user's interests: (a) by studying his/her tweets or messages, (b) by studying his/her hashtags, or (c) by studying which lists the user was listed in. This paragraph will show why the data traces of Twitter lists are best suited to determine the user's cognitive social capital or his/her interests.

In order to determine the interests of users, the most obvious idea would be to look at the tweets these users have created. Assuming that each user is part of an interest group, the tweets of each member might be searched for keywords that represent this interest group, in order to reveal a shared language and vision. So, for example, if a certain interest group revolves around the topic of "wedding(s)", one could count how often each individual has mentioned the word "wedding". One might expect that the more often a user mentions this keyword, the more he shares a language or vision with this interest group. A more sophisticated approach to determine user interests uses the variants of probabilistic topical models (Blei, 2012) to analyzes the content of the user's tweets. A number of authors (Counts, Scott, Pal, 2011; Hong

and Davison, 2010; Ramage et al., 2010; Tunkelang, 2009; Weng et al., 2010; Zhao et al., 2011) have focused on classifying[4] users according to a set of given topics in their tweets. The simplest and most common topic modeling method is called the latent dirichlet allocation (LDA). It is based on the idea that since documents exhibit multiple topics, then users can also exhibit multiple interests. Researchers have applied this exact method to the analysis of tweets. Unfortunately, they have reported mixed results (Counts, Scott, Pal, 2011; Ramage et al., 2010), mainly because of the short length of tweets and the inherent use of acronyms, slang (Weng et al., 2010) and spam[5]. There are additional disadvantages to using the LDA approach: a majority of cognitive language details (such as style, humor, irony) cannot be captured by this nevertheless sophisticated method. In order to conceptualize an interest-based cognitive social capital, the LDA method could be used but it would suffer from the problems detailed above, which is why this study has decided to use a different method.

As shown in the literature review in section 3.4, the second way to determine the users' interests is by studying their hashtags. Hashtags in Twitter are a valid representation of often short-lived user interests that revolve around current trends and recent events (e.g., "#election2012", "#iranrevolution", or "#haiti_quake"). The corpus of all the hashtag keywords used could therefore be gathered, clustered and analyzed in a similar way to users tweets. Researchers (Charlie Nesson, 2009; Romero et al., 2011) have used hashtags to capture user interests and the quickly emerging and disappearing interest networks revolving around these hashtags, yet reported a number of operational and theoretical drawbacks. First, the researcher would need full access to the Twitter stream in order to capture all quickly emerging and disappearing hashtags. This is an impossible task since Twitter's publicly accessible timeline, called the "gardenhose", only shows a snippet of the entire Twitter stream: it holds[6] only approximately 5% of all the tweets produced by Twitter users. Second, if a user mentions certain hashtags, this represents only the very current interests of this user and does not allow researchers to capture the other or past interests of a user (Huang et al., 2010). Huang et al. (2010) note that hashtags are indeed used for "emergent topics for which a tag is created, are used widely for a few days and then disappear"[p.8]. Indeed, the interests found using this method only represent a small snapshot in time and refer to current topics that are being highly discussed while neglecting the other interests of the user.

A third way to infer the interests of a user is to look at Twitter lists. Indeed, a user can be interested in a multitude of topics, which can be revealed by which lists the user is listed. This concept has already been used by a number of researchers (Wu et al., 2011; Kim et al., 2010) to determine the interests of users: Wu et al. (2011) have determined the interests of users in four

[4] Blei (2012) explains such topical models as "topic modeling algorithms [which are] based on statistical methods that analyze the words of the original texts to discover the themes that run through them [...]"[p.2].

[5] It turns out that bots and spammers are actually using the "keyword-repetition" technique to score better on search results. Indeed, they constantly produce tweets that contain keywords representative of a certain interest group in order to be quickly found by users. The links in such Tweets then lead to dubious websites. It is obvious that here the users do not share a cognitive dimension of social capital with the group, but rather that they are trying to scam users.

[6] `https://dev.twitter.com/docs/streaming-apis/streams/public`

different categories, while Kim et al. (2010) used lists to determine a user's interest in multiple categories. Kim et al. report that lists can reveal the "characteristics and interests of the users [...] *even* if the users do not tweet about those interests"[p.1]. In a survey, where they compared the interests that users have with the one's that they were able to compile from Twitter lists they confirmed that "Twitter lists reflect well the perceived characteristics and interests of the users' in those lists"[p.7]. They also note that "Twitter lists are unique in that when we look at a user and the names of the lists that he is in, those list names represent what other Twitter users think of that user"[p.2]. So, in general, the list-based method additionally corresponds to asking *other* people on Twitter who they think is relevant to a given interest. Finally, numerous third party providers (see overview in table 6.2) contain directories of such lists and allow Twitter users to find other users that exhibit a certain interest or include themselves in a certain category. Indeed, lists are not only created to categorize other users into lists: users often self-categorize themselves, using specific keywords, by adding themselves to a list with other users.

The approach to capture a user's interests using Twitter lists can be applied in straightforward manner. Given that a large number of topic lists to measure[7] a person's interest[8] is available: By going through each of the lists for a given topic and counting how often each person is listed for this topic, we can obtain a measure of how strongly a person exhibits an interest in this topic. Details of this approach will be provided in section 5.4.2 and 5.2.2 in chapter 5.

In conclusion, the concept of the list-based method to capture the cognitive dimension of social capital has a number of advantages: First, using lists to determine a user's interest is cheap and relatively easy to perform, since no interviews or polls with the individual Twitter users have to be conducted. Second, since a user's interest is determined by a high number of votes by other users it is less prone to errors (Kadushin, 2004; Kim et al., 2010): Indeed, we do not have to rely on self-designation (Rogers and Cartano, 1962), where respondents are asked to what extent they are interested in a certain topic. Instead, this democratic approach is based on the intuition of many Twitter users. In comparison to the so-called key informant technique (Iyengar et al., 2010), where only a selected number of people are asked to name the key members of a given interest group, this method harnesses the collective intelligence of Twitter users. Third this method allows researchers to determine the multiple interests of a given person, and so works in a similar way like the LDA technique which is based on tweets, but does not suffer from the underlying deficiencies of this method. In comparison to the usage of hashtags it captures also the long-lived nature of a user's interest. Finally the list-based method has been proven as reliable and reproducible in determining the user's interest in previous studies (Wu et al., 2011; Kim et al., 2010; Greene et al., 2012). This is why the data traces contained in Twitter lists were chosen to conceptualize the cognitive social capital dimension of a person. The next paragraph will show that people who share the same cognitive social capital dimension also form interest-based social networks.

[7] Section 6.4 will describe how these lists will be collected.

[8] Chapter 6 will explain how we will determine these interests.

Cognitive social capital as a basis for groups

When trying to define groups on Twitter one has to ask the question what constitutes a community or group on Twitter. Generally given the very open nature of the network there are currently no clearly defined group concepts. On Facebook for example one can become a member of a group and so signal his membership, on Twitter there is no such mechanism. Therefore one has to reuse existing more general community or group concepts in order to define groups on Twitter. Regarding such concepts it is important to note that there are generally two different ways to define a group (Wasserman and Faust, 1994): One type of definitions is rather based on explicit social ties, while the other is based on similarities between members. These dualist conceptional approaches of online communities have also been noted by Etzioni and Etzioni (1999) by describing communities as groups that share the underpinnings of a) social relations between members and b) a common shared cognitive element between members.

Communities based on explicit ties

Definitions that are based on explicit social ties define groups or clusters as a result of a computational division of the tie networks (e.g., friend and follower ties in Twitter). These attempts assume that the social network structure can be divided into communities in such a way that a node has to be member of at least one community. In the computer science community, this problem is also known as graph partitioning (Fiedler, 1973; Kernighan and Lin, 1970; Pothen et al., 1989). Sociologists have also developed methods that separate a network into cohesive communities: examples of this are the hierarchical clustering method (Scott, 2000), the Girvan and Newman (2002) method, or the modularity method (Flake et al., 2004; Newman et al., 2006). On Twitter there are also a number of studies (Gonçalves and Ballon, 2011; Grabowicz et al., 2012) that have applied such methods on large corpora and found that most groups on Twitter have a size of 100-200 members.

However all of the approaches are based on the division of explicit ties and assume no prior knowledge about the similarity backcloth. This means that such approaches do not assume that groups are being formed because the group members have something in common. Instead such approaches infer possible similarities between members using post-hoc reasoning[9]. This means that for a community, that has been suggested by one of these algorithms, the researcher has to decide which attributes these persons have in common. Often such a task is rather unfeasible because the researcher does not possess enough information about the individuals to detect the attributes that the individuals have in common. A correct classification of the shared cognitive dimension of the group often involves a lot of domain expertise or prior knowledge about the group members in order to be meaningful. Moreover, the community detection methods are

[9]This type of reasoning is similar to that used in statistical clustering procedures where multi-dimensional scaling (MDS) is used and meaning is subsequently attached to the resulting dimensions.

unable to model overlapping social circles or group memberships, which are highly present in online social networks such as Twitter (McAuley and Leskovec, 2012).

Connecting this discussion on the existing research on communities we rather find evidence that this structuralist approach is in need for a cognitive shared dimension: Indeed, Etzioni and Etzioni (1999) defined online communities as being comprised of two attributes: "First it is a web of [...] relationships that encompasses a group of individuals [...]. Second a community requires a measure of commitment to a set of shared values, mores, meanings and a shared historical identity - in short a culture."[p.241]. Looking at communities of practice, Wasko and Faraj (2005) also find that "communities of practice [are] self-organizing, open activity systems focused on a shared practice"[p.37], suggesting that it is the shared cognitive dimension which then leads to the formation of communities. One of the main characteristics of these emerging online social communities is mentioned by Pipa Norris as being that "online communities indeed bring together like-minded souls who share particular beliefs hobbies or interests"[p.37] (2002).

Communities based on similarities

For these reasons, in this dissertation, interest-based groups will be defined not by explicit social ties, but by the similarities between members. This means that the cognitive social capital dimension is also the basis for the definition of groups in this work: Looking at the network flow model, we have seen that the network phenomena of social ties and interactions that lead to information diffusion start with certain similarities between persons. As we have seen the cognitive dimension of social capital captures exactly this similarity. Indeed, homophily is a major driver of social network group-forming processes (McPherson et al., 2001), which means that groups start to form around people sharing the same interests in Twitter (Kwak et al., 2010; Bakshy et al., 2011; Golder and Yardi, 2010; Weng et al., 2010; Wu et al., 2011; Kim et al., 2010). All subsequent phenomena, or social capital dimensions are based on the premise[10] of the existence of this shared cognitive dimension: Individuals with similar interests start to create network ties, which then result in interaction. Finally this leads to information diffusion among members of this group. Simply put, a group is a collection of individuals that all exhibit the same interest[11] and an individual is considered a member of a group if the

[10] It is important to distinguish between selection mechanisms (Kossinets and Watts, 2009) and influence mechanisms (Anagnostopoulos et al., 2008; Aral et al., 2009) when talking about group concepts. In this work the assumption is made that only the selection mechanism is existing in the Twitter network. This means that because of the users' homophily persons with similar interests create network ties. This work neglects the influence mechanism, which describes how network ties lead to more similarity between individuals because of two reasons: First because this mechanism is reported to be significantly less observed in online social networks than homophile selection (Aral et al., 2009; Aral, 2010; Lewis et al., 2010) and secondly because the outcome of the influence process is a change in the cognitive dimension of social capital of a user which is not considered an outcome in the social capital view.

[11] Since we analyze the selection process using the network flow model, we also assume that once an individual has acquired group membership, then it remains stable throughout the analysis.

individual shows enough similarity with the group members.

Interest-based communities defined by Twitter lists

This group definition has also been used by a number of researchers who studied interest-based groups on Twitter. Due to the lack of an existing group signaling mechanism on Twitter researchers have suggested to make use of Twitter lists which serve a similar purpose. They found that Twitter lists were a very useful data trace to determine people that belong to the same interest community: Wu et al. (2011) suggested that user list memberships can be used to organize users into a predefined set of interest groups where a user's group can be identified based on the number of lists to which they are assigned. Kim et al. (2010) suggested that latent interest-based groups in Twitter can be extracted by examining list data. Greene et al. (2011) showed that user list aggregation can be used to define topical Twitter communities. Zhao et al. (2011) confirmed that "users [that] are connected implicitly by being added to the same list"[p.1] form topic-based communities on Twitter and Yamaguchi et al. (2011) found in these interest-based groups "users included in the same list were likely to have posted on the same topic."[p.2]. Garica-Silva et al. (2012) confirmed that persons on the same lists also used similar terms and so have a shared language. Finally, Greene et al. (2012) showed that "identifying topical communities on Twitter by aggregating the "wisdom of the crowds", as encoded in the form of user lists, can detect and label coherent overlapping clusters of users"[p.8], and performs well in highly overlapping weighted networks (such as Twitter).

Generally there are two ways of using Twitter lists in order to classify members into interest communities. Twitter users can either a) follow one or many certain list that already has members that belong to a certain interest or b) they can be listed on such lists. In the first case the member actively signals his membership to the community, similar to confirming his interest in a certain topic. The advantage of this community definition is that the member is aware of being a member of this community, since he actively has listed himself on such a list. The drawback of this community definition is that the community might not be aware of the existence of this member in this community. Creating a similar situation in real life we can think of a community of experts on a specific topic. If we considered everybody that has enlisted his interest in this topic on his resume we would not be able to find the core members of this community that are aware of each other.

On the other hand by using the second approach and considering members of a community that have been suggested multiple times to belong to this community we can overcome this problem. By asking everybody to list members that belong to certain expert community we can be sure that the persons that have been listed most often are also forming the core community around this topic. Transferring this approach to the existing literature on social capital in chapter 2 we find a parallel to the different conventional collection methods of network data in real life. The questionnaire-based methods are based on the aforementioned name or position generators (Campbell, 1991), which were used to construct the group boundaries. Thus transferring this

concept to Twitter, this means that people that were listed multiple times on lists for a specific topic must constitute the core of this interest group or community (see figure 4.3). Regarding the problem that members have to be aware that they are part of a certain community, we have situation that Twitter notices each member if he has been added to a certain list, which means each Twitter user is well aware of the communities his is a member of. Of course the situation can occur that a person is listed in a community that he does not believe to be a member of, but by considering not only one list but a multitude of lists, the majority rule results in a community classification that the users very much agree with (Kim et al., 2010). In conclusion in this framework the later definition was used to construct the communities or group boundaries of the analyzed social systems.

Bonding and bridging cognitive social capital

After using the cognitive dimension of individuals to define the group concept, and defining Twitter lists as the main data trace to capture individual cognitive social capital, it becomes possible to define the cognitive dimension of bonding and bridging social capital. As described in chapter 2, bridging and bonding social capital describe the two different strategies individuals use to acquire social capital: when the cognitive bonding strategy is applied, the user exhibits a strong topical interest that matches the group interest, and when the cognitive bridging strategy is applied, the user exhibits interests in a number of different topics. According to this, bonding and bridging cognitive social capital can be defined in the following manner:

- The more often a person is listed for a certain topic, the more this person can be considered a core community member, and thus more creates cognitive bonding social capital.
- The more a person is listed on different topic lists, the more this person can be considered as a member of different communities, and thus creates more cognitive bridging social capital.

The two types of cognitive social capital are not exclusive: it is thus possible for a person to be highly listed for a certain topic and simultaneously be interested in a multitude of different topics.

In order to demonstrate the two different conceptions of cognitive social capital, an example is depicted in figure 4.3. Here person one has received three nominations (i.e., has been listed on three lists) for swimming, and person two has received two nominations for swimming. This means that both person one and two exhibit an interest in swimming and are group members of the swimming community. While both have a high cognitive bonding social capital for the swimming interest group, person one exhibits a stronger bonding cognitive social capital than person two. Generally, the more votes (or listings) a person receives for a certain interest group, the more cognitive overlap this person will have with other members of this interest group. In contrast to person one and two, person three exhibits a more diverse interest: this person has received nominations, or listings, for two different topics. Indeed, person three

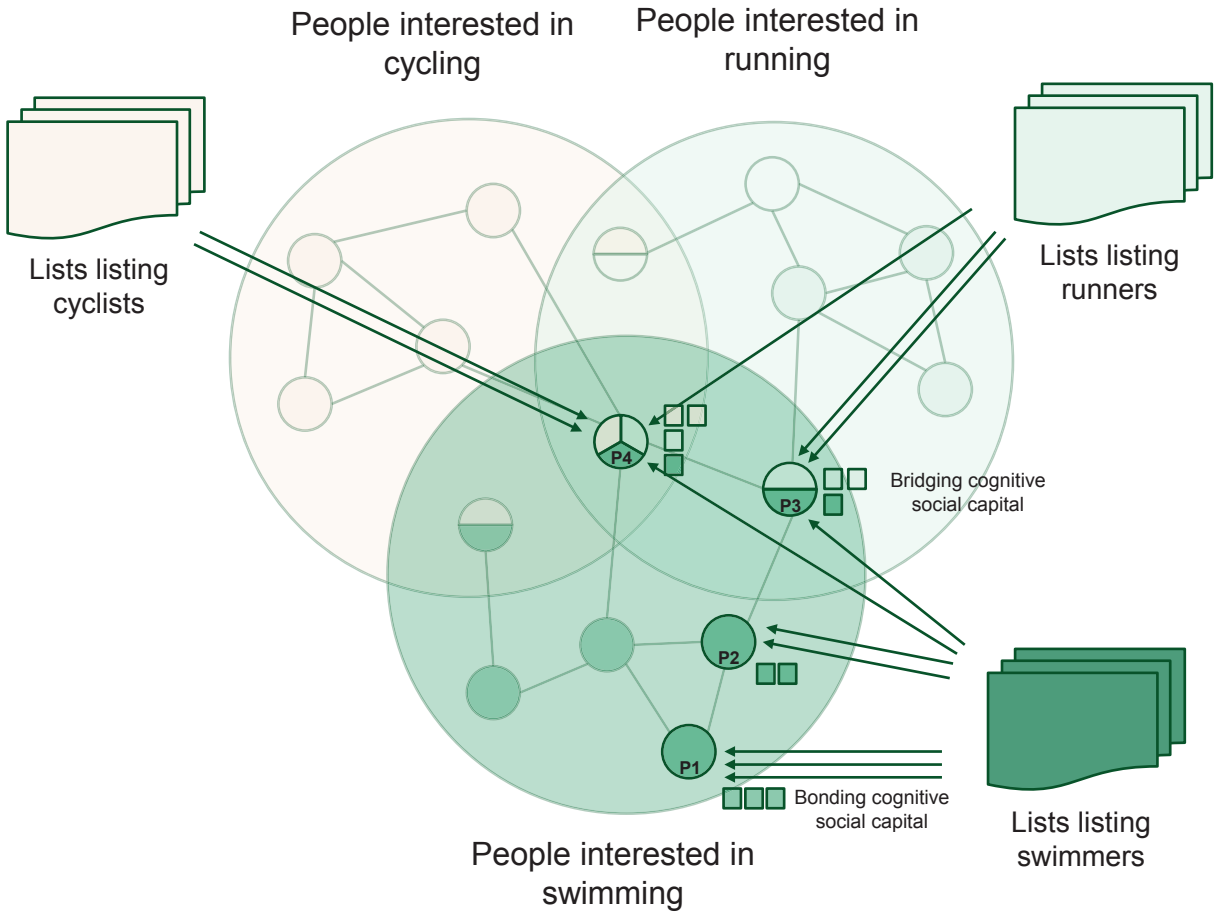


Figure 4.3: Overview of how Twitter lists can be used to determine cognitive social capital for a certain topic. The listings of persons 1-4 have been depicted with arrows. Non-directed ties represent social ties. Directed ties represent listings.

exhibits an interest in running and swimming. Person three is both a member of the running and swimming community. This means this person is building cognitive bridging social capital, which stems from receiving *multiple nominations for different lists*. The more a person receives nominations for different lists, the more this person is creating cognitive bridging social capital. For example, person four has received three nominations from three different lists, and thus has a higher bridging cognitive social capital than person three. The details of the resulting metrics will be provided in section 5.4.2 and 5.2.2 in chapter 5.

4.2.4 The structural dimension of social capital

This paragraph will show how the most obvious explicit friend and follower ties on Twitter (FF ties) define the structure of the social network and thus provide the infrastructure through which information flows. This paragraph will highlight why friend and follower ties should be used to express the structural dimension of social capital for Twitter.

Using friend and follower ties to determine the structural dimension of social capital

A variety of studies have analyzed the network structure of the friend and follower ties in Twitter. Observing existing network structures they have provided valuable information on how this type of ties are used by Twitter users to create social capital. Indeed, researchers have shown that this network structure influences the way individuals share information with each other in Twitter and that users have a tendency to create small dense networks of individuals, where they can shortcut connections in order to follow their interests. Researchers agree that the characteristics of small-world networks (see section 5.1.2) in Twitter facilitate information diffusion: Java et al. (2009) found that the Twitter network shows properties of a small-world network, with a power law exponent of -2.4, which is similar to the power law exponent found in the blogosphere. Galuba et al. (2010) found that the follower graph has properties of a small world with a giant connected component encompassing the majority of the users, which is characteristic of many other social networks both online and in the real world. They also found that the degree distribution of the indegree (followers) and outdegree (friends) follows a log-normal fit, which is supported by other analyses (Gayo-Avello, 2010; Weng et al., 2010; Bakshy et al., 2011). Kwak et al. (2010) found that very few people follow more than 10,000 users. They also report that the distribution of followers fits a power law for up to 105 users. Studies have also shown that the average path length between members (see section 5.1.2) is extremely small: Kwak et al. found that it is 4.12 hops. Galuba et al., who report a mean shortest path of 3.61, have also confirmed this finding. Galuba et al. interpret it as an indication that people follow others not only for social networking purposes, but also to acquire information, since the act of following someone represents the desire to receive all of their tweets. They note that indeed, on average, information flows over less than 5 hops between 94% of user pairs.

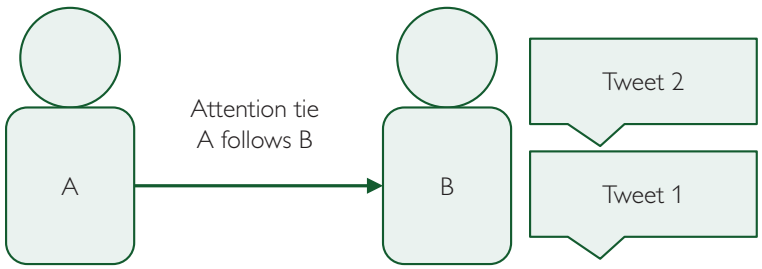


Figure 4.4: A friend-follower or attention tie. A high indegree means that a lot of attention is given to a specific person.

Each person's friend and follower ego-network shows a unique structure. Twitter researchers have tried to use this structure in order to classify Twitter users. Indeed, Java et al. (2009) divided users into three types: (1) information-sharing users, (2) information-seeking users and (3) users with friendship-wise relationships. Krishnamurthy et al. (2008) distinguish between (1) broadcasters, which are users with a much larger number of followers than followees, (2) acquaintances, which are users that exhibit reciprocity in their relationships, and (3) evangelists, which are users that "contact everyone and hope that some will follow them"[p.2]. Finally, Choudhury (2010) defined the three basic information roles of users: "generators", "mediators" and "receptors". Generators are users who create several posts but few users respond to them. Receptors are users who create fewer posts but receive several posts as responses. And mediators are users who lie between these two categories.

Since friend and follower ties significantly represent the attention of Twitter users and shape the overall structure of the social network, they can be used as a core concept for the structural component of social capital. Additionally, friend and follower ties represent a big proportion of how people play certain structural roles in this environment by representing key mechanisms in the information diffusion process: while some people focus on consuming information, others focus on generating news, and others serve as mediators between these two modes of information processing.

Structural bridging and bonding social capital

The definition of cognitively similar groups detailed above will help distinguish between structural bonding and structural bridging social capital. Users who build ties to users that share the same cognitive dimension (i.e., the same interests) are bonding, or creating structural bonding social capital (e.g., see P1 and P2 in figure 4.3). Twitter users that are dissimilar (that is, they are users that do not share a similar interest) yet create friend and follower ties are

creating structural bridging social capital (e.g., see P2 and P3 in figure 4.3). Since there are multiple ways of determining the different tie configurations within a network, the sections 5.2.2 and 5.4.2 in chapter 5 will describe how structural bonding and bridging social capital can be operationalized for groups and individuals.

4.2.5 The relational dimension of social capital

In Twitter, the publicly expressed @-interactions can be used to determine this relational dimension of social capital. It is important to distinguish between the cognitive dimension of social capital and its relational component, which is the tie strength between people (Gatignon and Robertson, 1985; Rogers, 1995). Brown and Reingen (1987) have noted that the concepts of 'strength of tie' and 'homophily' are often erroneously treated as synonymous, yet they differ greatly: homophily refers to the attribute similarities (e.g., the interests, language, gender, etc.) of individuals who interact with each other, whereas tie strength is a relational property that manifests itself in different types of social relations (e.g., best friend, close friend, acquaintance, etc.).

Using interactional ties to determine the relational dimension of social capital

As depicted in figure 4.5 interactions on Twitter can also be expressed as interactional ties. Due to their institutional nature and their ability to be weighted, these ties are well-suited to express the relational component of social capital. Interactional ties are highly valued to determine the strength of the connection between individuals (Yang and Counts, 2009; Huberman et al., 2009). Romero et al. (2011) confirm the value of interactional ties in determining the strength of the connections between individuals, noting that "users often follow other users in huge numbers and hence potentially less discriminately, whereas interaction via @-messages indicates a kind of interaction that is allocated more parsimoniously and with a strength that can be measured by the number of repeated occurrences"[p.689]. Such actions in Twitter can be interpreted as valued ties that publicly indicate the strength of a connection (Huberman et al., 2009; Gonçalves and Ballon, 2011) between users and the content of their discussion.

The conversational aspects of interactional ties have also been the focus of several Twitter studies: Honeycutt et al. (2009) analyzed the function and use of the '@'-sign in Twitter[12] as well as the "coherence" of the tweet exchanges. Their findings show that Twitter contains a high degree (approximately 30% on average) of conversationality (defined as the use of the @ sign), which has also been confirmed by other researchers (Java et al., 2009; Boyd et al., 2010). Honeycutt et al. found that 91% of '@'-signs were used to direct a tweet to a specific person. They also found that 30% of these direct messages received a public response — a phenomenon which is also reported by Huberman et al. in their study (2009).

[12] It is interesting to note that the usage of the @ sign to direct messages to other Twitter users has long been used in Internet Relay Chat (IRC) in order to address particular members of the chat room.

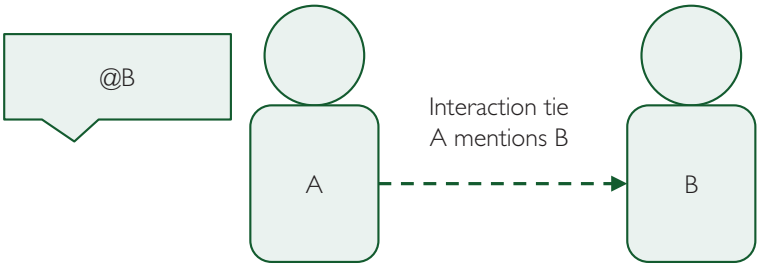


Figure 4.5: An interaction tie. A high indegree means a person gets mentioned often or is at the center of the discussion.

In conclusion, interactional ties are well-suited to capturing the relational interactions between Twitter users, i.e., the relational component of social capital. Indeed, this is because interactional ties lead to high levels of conversationality and because they possess a highly institutional character to address certain people. Additionally, due to the weighted nature of interactional ties in comparison to friend and follower ties, they are able to specify the various levels of tie strength between individuals. Moreover, the network model has shown that interactional ties represent a non-permanent structural network between individuals and can therefore be used to capture the less stable non-permanent structural component of social capital. This means that the interactions that people create also represent a non-permanent discussion network where certain individuals can emerge as highly central, because they are addressed by a high number of people.

Bridging and bonding relational social capital

In a similar way to friend and follower ties, interactional ties created between members of the same group of people that share the same interest contribute to the creation of bonding relational social capital. In comparison, interactional ties created between members of different interest groups contribute to the creation bridging relational social capital. The sections 5.2.2 and 5.4.2 in chapter 5 will focus on the relational social capital for groups and individuals and the operationalization of network-specific relational ties.

4.2.6 Outcome of social capital or information diffusion

Using diffusion ties to determine information diffusion

The literature review on information diffusion in section 3.4 highlighted that retweets have been used as a direct representation of information diffusion or knowledge sharing in a majority of studies. Indeed, the main purpose of retweets in Twitter is to forward interesting content from one's followees to one's followers. Retweeting, which is a core mechanism of information sharing, has become so widespread that Twitter added a feature in 2009 to allow users to easily retweet using only one click. Moreover, the word "retweeting" is now synonymous to 'sharing information' in the English language. The review on the factors affecting retweets (see chapter 3.4) has also shown these can largely be subsumed in the three different social capital dimensions, may it be the users structural position in the network (Gayo-Avello, 2010; Kwak et al., 2010; Ediger et al., 2010), his relational efforts to maintain reciprocal conversations with other users (Bakshy et al., 2011; Yang and Counts, 2009; Suh et al., 2010) or his cognitive dimension (Cha et al., 2010; Wu et al., 2011; Choudhury et al., 2010). Boyd et al. (2010) also note that "people often retweet as an act of friendship, loyalty, or homage by drawing attention"[p.6] and that people are well aware of the different social capital strategies that they occupy noting that "retweeting [...] is most successful when the retweeter has a large network and occupies structural holes, or gaps in network connectivity"[p.7].

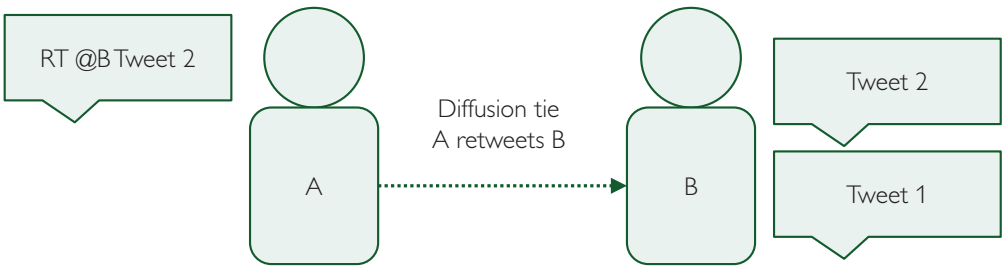


Figure 4.6: A diffusion or retweet tie. A high indegree means a person gets retweeted often.

As shown in figure 4.6, retweets can be expressed as explicit ties that are created between two users whenever a retweet takes place. These ties will be used as an adequate measure of information diffusion in Twitter. Therefore, the diffusion ties will be used to measure the *outcome* or return that people harness, based on the social capital they have built up. This

return or outcome can also be expressed in network terms: the more retweets individuals obtain, the more incoming retweet ties they possess. At the group level, it is important to express information diffusion as a property of the group, rather than as an aggregate of the ties created between individuals. Therefore, network measures describing all of the network ties between individuals of a group will be presented in the next chapter. Finally, the outcome has to be defined differently if we want to measure bonding social capital or bridging social capital: Studying the influence of bonding social capital, we are interested in knowing how many retweets an individual has received from group members, because of his social ties to these members. Studying the influence of bridging social capital, we are interested in knowing how many retweets the individual has obtained from members outside the group. These operationalizations will be described in detail in sections 5.1.2 and 5.3.2 in chapter 5.

4.3 Conclusion and goals

Reviewing the literature on social capital revealed the absence of a consensus on how social capital should be measured or which data traces in Twitter should be used to measure its dimensions. This stood in stark contrast to the widespread consensus of the research community on the perspectives of social capital (bonding vs. bridging social capital, and group vs. individual social capital) and its dimensions (cognitive, structural, relational). In order to fill this gap, this chapter has provided a framework of social capital for Twitter. The framework expressed how the unit of analysis (the individual or the group) can be differentiated in Twitter: the group is defined as individuals who share the same interest, while the individual is embedded in a group. It also described the difference between bonding and bridging social capital: bonding social capital is created when individuals form ties with others that are similar to them, while bridging social capital is created when individuals form ties with non-similar members from other groups. The social capital framework has also shown how the available data traces in Twitter can be expressed through the notion of ties, which then in turn can be used to express the structural, relational and cognitive dimensions of social capital for Twitter. The cognitive, structural and relational dimensions of social capital for Twitter will be summarized in the following paragraphs.

The cognitive dimension of social capital in Twitter can be captured by the implicit ties that people posses: the more implicit ties individuals have with group members - the more they share the interest of their group - the higher their cognitive bonding social capital is. The more implicit ties individuals have with non-group members - the more diverse interests they have - the higher their cognitive bridging social capital is. In order to determine the users' interests the Twitter list data traces have been proposed. This chapter has shown that the Twitter list feature offers a very intuitive and academically agreed-upon way to conceptualize cognitive social capital and is able to better capture the users' interests than by looking at his tweets or his hashtags. Additionally it could be shown that lists create a notion of interest-communities where people are aware of the inherent similarities of users that surround them in

their interest-based networks.

The structural dimension of social capital in Twitter can be captured by users' explicit friend and follower ties as well as their interactional ties. These ties create network structures that can then be analyzed with the methods and metrics of social network analysis. While the networks created by users' friend and follower ties are directed and binary, the networks created by users' interactional ties are directed and valued. The structural network positions, which capture the structural component of social capital, can be evaluated using network metrics. In the following chapter 5, the ways in which network metrics can capture the structural dimension of social capital will be discussed. These metrics will allow us to quantify it on an individual and group level, and distinguish between structural bonding and bridging social capital.

The relational dimension of social capital in Twitter can be captured by users' explicit friend and follower ties and their interactional ties. While the friend and follower network can offer insights into the reciprocity of relations, the interactional network can offer insights into the tie strength and reciprocity of relations. The relational dimension of social capital will be operationalized in chapter 5, at an individual and group level, and for bridging and bonding social capital.

The outcome variable will be measured by the retweets that people receive as a return of their activities. Retweets have been proposed to be operationalized as explicit retweet ties that are created as a result of a retweet happening between two persons. This allows us to measure retweets in a network structure. The details on how this network structure can be evaluated for groups and individuals, and also serve as a basis to measure retweets from group members and from members outside the group will be provided in chapter 5.

The resulting final social capital framework for Twitter is shown in figure 4.7. The framework shows that the social capital concept for Twitter contains the same cognitive, structural and relational dimensions of social capital as in Nahapiet et al.'s (1998) model. The framework also shows that these dimensions are expressed through the implicit and explicit ties in Twitter, which were introduced in this chapter. The model maps these dimensions to attentional (FF ties) and interactional ties (AT ties), and expresses the outcome variable through diffusion ties (RT ties). In conclusion, the model shows that the theoretical building blocks of the network flow model (similarities, social ties, interactions and flow) map exactly to the three cognitive, relational and structural of social capital dimensions and its outcome.

Now that a social capital framework for Twitter has been built, this study will focus on formulating hypotheses about the influence of social capital on the information received. These hypotheses will be operationalized using the building blocks from the model. The next chapter will provide an operationalization of social capital for each of the four quadrants defined in chapter 2 (see figure 2.3). All quadrants will contain the same social capital dimensions shown in figure 4.7 and, for each of the four quadrants, quantifiable measures for the cognitive, structural and relational dimensions of social capital will be provided. The next chapter will thus provide an operationalization of (a) bonding social capital for individuals (section 5.2.1), (b) bonding social capital for groups (section 5.1), (c) bridging social capital for individuals

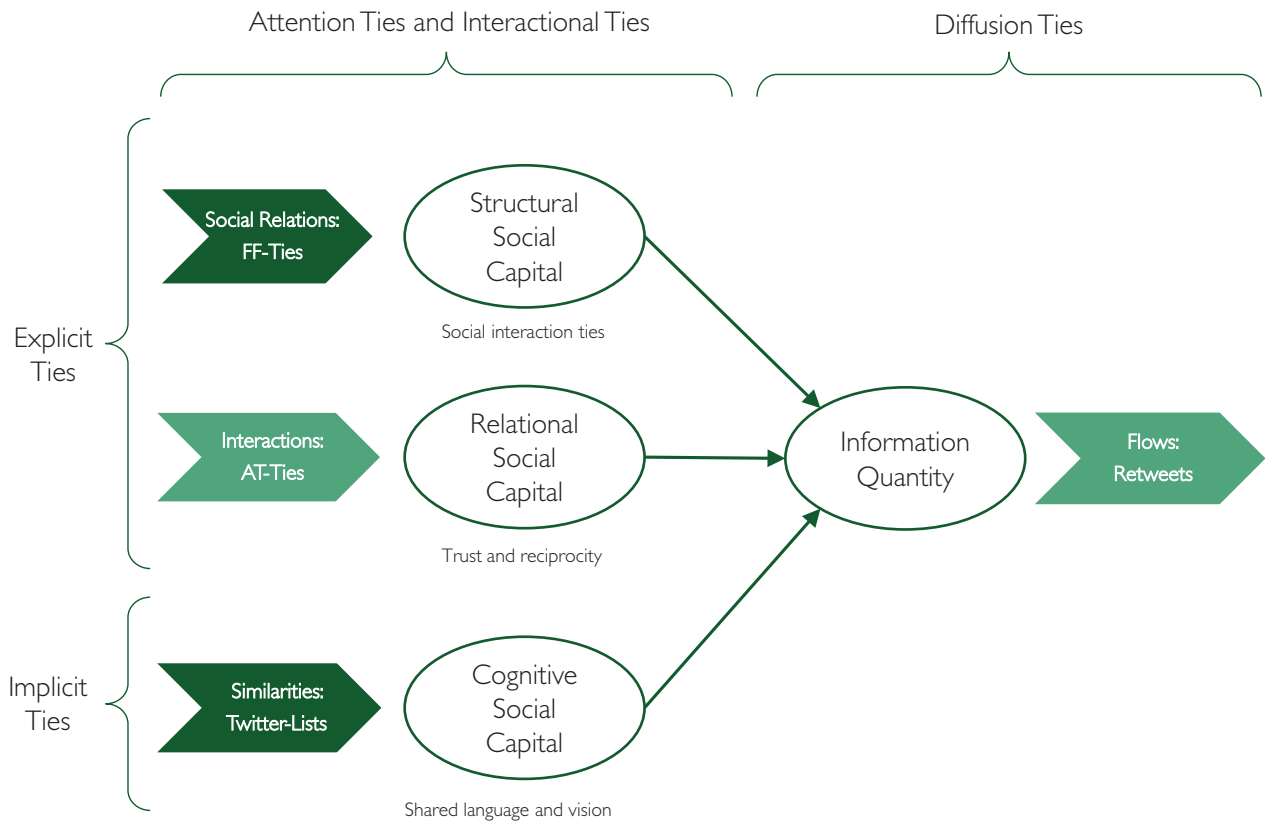


Figure 4.7: Adapted social capital model of Nahapiet. The framework integrates the three types of social capital, the implicit and explicit ties and the underlying network flow model.

(section 5.4), and (d) bridging social capital for groups (section 5.3).

5 Hypotheses and measurements of the influence of social capital on information diffusion in Twitter

After providing a social capital framework for Twitter, we return to the main question of this thesis:

> **Research Question:** How does online social capital influence information diffusion in Twitter's interest-based social networks?

Drawing on the notions of individual social capital, group social capital, bonding social capital, and bridging social capital, as well as on the social capital framework for Twitter presented earlier, in this chapter I will form hypotheses explaining how social capital influences information diffusion. In table 5.1 and figure 5.1, the four research questions were translated into sets of hypotheses. Hypotheses sets H1 and H2 predict that bonding social capital has positive effects on information diffusion within the group, while hypotheses sets H3 and H4 hypothesize that bridging social capital has positive effects on information diffusion outside of the group. Finally, hypotheses sets H5-H8 postulate that bridging and bonding social capital have negative interaction effects on information diffusion.

This chapter will first introduce the hypotheses related to groups and then introduce those related to individuals, since groups are the environment in which individuals are perceived. The analysis of a social network group structure, which is part of research question one, plays an important role in the plausibility assumptions of the empirical results in chapter 7. For this reason, I will start out by presenting the hypotheses and measures for RQ1. Then, the hypotheses and measures answering RQ2 will be laid out. Finally, I will describe the hypotheses and measures for RQ3 and RQ4.

Measurement indicators are needed in order to operationalize our hypotheses. Thus, the measurement indicators of the type of social capital presented will be introduced following each set of hypotheses. These indicators will make use of existing network measures that capture certain social capital dimensions. The main benefit of using network measures is that for each of the described quadrants, the social capital metrics will come right "off the shelf"[p.1.] (Everett and Borgatti, 1999). This means that we can rely on established metrics that have been used by network researchers for decades. In most of the reviewed literature on social capital, the term "network" has often been understood as an informal, face-to-face interaction or as a synonym of membership to civic associations or social clubs. In this study,

	Bonding	Bridging
Group	**RQ1:** How does group bonding social capital influence the overall diffusion of tweets inside the group? Hypotheses: **H1a, H1b**	**RQ3:** How does group bridging social capital influence the diffusion of the group's tweets to other groups? Hypotheses: **H3a, H3b, H3c**
Individual	**RQ2:** How does a person's individual bonding social capital influence the diffusion of their tweets to other members of the group? Hypotheses: **H2a, H2b, H2c**	**RQ4:** How does a person's individual bridging social capital influence the diffusion of their tweets to members of other groups? Hypotheses **H4a, H4b, H4c**

Table 5.1: Tabular overview of the hypotheses on the influence of social capital on information diffusion

the formal definition of "network", as a collection of nodes and edges, will be used. Indeed, network theorists argue that an understanding of social capital requires a fine-grained analysis of the specific quality and configuration of network ties (Everett and Borgatti, 1999) and the formal definition of "network" allows this.

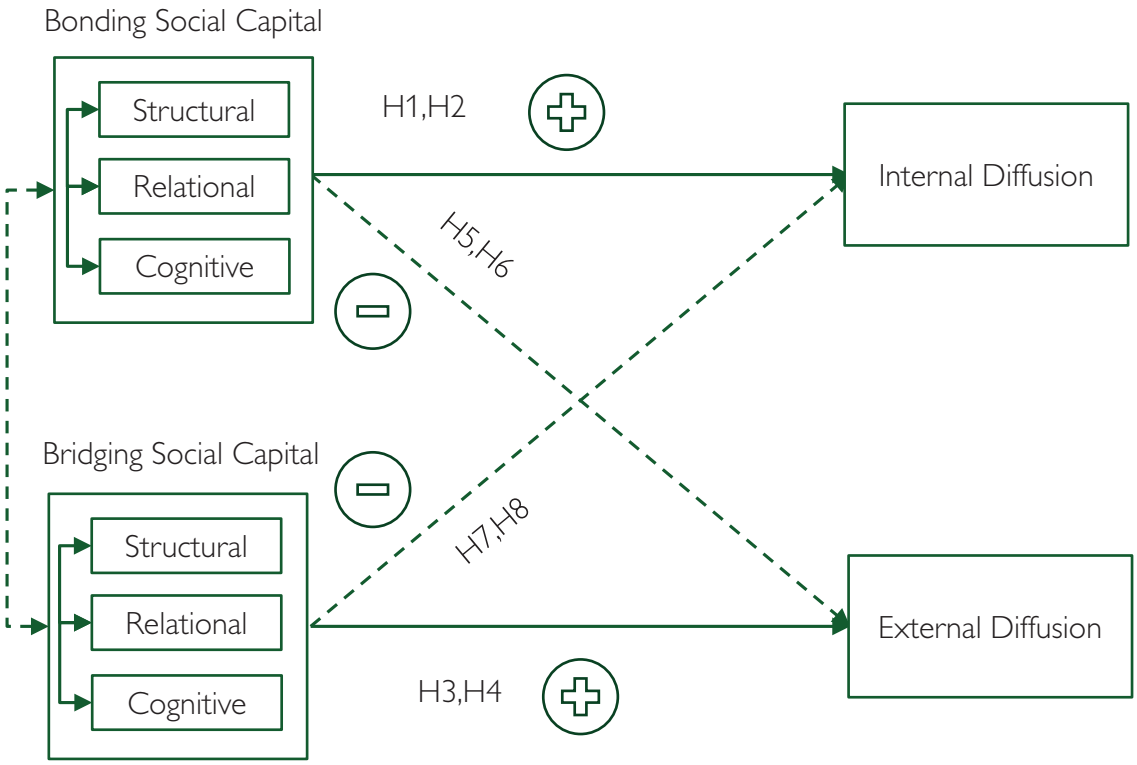


Figure 5.1: Overview of the hypotheses concerning the influence of social capital on information diffusion.

To get a better understanding of how network metrics describe social capital in each of those quadrants, the measurement of social capital will be demonstrated using the illustrative example network shown in figure 5.2. The network in figure 5.2 contains four groups A, B, C, D which are partly connected by strong ties (weighted with a symbolic value of 2 - dark tie color) and weak ties (weighted with a symbolic value of 1 - light tie color). Nodes with the same color also share the same cognitive dimension, and thus belong to the same group. This weighted network allows us to show the results of network metrics that will then be computed separately

for the AT and FF networks.

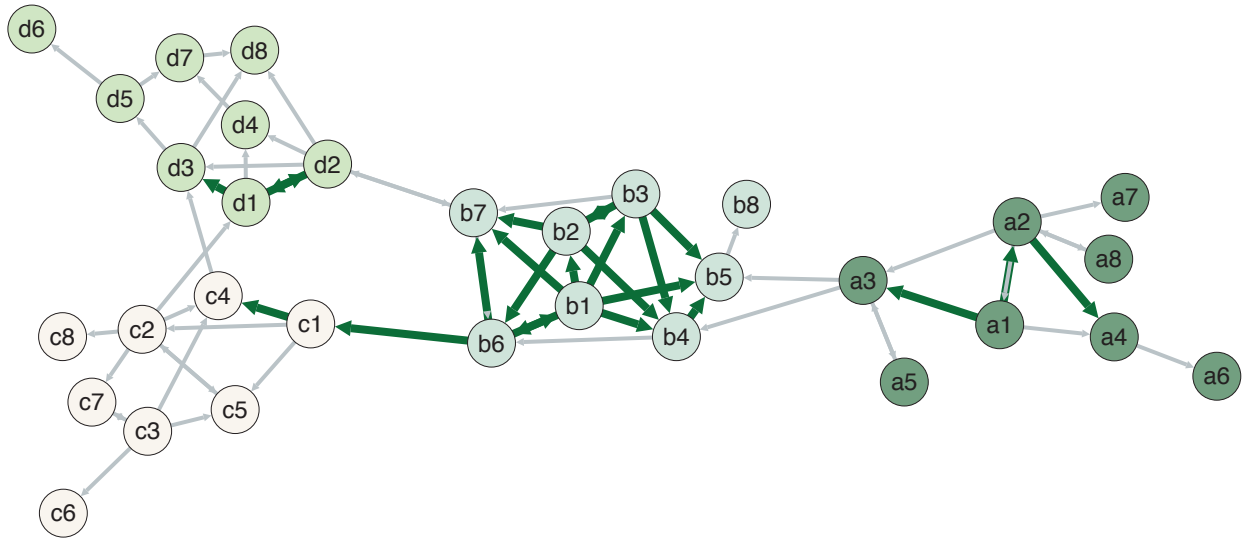


Figure 5.2: An example of a network of 4 groups with 8 actors each. Each color corresponds to one group. This network was created by the author for illustrative purposes.

5.1 Influence of group bonding social capital on information diffusion

5.1.1 Hypotheses

The basic assumption behind the first set of hypotheses is that creating bonding group social capital facilitates information exchange, where information becomes a resource that is exchanged between group members. At the internal group level, social capital is the type of group bonding social capital described by Bourdieu, Putnam and Coleman (Coleman, 1988; Putnam, 1995; Bourdieu, 1986). They differentiated between structural bonding group social capital and relational bonding social capital.

According to Lin (2002), structural bonding group social capital is seen as network closure, which "has a distinctive advantage of social capital because it is closure that maintains and enhances trust, norms, authority and sanctions. These solidifying forces may ensure that it is possible to mobilize network resources"[p.34], such as the basic resource of information. Lin further notes that "preserving or maintaining resources, denser networks may have a relative advantage"[p.34] for the group. It has also been suggested that the online social capital of groups (Ellison et al., 2006, 2007; Norris, 2002) fosters information diffusion between these

members. Moreover, the literature on information diffusion suggests that although information can be expected to spread from person to person, it will mostly circulate within groups before it circulates between groups (Burt, 1999; Granovetter, 1978). The research of the diffusion of innovation (Rogers, 1995) has also shown how cognitive group norms can lead to interpersonal networks that can enhance the diffusion process (Berth, 1993; Lapinski and Rimal, 2005) within the group. Finally, the literature on information diffusion in Twitter has shown that over time always a densification of user networks takes place (see chapter 3.4). At the group level, this means that more bonding social capital is created, in the form of denser attention (FF) networks. To explore this phenomena, hypothesis H1a will test if the structural group bonding social capital created in interest-oriented groups also leads to increased information diffusion between members of the group (H1a).

In the literature, relational bonding group social capital is also known to enhance information diffusion within groups. For example, some studies suggest that groups with higher levels of trust (Menzel and Katz, 1955; Rogers, 1995) also lead to a better diffusion of innovation. Other authors, such as Krackhardt (1992; 2002), have also emphasized the importance of strong ties in enabling group collaboration and knowledge exchange. Therefore, hypothesis H1b claims that the more relational social capital the group possesses, the more information will be diffused within the group. In summary, the set of hypotheses (H1a,H1b) will evaluate if group bonding social capital will lead to more information diffusion within the group.

Hypothesis 1a: The more structural bonding group social capital the group realizes, the more retweets are exchanged in the group network.
Hypothesis 1b: The more relational bonding group social capital the group realizes, the more retweets are exchanged in the group network.

At the group bonding level, the cognitive dimension[1] of a group should not be expressed as a hypothesis since it is the basis for the existence of interest-based groups that are collected (see the considerations network flow model in chapter 4). This means that the interest group collection process (see chapter 6.4) will optimize the memberships to groups in a way that minimizes the variance inside their cognitive dimension. In other words, we will obtain groups of people that share the same interest. In order to statistically prove the hypotheses presented earlier, multiple groups need to be collected and their social capital compared[2]. The collection of groups will be performed in section 6.4 of chapter 6, whereas the measurement of structural and relational group bonding social capital will be provided below.

[1] Instead to varying the cognitive dimension inside the group, the influence of the group's cognitive dimension on information diffusion will be expressed as part of the group bridging hypotheses (see section 5.3).

[2] This empirical observation of different groups is markedly different from the single case observation (Adar et al., 2004; Leskovec et al., 2007) of groups, where only one group is studied and the group always remains on the outer boundary of data collection, or perishes completely as a structural component.

5.1.2 Measurements

Structural dimension

From a network perspective, group bonding social capital corresponds to a number of simple network metrics such as network cohesion or density of the network. Indeed, Van der Gaag and Snijders (2005) argued that group social capital could easily be measured using the volume or density of the group's ties. More sophisticated measures such as the average degree of nodes in a group or network clustering allow group social capital to be measured at a finer level of detail. It is generally assumed by researchers that the more cohesive the network is, the more bonding capital that group possesses. The metrics evaluate the social capital of a group as a whole and indicate that it is independent of the direction of ties: person A following person B contributes to the group's social capital just as much as when person B is following person A. The next sections will explain what I used as indicators of structural group bonding social capital.

Measurement by network density

Network cohesion can be measured in a number of different ways. *Network density (D)* (Wasserman and Faust, 1994)[p.129/143] is seen as the most prominent network measure of social capital (Gaag, 2005). Friedkin (1982) showed that a higher number of ties is better for building social capital than a lower number of ties. Indeed, a high network density leads to cohesive embedded ties, which lead to greater information exchange between group members. According to Coleman (1988), this cohesion motivates individuals to devote time and effort to communicating with each other.

Network density is described as the ratio of the total number of existing ties to the total number of all possible ties. The maximum number of edges $|E|$ is $\frac{|V|}{|V|-1}$, so the maximal density is 1 and the minimal density is 0. The more ties there are between members of a group, the denser the network is, and the higher the network density is.

$$D = \frac{2\,|E|}{|V|\,(|V|-1)}$$

Measurement by average distance on a group level

Another metric of group social capital can be determined by computing the *average distance (a)* (Wasserman and Faust, 1994)[p.111] between all pairs of members. This metric is based on the idea that smaller distances between group members lead to faster communication between members, and an increase in group bonding social capital. The average distance between all pairs of members is the average of the geodesics between any pair of nodes[3]. Assuming that

[3]The geodesic (Wasserman and Faust, 1994)[p.110] distance or geodesic between two nodes is a shortest path between those nodes.

communication flows on the shortest paths between people, Baker (1992) argues that networks with more direct connections improve communication between actors. Baker suggests that information quality deteriorates as it moves from one person to the next. The average distance is defined by

$$a = \sum_{s,t \in V} \frac{d(s,t)}{n(n-1)}$$

where V is the set of nodes in the group, $d(s,t)$ is the shortest path from s to t, and n is the number of nodes in the group. For example table 5.2 reports that group D has the lowest average path length in the example network depicted in figure 5.2. This means that information in this group does not have to travel a long distance to reach all members of the group. This results in a higher group bonding social capital.

Measurement by clustering on a group level

A third way of measuring group bonding in networks is to use global *network clustering* (Watts, 1999) or *transitivity* (Wasserman and Faust, 1994)[p.150/165]. Granovetter (1973) intuitively observed the tendency for networks to cluster and reasoned that two people that are already friends both have a high chance of becoming friends with a third partially-known person. This tendency towards closing triads leads networks to create densely knit pockets of friends that were introduced to each other. A historical measure that describes transitivity is the number of transitive triples (Holland and Leinhardt, 1981), yet a more commonly used measure of group bonding through transitivity is network clustering (Adar et al., 2004; Gruhl et al., 2004; Leskovec et al., 2007). Researchers also postulate that group bonding has an influence on information diffusion. In one of the most prominent works in the literature, Duncan Watts (1999) analyzed “small worlds”: he observed that society is divided into groups with smaller denser connected subgroups. He introduced the measure of a “clustering coefficient” for networks, which serves as a good indicator of how fast information spreads in networks. Whereas a random network (Albert et al., 1999) shows the same structure throughout, a clustered network contains more and less strongly connected pockets of connected people. A high clustering coefficient indicates a high probability that a person’s friends know each other; whereas a low clustering coefficient means that there is a relatively low probability that the friends know each other. Thus, the clustering coefficient describes how clustered the network is. In highly clustered networks (which are a result of structural group bonding), information is, in theory, supposed to diffuse faster than in networks with lower clustering coefficients. This will be tested as part of hypothesis one.

The *local clustering coefficient* (C_l) (Watts, 1999) of a node in a network quantifies how close its neighbors are to being a clique (complete graph). In order to describe the coefficient for groups, I will make use of the average local clustering coefficients of all members of the group.

The local clustering coefficient C_i for a node n_i is given by the proportion of links between the node within its neighborhood, divided by the number of links that could possibly exist between them. For a directed graph, e_{ij} is distinct from e_{ji}, and therefore, for each neighborhood N_i, there are $k_i \cdot k_{i-1}$ links that could exist between the nodes in that neighborhood. The local clustering coefficient is defined as:

$$C_l(v) = \frac{|\{e_{jk}\}|}{k_i(k_i - 1)} \quad for \quad n_j, n_k \in N_i$$

where $\{e_{jk}\}$-is the number of edges present.

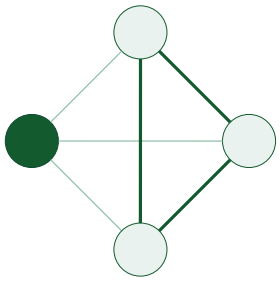


Figure 5.3: A node's (dark color) clustering neighborhood

Figure 5.3 provides an example of a local clustering coefficient on an undirected graph. The local clustering coefficient of the dark node is computed as the proportion of actual connections between its neighbors in relation to the number of all possible connections. In figure 5.3, the dark node has three neighbors, which can be linked by a maximum of 3 connections. In figure 5.3, all three possible connections are realized (thick black lines), giving a local clustering coefficient of 1. If we remove two of the black lines and only one connection is realized, the local cluster coefficient is 1/3. Finally, if we remove all three connections between the neighbors of the dark node, we obtain a local clustering coefficient value of 0 for the dark node. By averaging all individual local clustering coefficients, we can obtain a measure of how cohesive the network is, and thus how much structural bonding social capital the group has acquired. For example, figure 5.2 shows that group B has a high number of ties between its members, and that these members are very often in a clustered neighborhood. This results in a high clustering coefficient (0.72) for group B, as reported in table 5.2.

Relational Dimension

Measurement by reciprocity on a group level

Generally, the relational dimension of social capital for a group can be captured by the amount of *reciprocity (r)* (Rao and Bandyopadhyay, 1987; Wasserman and Faust, 1994) present in the group. The relational dimension at the group level can be determined either by measuring the average individual reciprocity of each group member (see section 5.2.1) and then averaging this metric to obtain the average group reciprocity; or, it can be determined simply by computing it at the group level: In the arc method (Wasserman and Faust, 1994)[p.507] (see below), the number of reciprocated ties (where A and B share information with each other) is counted and then expressed as a ratio (r) of all possible ties in the group. For example, group A has 3 reciprocated ties and 12 total ties, thus a reciprocity of 3/12 or 0.25 as indicated in table 5.2. Reciprocity can both be measured for the network of interactions (AT) and the simple attention network consisting of friend and follower ties.

$$r = \frac{T^{<->}}{T}$$

Outcome

Measurement of information diffusion on a group level

The information diffused between the members of a group can be measured by the group density[4] in the retweet (RT) network. This means that the ratio of the total number of existing retweets is compared to the total number of all retweet traces. Group density is used as the outcome variable because it is the ratio of the total number of existing ties and the number of all possible ties. This means that it takes into account how much information could have been exchanged between group members if all members did exchange information. Using group density to measure the outcome is slightly different than simply counting the volume of retweets: indeed, the volume of retweets does not take into consideration that information should be exchanged equally between group members in order to obtain a high score.

Conclusion

Using a standard network toolkit[5], it is possible to compute the structural and relational network metrics of the FF and AT network. The outcome is computed using the RT network.

[4] Since all groups in the example network and in the final data set contain the same amount of nodes, the group density does not need to be normalized for group comparisons to be made.

[5] In this work, all network measures were computed using NetworkX, which is a free open-source python network library and can be downloaded at: `http://networkx.lanl.gov/`

Group Name	Number of Nodes	Structural Metrics			Relational Metrics
		Density	Average path length(bin)	Clustering (bin)	Reciprocity (bin)
A	8	0.21	0.68	0.38	0.25
B	8	0.36	1.38	0.72	0.15
C	8	0.23	1.07	0.38	0.15
D	8	0.23	0.66	0.43	0.07

Table 5.2: Group bonding social capital measures for the example network

To determine these group bonding social capital measures, I computed them for the example network shown in figure 5.2[6]. The results were summarized in table 5.2 below.

Table 5.2 shows that group B has the highest density (0.36) and the highest clustering coefficient (0.72), although it also has the highest average path length (1.38). One can therefore consider that this group has the overall highest structural bonding social capital (with the exception of its average path length). However, group A has the highest relational social capital, because it the highest amount of reciprocal ties (0.25).

Figure 5.4 shows an overview of the measures proposed as well as their influence on the outcome variable. It shows how the relational and structural metrics are expected to positively influence the amount of retweets within the group's network.

5.2 Influence of individual bonding social capital on information diffusion

5.2.1 Hypotheses

Influence of individual structural social capital on information diffusion

The review of social capital has highlighted how the embedding of individuals in the group creates individual bonding social capital, which, in turn, allows these people to obtain certain

[6]Whenever a metric was computed on the binarized version of the network, the variable was marked with (bin) postfixes.

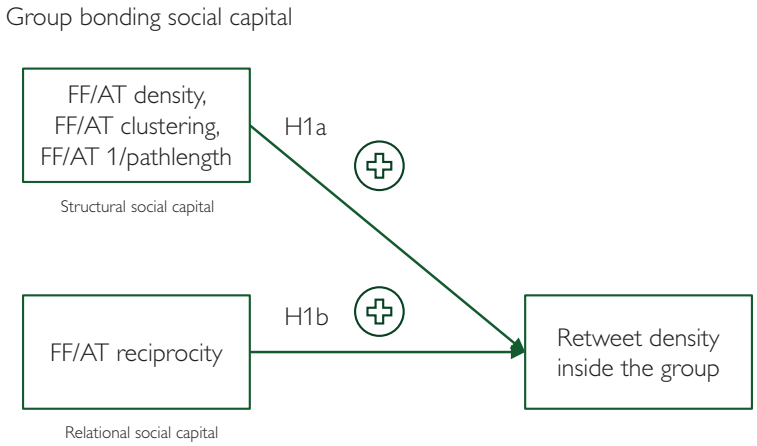


Figure 5.4: Overview of the group bonding hypotheses

goals or goods (Coleman, 1988; Adler and Kwon, 2002; Borgatti et al., 1998; Burt, 1999). Yet there is also overwhelming evidence that individual structural social capital should have a positive effect on information diffusion within the group. In the discussion of how individual social capital leads to diffusion, Burt (1999) referred to Rogers' observations (1995) on the spread of information in communities, and noted: in order to reach the entire system, i.e. the group, structurally central individuals need to maintain connections to less central individuals. Indeed, by maintaining these connections, opinion leaders ensure that their information will be diffused rapidly and effectively. Similarly, Wasko and Faraj (2005), who studied online social capital, noted that "[individual] structural capital is relevant for examining individual actions such as knowledge contribution, within a collective. Individuals who are centrally embedded in a collective have a relatively high proportion of direct ties to other members, and are likely to have developed this habit of cooperation. Thus, an individual's structural position in an electronic network of practice should influence his or her willingness to contribute knowledge to others."[p.39]. Similarly, the literature review on information diffusion revealed that numerous studies also predict that structurally central individuals can positively influence information diffusion within a group (Bass, 1969; Gladwell, 2000; Katz and Powell, 1955; Lazarsfeld et al., 1965; Merton, 1957; Schenk, 1993; Valente, 1993; Weimann, 1994). Finally, the existing research on Twitter highlights the structural role of individual social capital by offering different indicators, such as the global number of followers (Cha et al., 2010), the

PageRank index (Kwak et al., 2010) or the TunkRank (Tunkelang, 2009)[7]. These studies all agreed that the network centrality of individuals has a positive influence on the amount of retweets they are able to receive. However, the studies neglect the fact that the potential influence of structural opinion leaders or influentials might only be limited to their own group. Only Weng et al. (2010) acknowledge that a person's influence might be limited to a certain topic. In order to shed more light on topical opinion leaders, the following hypothesis will focus on the influence of an individual's structural capital on the amount of retweets they receive from members of their interest group:

Hypothesis 2a: The more structural individual bonding social capital one realizes in his own group, the more one's tweets are retweeted inside the group.

Influence of individual relational bonding social capital on information diffusion

The review of social capital also suggests that the relational dimension of social capital strongly influences information diffusion: Rogers (1995) stated that the diffusion of innovation through strong connections can lead to a higher chance of adoption. Such ties raise the strength of social influence during the interpersonal exposure of the individual. This means that information will have a higher change to diffuse through such strong connections. Nahapiet and Ghoshal also argued that "strong network ties influence both access to parties for combining and exchanging knowledge and anticipation of value through such exchange"[p.252] (1998). Burt noted (2000) that "a leader with strong relations to all members of the team improves communication and coordination"[p.393] and is based on the premise that people will exchange information because they have strong ties. The diffusion of innovation and information also suggests that the relational dimension in the form of strong ties leads to certain outcomes: Wejnert stated that strong ties drive network-based decisions (2002), because they exert pressure towards conformity[8]. Granovetter (Granovetter, 1973) also noted that inside groups, one will find a "lively discussion" that will remain inside the group, because the majority of information is only exchanged between individuals with strong ties. Moreover, recent studies on the strength of ties in electronic networks (Onnela et al., 2007) showed that stronger ties lead to an increase in information diffusion. Finally, the research on the influence of strong ties on information diffusion in Twitter showed that the relational dimension also plays a role due to Twitters strong conversational character (Galuba et al., 2010; Crawford, 2009). In the review of relational aspects driving information diffusion on Twitter[9], multiple studies (Suh et al., 2010; Yang and Counts, 2009; Cha et al., 2010) highlighted the influence of mentions on the number of

[7] Other attempts included the temporal order of information adoption (Lee et al., 2010), or algorithmic solutions (Romero and Kleinberg, 2010; Gayo-Avello, 2010).

[8] Although people differ in their tendency to follow social norms, cohesively-tied individuals have a higher chance of influencing each other, thus creating a common norm they believe in.

[9] See section 3.4.2 in chapter 3.4

retweets obtained. Interestingly, only a few of these studies investigated such ties in the context of homogenous groups of users with the same interests. In one study, researchers found that reciprocity between users had a positive effect on the number of retweets received (Kwak et al., 2010). In another study, researchers found results that contradicted this assertion and suggested that reciprocal relations lead to less information diffusion (Weng et al., 2010; Cha et al., 2010). In order to further investigate this question, the following hypothesis will explore to what extent relational individual social capital influences the information diffusion of individuals from the same interest group:

Hypothesis 2b: The more relational individual bonding social capital one realizes in his own group, the more one's tweets are retweeted inside the group.

Influence of individual cognitive social capital on information diffusion

The literature review also showed that the individual bonding cognitive dimension positively influences information diffusion (Coleman and Katz, 1966; Putnam, 1995): For example, Burt showed that individuals who share the cognitive dimension of the group are likely to help diffuse innovation inside the group (1995). From an information diffusion perspective, Becker noted that such opinion leaders are able to successfully diffuse information within the group because of their personal alignment with the group norms and their high compatibility with the group (1970). Similarly, research from the field of interpersonal information diffusion suggests that the opinion leader position (Weimann, 1994; Schenk, 1993; Katz and Powell, 1955) stems from a cognitive social capital dimension where opinion leaders tend to be more experienced in a certain topic (e.g., politics), and this cognitive advantage has an influence on the consequent diffusion of information. The network researchers Rogers and Kincaid (1981) showed that individuals who are more likely than others to perceive cognitive group norms and expectations might be better at diffusing information within the group. Interestingly, the literature on information diffusion in Twitter so far offers mixed results regarding the influence of cognitive social capital on information diffusion: while some researchers (Zaman et al., 2010; Bakshy et al., 2011) found that the contents of a message (which corresponds to the expression of an individual's cognitive dimension) had no influence on retweets, others found that sentiment (Hansen et al., 2011) or certain message contents (Wu et al., 2011; Suh et al., 2010) can positively influence information diffusion. The following hypothesis will therefore explore to what extent the cognitive individual capital dimension can predict how well that person's tweets are retweeted within the interest group. The more the person shares the group's interest, the more his information will be retweeted by other group members. Therefore, the third hypothesis in this set can be formulated as:

Hypothesis 2c: The more cognitive individual bonding social capital one realizes in his own group, the more one's tweets are retweeted inside the group.

The measures that will be used to capture the structural, relational and cognitive individual bonding social capital will now be introduced. All of these measures are based on standardized network measures that have been widely used in the studies presented earlier.

5.2.2 Measurements

Structural dimension

The structural bonding social capital of an individual can generally be defined as the sum of the ties an individual shares with members from the same group. There is a standard set of network metrics that can express the structural individual social capital of actors in groups as measures of network centrality. In this case, it becomes important to distinguish between incoming and outgoing ties, since only one sort of tie contributes to the creation of social capital:

- Outgoing ties: If person A follows a person B, A creates an outgoing tie. A gets the chance to read what B writes and so obtains more information from that person's network. Yet this tie does not create individual social capital for person A that could help A in the diffusion of his or her information. Why is that so? Since B is not following A, no matter what A writes, B will never have the chance to read it and forward it to others. Thus, it is highly unlikely that A will obtain a retweet from B.
- Incoming ties: If person B follows person A, A receives an incoming tie. This tie means that B potentially reads what A is writing and can thus choose to forward A's message to his followers. Therefore, only the incoming ties of person A actually contribute to his or her individual bonding social capital. The more incoming ties A receives, the more individual bonding social capital he or she has on Twitter.

Following these two definitions, all centrality metrics will be computed on a directed and weighted network, with the *incoming* centrality metrics expressing the social capital of an actor. This means that, for example, instead of computing a general degree for an actor, we compute the indegree, because this measure contains more information about an actor's social capital. If not otherwise possible (since some measures cannot be computed on directed networks), the measures will be computed on symmetrical versions of the networks. The standard centrality measures used to express this type of centrality are degree-, closeness- and Eigenvector- or PageRank-centrality (Wasserman and Faust, 1994), which will be now introduced in detail[10]. As introduced in chapter 4, the centrality metrics will be computed for the friend and follower

[10]For a review of centrality metrics for weighted networks, see Opsahl et al. (Opsahl et al., 2010)

(FF) network, since it captures the amount of attention an individual receives, and for the interactional (AT) network, since it captures how central the actor is in the discussion.

Measurement by indegree centrality

The *indegree* (C_D) (Wasserman and Faust, 1994)[p.126/127] of an actor is defined by the direct incoming connections he or she maintains with other actors[11]. In the social capital literature, indegree is considered as an indicator for this actor's communication activity and an indicator of his social capital (Rogers and Kincaid, 1981). In order to compare the degree between different networks, it must first be normalized. The normalization of the indegree is called the degree centrality C_D (shown below). If networks have the same amount of actors, the direct degrees can be compared. For the FF network, the degree was normalized, whereas for the AT network the direct weighted degree was used.

$$C_D = \frac{deg(V)}{|V| - 1}$$

For example, in figure 5.2 in group A, actor a3 has the highest direct indegree of 4, since he has two incoming ties with the strength of one, and one incoming tie with the strength of 2, which comes out to a total indegree of 4.

Measurement by closeness centrality

Another form of structural network centrality is expressed by the *closeness* (C_C) (Wasserman and Faust, 1994)[p.184-186] centrality metric of a node. Closeness centrality is defined as the mean geodesic distance (d_G) (i.e., the shortest path) between a node v and all other nodes it can reach. It can be used to describe how efficiently an actor can access information in the network. Nodes that are "shallow" in comparison to other nodes (that is, those that tend to have short geodesic distances to other nodes in the graph) have a higher closeness. Closeness centrality gives *higher* values to more central nodes.

$$C_C = \frac{1}{\sum_{t \in V/v} d_G(v,t)}$$

For example, in figure 5.2 in group A, actor a2 has a closeness centrality of 0.78, which indicates that he is the most central actor of this group. He is closest to all of the other members in the group. In contrast, while actor a3 has the highest indegree, he is less central, with a closeness centrality of (0.14).

[11] In directed networks, we can distinguish between an outdegree which is defined by the outgoing connections, and an indegree which is defined by the incoming connections.

Measurement by PageRank

The two metrics that were presented are able to measure the direct social capital of an actor, meaning that they only take into account the direct followers of a Twitter user (e.g., indegree). In contrast, the PageRank metric (C_P) (Page et al., 1998) is able to measure the *indirect social* capital of a Twitter user, by intuitively taking into account the social capital of his or her followers, and ensuring that followers with more social capital contribute more to it than the ones that have less[12]. Generally, the PageRank computes a ranking of the nodes in the group based on the structure of the incoming links. The PageRank is a variant of the Eigenvector centrality measure (Bonacich, 1987), a network metric that is also able to compute the indirect social capital, but does not guarantee that it can be computed for all nodes. In contrast, the PageRank was specifically designed for directed graphs and has the advantage of always converging. The idea behind the PageRank algorithm is that it gives high scores to nodes that exert a "hidden" influence, which means that they are connected to other nodes that are very central. Thus, highly structurally central actors with a high indirect social capital also obtain high PageRank scores. In group A of the example network in figure 5.2, actor a4 has a relatively high PageRank (0.30) although this actor neither has a high indegree (1.0) nor a high closeness centrality (0.14). His high PageRank reveals his high indirect social capital, which results from the fact that he is followed by actor a3, who in turn, is followed by the most central actor a2.

Relational dimension

The relational dimension of individual social capital depends on the reciprocity and tie strength of the individual in the network (Romero et al., 2011; Weng et al., 2010; Cha et al., 2010; Kwak et al., 2010). In this case, the amount of individual relational social capital in the Twitter network can be captured by analyzing reciprocity in the friend and follower (FF) network and in the interactional (AT) network. In the FF network, we speak of reciprocal friend and follower connections, and in the AT network, we speak of reciprocal interactions. Additionally, we can study the average amount of interactions in the interaction network (AT), also described as the average tie strength of an individual.

Measurement by average tie strength

The measurement of tie strengths in virtual social networks (Petroczi et al., 2010; Kumpula et al., 2007) relies on the frequency of interactions. As introduced in chapter 4, in the case of Twitter, the average tie strength relies on the conversational aspects of Twitter users. The social capital framework for Twitter has already shown that there are important differences

[12]The details of this metric can be found in the original works of Page et al. (Page et al., 1998).

between the declared Twitter user graph (FF ties) and the actual interaction graph (AT ties). According to the social capital framework, the more @ conversations flow between individuals, the stronger the tie between them. In order to compute the *average tie strength* (a_t) (Petroczi et al., 2010; Kumpula et al., 2007) for an individual i, the weighted indegree is added to the weighted outdegree of this individual and the total is divided by the number of total edges $|E(i)|$ the individual holds.

$$a_t(i) = \frac{indegree(i) + outdegree(i)}{|E(i)|}$$

In group A of the example network in figure 5.2, actor a1 has the highest average tie strength since two out of his four ties have the strength of two, resulting in an average tie strength of 1.5.

Measurement by reciprocity of ties

Researchers (Romero et al., 2011; Weng et al., 2010; Cha et al., 2010; Kwak et al., 2010) have also used the definition of reciprocal ties in Twitter to express the relational dimension of individual social capital. The higher the overall individual reciprocity of Twitter users, the more relational bonding social capital they possess in the group. Using an actor-centric approach, the *reciprocity* (r_t) (Wasserman and Faust, 1994)[p.507] of the actor's ties can be formulated as a ratio. This ratio is expressed as the proportion of all reciprocated ties of all internal group ties $|E(i)|$ minus the reciprocated ties of this individual. Here the reciprocated tie always consists of two edges (one outgoing, one incoming). Higher values of reciprocity express higher relational individual bonding social capital.

$$r_t(i) = \frac{reciprocated\ ties(i)}{|E(i)| - reciprocated\ ties(i)}$$

For example in figure 5.2 in group A actors a5 and a8 have the highest reciprocity as they have only one tie and this tie is reciprocated. Actor a3 has a low reciprocity because it has only one reciprocal tie, among his 4 edges. Thus his reciprocity is 1/(4-1) or 0.33 as reported in table 5.3.

Cognitive dimension

Measurement by ranking in interest

The individual cognitive bonding social capital dimension will be measured by the number of listings the individual collects for a given topic (see figure 5.5). This concept was introduced

with the social capital framework in chapter 4. Individuals are ranked according to the number of votes or listings for a given interest (e.g., an interest in running). These votes stem from lists that were collected for each interest. The details explaining how these lists were obtained are provided in chapter 6. Figure 5.5 shows an example: Among the three considered persons (P1, P2, P3), person P3 has the highest amount of listings because he is listed on three lists that list runners (L1, L2, L3), and is thus ranked number one for the interest in running. Person P2 has collected the lowest amount of listings (1) and is thus ranked third. The more listings an individual obtains from such lists with the same topic, the higher his ranking position is among similar individuals and the higher his *individual cognitive bonding social capital* is. Since the metric is described as "Ranking in interest", it indicates that the highest amount of cognitive overlap comes with the first rank in such rankings. Therefore, its influence on information diffusion is encoded as negative in figure 5.6.

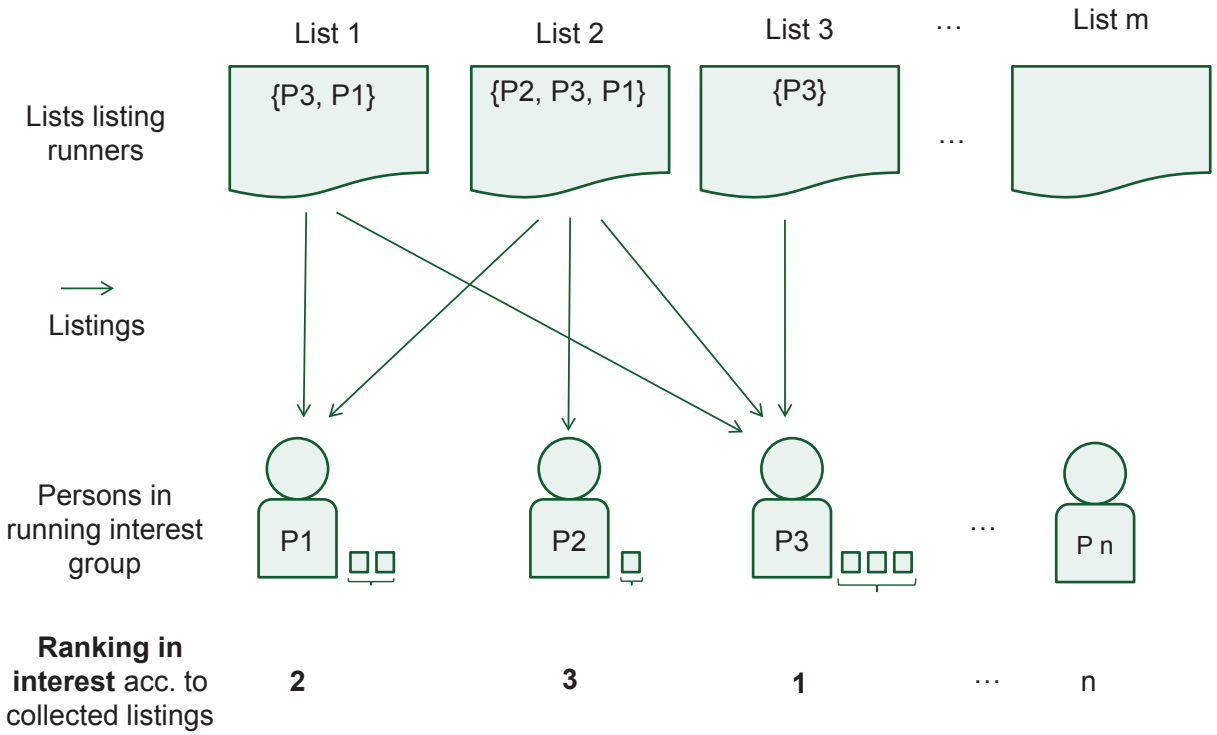


Figure 5.5: Overview of how the ranking in interest measure is computed. Figure shows lists that list runners for the running interest group.

Outcome

Measurement by received retweets

The outcome variable representing individual information diffusion is measured by using the number of retweets each individual obtained from members of his group. This corresponds to the weighted direct indegree in the RT-network for each individual in his corresponding group. Only the retweets of group members are considered here. As an example, in order to measure the internal information diffusion in group A, one would count the retweets a member obtained from group members a1 to a8.

Conclusion

In this section, I defined the metrics for structural, relational and cognitive individual social bonding capital. They are summarized in figure 5.6. The figure shows that the according hypotheses postulate their positive influence on the outcome variable. The outcome variable is expressed as the number of retweets that each individual obtains from other group members. The structural and relational network metrics were computed for the example network and are displayed in table 5.3.

Table 5.3 highlights the actors in the groups who have the highest structural and relational social capital. For example in group A, actor a2 has the highest closeness degree (0.78). Actor a5 and a8 have the highest reciprocity of 1, which means that all their ties are reciprocated, whereas actor a1 has the highest average tie strength of 1.5. Finally, actor a3 has the highest indirect social capital, resulting in a PageRank of 0.32. Overall actor a3 also has a high and indegree of 4, indicating that he possesses a strong structural social capital. Using these indicators of individual bonding social capital, the hypotheses in this section predict that the actors with the highest structural and relation social capital (marked in bold in table 5.3) will have the highest chances of obtaining retweets from their group members (group A in this case).

5.3 Influence of group bridging social capital on information diffusion

5.3.1 Hypotheses

The next section will reveal how I arrived at the hypotheses concerning the influence of group bridging social capital on information diffusion. Although many studies, especially those by Ronald Burt (2010; 1995), investigate the influence of individual bridging social capital on information diffusion — to see if structural brokers facilitate information diffusion between certain groups — a much lesser number of studies focus on this phenomenon at the group

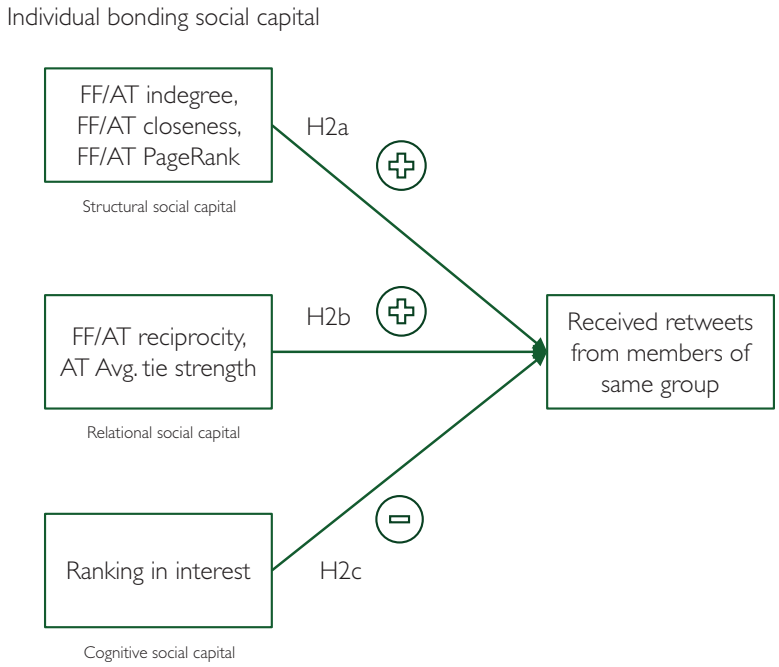


Figure 5.6: Overview of the individual bonding hypotheses

		Structural metrics			Relational metrics	
Group	Person ID	Indegree	Closeness	PageRank	Reciprocity	Average tie strength
a	**a1**	1.00	0.64	0.05	0.33	1.50
a	**a2**	3.00	0.78	0.09	0.40	1.29
a	**a3**	4.00	0.14	0.32	0.33	1.25
a	a4	3.00	0.14	0.07		1.33
a	**a5**	1.00	0.14	0.30	1.00	1.00
a	a6	1.00		0.09		1.00
a	a7	1.00		0.05		1.00
a	a8	1.00	0.47	0.05	1.00	1.00
...	...	...	...	...	...	...

Table 5.3: Excerpt of the individual network bonding metrics for the example network. Zero values not shown.

level. Most of the evidence for the existence of *bridging social capital* at the group level comes from the field of organizational social capital, where studies (Oh et al., 2006; Huber, 2009) analyze how teams, companies or departments create networks with each other and how these networks are beneficial to the group's success (see chapter 2). It is important to stress that this group social capital is different from individual social capital because it does not relate to the individuals but rather to the entire group, where no single member of the group can possess the group social capital. Linton Freeman (Freeman, 2004) pointed out that the entire group with all of its members must maintain heterogeneous connections to other subgroups in order to guarantee that the group's information is able to spread beyond the group to other groups. Relying on only few members to facilitate the spread of information to other groups might hinder the diffusion process. Once new information has reached a cluster, it rapidly spreads within the group because of the high interaction between members with strong ties and because of high group density. Borgatti and Everett similarly note (1999) that a group's bridging social capital is created by every member's relationship to the rest of the network. All these ties are an asset to the group. Indeed, if the members of a group hold enough ties to members of other groups, information that is contained within the group (and tends to be quickly exchanged among group members) is also diffused to these other groups (Granovetter, 1978; Burt, 1999). Therefore, the more brokers such a group possesses, the higher its group bridging social capital is, and the more this group's information will be spreading to other groups.

If a social network such as Twitter can be seen as consisting of multiple thematic groups that evolve around the users' interests (Wu et al., 2011; Kim et al., 2010; Zhang Y, 2012; Greene et al., 2012), then the most interesting questions relate to how these groups' bridging social capital influences information diffusion between such groups. Drawing analogies on to the observations of brokers (Burt, 2010, 1995) on an individual level, it is feasible to assume that whole groups in Twitter also tend to acquire brokerage positions among other interest groups (Rogers and Kincaid, 1981). In the literature on information diffusion in Twitter, this assumption has not been explored in great detail. The following set of hypotheses will seek to explore if groups that gain more structural, relational and cognitive group bridging social capital are able to better disseminate their information to other groups. Hypothesis H3a relates to the influence of the structural dimension of the group's social capital on information diffusion. It will explore how the group's position in the network influences the amount of retweets this group obtains from other groups. The following hypothesis postulates that groups that are highly central in the network of groups will have a better chance at obtaining retweets from other groups. Hypothesis H3b explores the groups' relational social capital as an aggregate of the relational social capital of its members to members of other groups. The more relational social capital the group's members can obtain, the higher the chances that the group's information will be retweeted by members of other groups. Finally, hypothesis H3c postulates that the groups' cognitive social capital will positively influence information diffusion. This means that the diversity in interests of the group members affects the chances of obtaining retweets from other groups. The more a group contains members with a high bridging cognitive social capital, the more retweets such a "cognitive brokerage group" is

expected to obtain. Since this thesis analyzes interest groups, the outcome of such an analysis not only reveals which group bridging social capital dimensions affect information diffusion at the group level, it also shows which interest groups are at the center of the Twitter ecosystem.

Hypothesis 3a: The more structural group bridging social capital the group realizes, the more its tweets are retweeted by other groups.
Hypothesis 3b: The more relational group bridging social capital the group realizes, the more its tweets are retweeted by other groups.
Hypothesis 3c: The more cognitive group bridging social capital the group realizes, the more its tweets are retweeted by other groups.

In order to statistically prove these hypotheses, multiple interest groups were collected and their social capital was compared. Whereas the group collection process will be described in section 6.4 of chapter 6, the network measures of group bridging social capital will be described below.

5.3.2 Measurements

Structural dimension

Generally, the bridging capital of a group is the sum of the bridging capital created by its members. This corresponds to the number of ties between groups. This observation was described in the theoretical network literature as intergroup centrality (Everett and Borgatti, 1999). Borgatti and Everett introduced a number of group centrality measures that describe the group's centrality according to the individual connections of group members to outsiders. These measures are: group degree[13], group closeness[14], group distance[15], and group betweenness[16]. Although these metrics have been suggested by various researchers for more than fifteen years, they have never been prominently used in social network analyses due to their complexity and lack of implementation (Borgatti et al., 2002).

Instead, group structural metrics are usually computed on a *blockmodel* (Heidler, 2006) version of the network. The blockmodel network is an intuitive alternative approach to the centrality metrics presented in the previous paragraph. It replaces all members of a group with a single vertex whose neighborhood is the union of the neighborhoods of all group members. The

[13]The number of outsiders tied to *at least one* group member.
[14]The distance of the group to all non-members.
[15]The distance of the group to outsiders is defined as the *minimum* distance between any insider and any outsider. The greater the distance to outsiders, the later information is available to the group.
[16]The number of times that the shortest path between any two outsiders includes a group member. Groups that score high on group betweenness generate exploitable structural holes.

internal group structure is then neglected and therefore does not affect the group's centrality. This also means that group membership is dichotomized (Heidler, 2006), which means that a member can either belong to a group or not. It is important to note that this block modeling approach can be performed according to the actors' attributes, but also according to the structural similarity of the actors[17]. In this work, actors are united according to their attributes (Heidler, 2006), namely their topical interests. This approach was used on the example network (see 5.2) and the results can be seen in figure 5.7. The color of the node group represents the color associated with the individual nodes in the example network. We see that the initial network of 32 nodes was reduced to a network containing only four nodes. Each of these nodes represents members with the same topical interest (in our case, represented by their color).

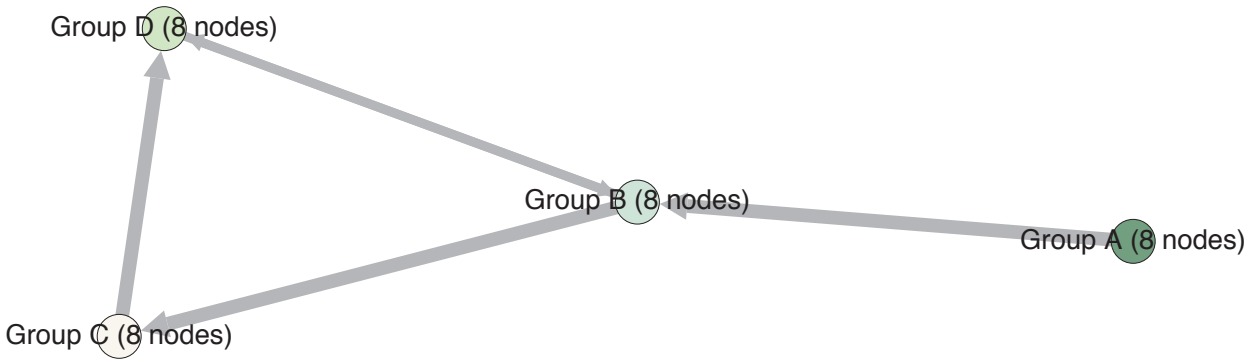


Figure 5.7: The blockmodel network in which nodes of the same group have been grouped into a single node. The individual ties between group members of different groups have been summed up into weighted ties between groups.

Centrality measures

In the resulting blockmodel network, (see figure 5.7) the same set of centrality measures can be computed as for the individual group bonding social capital. The only difference is that now the group is the unit of analysis, and the outer limit of observability is now the entire network of groups. This means we are now observing a network of interest-based groups in the Twitterverse, instead of observing individuals in their interest-based groups. The centrality degree measures for those groups are also computed on the directed network, since

[17] The search for equivalent actors is known under algorithmic methods such as CONCOR (Harrison et al., 1976) or REGE (Everett and Borgatti, 1994). Such structurally-equivalent actors are then united into one node. In-depth explanations can be found in the works of Wasserman & Faust (1994), Dorothea Jansen (2006)

the considerations about the direction of ties still hold in a blockmodel version of the network. The resulting measures of indegree centrality, closeness centrality and PageRank centrality are then computed for each group node and represent the overall structural group bridging social capital. The computation of these measures on the example network can be seen in table 5.4.

Since the betweenness measure expresses how certain groups or individuals can position themselves in the network as brokers that control information, the group's *betweenness centrality* (Freeman, 1977) in the network was calculated. The betweenness centrality of a group is measured in the same way as the betweenness of an individual: The betweenness of a node is defined by the amount of shortest connections between all members of the network that go through an actor, in our case a group. This measure describes the potential of a group to control the communication in the network. Groups that occur on many shortest paths between other groups have higher betweenness score than those that do not. In order to avoid duplication, the reader can find the formula definition of betweenness centrality in section 5.4, where the same metric is explained at an individual level. As an example, the metrics that were computed at the group level reveal that group B is one of the most central groups in the example network, with a closeness centrality of 0.67 while also having a high betweenness metric of 0.5. One would therefore assume that this group might not only succeed at spreading information to other groups because of its centrality but would also be able to broker information between groups, because of its high betweenness metric.

Relational dimension

The reasoning to compute the reciprocated ties and the average tie strength for groups follows the same previous considerations for individuals in chapter 5.2.1: In this case, the amount of group relational social capital in the Twitter network can be captured by analyzing reciprocity in the friend and follower (FF) network and in the interactional (AT) network on a group level. In the FF network, we speak of reciprocal friend and follower connections, and in the AT network, we speak of reciprocal interactions. Additionally, we can study the average amount of interactions in the interaction network (AT), also described as the average tie strength of an individual. To compute these metrics at the group level, the groups are reduced into one single node. This means that the metrics used for relational group bridging social capital are computed in exactly the same way as the metrics used for individual bonding social capital, except that in this case, the unit of analysis is now the group node. This means that the ties considered are those that exist between groups, rather than those that exist between individuals.

Average number of reciprocated ties

The average number of reciprocated ties is computed by counting how many reciprocal ties the group has with other groups. This metric is computed in exactly the same way as it is

for individuals (see reciprocal ties in section 5.2.1). For example in figure 5.7, group D has the highest reciprocity since it has one reciprocated tie to group B, and two non-reciprocated incoming ties from group C (see figure 5.2). Following the block modeling procedure, the ties to group C are merged into one non-reciprocated tie in the blockmodel network. This results in a reciprocity of 1/((2+1)-1) or 0.5 as reported in table 5.4 for group D. Group B has a reciprocity of 1/((2+2)-1) or 0.33 since it has 2 incoming and 2 outgoing ties in the blockmodel network and 1 reciprocal tie.

Average tie strength

The average tie strength of a group to other groups is computed in the same way as for individuals (see reciprocal ties in section 5.2.1), the only difference being that in the blockmodel network the group's tie strengths are an aggregate of the members' tie strengths. For example in figure 5.7, group A has one of the highest tie strengths since it has only one outgoing tie to the other groups (A $\rightarrow$ B). This tie has a total strength of 2, which is the result of the aggregation of the two individual ties to group B with strength 1.

Cognitive dimension

Measurement by aggregate number of interests

The group's cognitive bridging social capital dimension is measured by the *aggregate number of interests* that the group's members possess. At this point, it is important to point out that similarly to the structural and relational group bridging measures, this metric is simply an aggregate (a sum) of the attributes of the group's members interests: The more interests the individual has (as measured on how many different lists for different interests he is appearing), the higher the individual's cognitive bridging social capital, as this individual is perceived to be expressing multiple interests. By aggregating the interests of each individual in the group we can compute the overall cognitive bridging dimension of the group. The higher the number of interests the group's members possess, the more cognitive group bridging social capital the group has as a whole. While the cognitive bridging dimension of social capital indicator was already introduced in section 4.2.3 of chapter 4, the reader is also invited to skip forward to section 5.4.2, to find an example how this measure is explained in the context of individuals.

Outcome

Measurement by received retweets from other groups

The outcome variable is measured by the *number of retweets the group is able to obtain from other groups*. This means that all retweets a tweet might have collected from within the group are disregarded and instead, only the retweets obtained from members of other groups are counted. This measure automatically corresponds to the indegree in the blockmodel network, since self-loops (i.e., the retweets made by group members) are removed.

Conclusion

This section defined metrics for the structural, relational and cognitive group bridging social capital, which are summarized in figure 5.8. The structural and relational metrics were computed on the blockmodels of the FF and AT network, while the outcome variable was computed on the blockmodel of the RT network. The structural and relational network metrics were computed for the example network and are displayed in table 5.4.

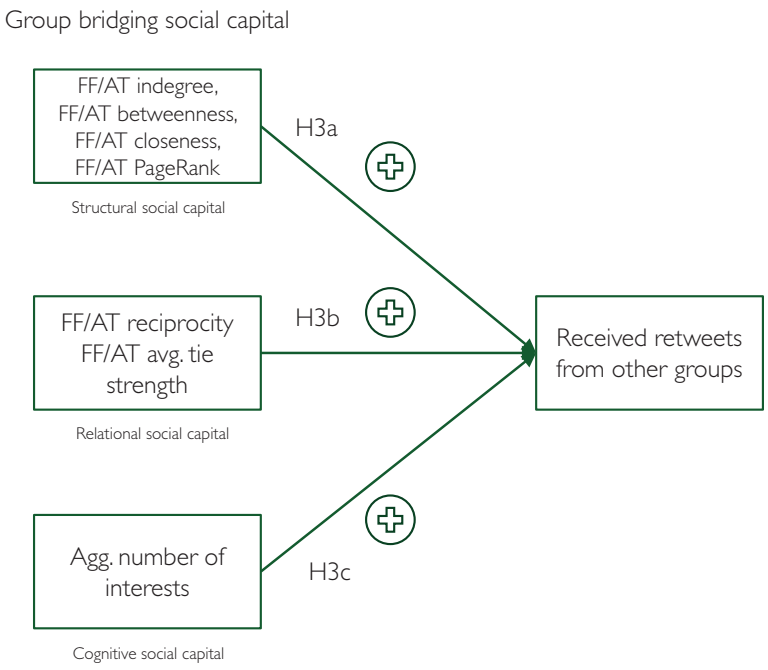


Figure 5.8: Overview over the group bridging hypotheses

	Structural metrics				Relational metrics	
Group name	Indegree	Betweenness centrality	Closeness centrality	PageRank	Reciprocity	Tie strength
A			0.60	0.04		2
C	2.00		0.44	0.25		2
B	3.00	0.50	0.67	0.37	0.33	1.5
D	3.00	0.17	0.44	0.35	0.5	1.33

Table 5.4: Summary of the group network bridging metrics. Zero values not shown.

Using the blockmodel approach and reducing the groups in the example network to only 4 nodes, it becomes apparent that group B has the highest betweenness centrality (0.5) of all the groups since it connects group A with the rest of the groups. Group B has the highest indegree of 3, the highest closeness centrality (0.67) and the highest PageRank (0.37) of all the groups. These metrics indicate that group B has the highest structural group bridging social capital. Regarding the relational group bridging social capital, group A and C obtain high values because they have higher average tie strengths to other groups, while group D has the highest reciprocity (0.5) in its ties to other groups.

5.4 Influence of individual bridging social capital on information diffusion

5.4.1 Hypotheses

Influence of structural bridging social capital

The next set of hypotheses deals with the influence of an individual's bridging social capital on individual information diffusion. The review on social capital shows that there is a long tradition of studies that explore individual bridging social capital (Coleman and Katz, 1966; Coleman et al., 1957; Lin, 2002) and the "bridges or access to bridges that facilitates returns in actions"[p.15] (Lin, 2002).The social capital literature has shown that individuals who are connected to other subgroups help transfer information between these subgroups. This was formalized by Ronald Burt (1999), who called these connections "structural holes". The

recent studies of Burt (2010) suggest that these patterns and mechanisms can also be found in electronic social networks on the internet and that people trade off such positions in terms of the amount of bandwidth they are able to process (Aral and Van Alstyne, 2009). Granovetter (1978) comes to similar conclusions from an information diffusion perspective, showing that weak ties between groups behave similarly to the ties that close the structural holes in networks Burt referred to. People that hold such weak ties and fill a network's structural holes and help transfer information between groups. Structural holes are therefore an opportunity to broker the flow of information between people belonging to different groups. The further a broker's connections reach into a network, the more easily this person can obtain new sources of information, and the less this person is constrained by his or her social surroundings in his or her group. In the literature on information diffusion, several researchers made similar observations, using the terms of Two-Step-Flow and Multi-Step-Flow of information (Katz and Powell, 1955; Lazarsfeld et al., 1965; Berelson et al., 1954). They observed that information is transmitted from the broadcasters (e.g., radio, television, newspapers) to the audience through intermediaries, which results in a Multi-Step-Flow of information. Moreover, research on information diffusion in Twitter has shown that such brokers tend to acquire individual bridging social capital by reaching out to strangers (Wu et al., 2011; Van Liere, 2010). Several studies have also provided evidence that information networks on Twitter are formed by "short-cutting" a two-step path between the source and the destination, and observed that 90% of new links on Twitter are created with people who are just two links away (Romero and Kleinberg, 2010; Yin and Hong, 2011). Therefore, the following hypothesis will postulate that individuals who acquire structural bridging social capital between different interest groups are also able to obtain more retweets from members of other groups. Hypothesis H4a will test if the structural dimension of bridging social capital or, in other words, a brokerage position, has an effect on the number of retweets an individual obtains.

Hypothesis 4a: The more individual structural bridging social capital one realizes, the more one's tweets are retweeted by members of other groups.

Influence of relational bridging social capital

Individual relational bridging social capital is mostly considered in social capital theory by observations that have been made in this context with the influence of tie strength on information diffusion. They often refer to the relational component of information diffusion, which was captured by Granovetter in his study of the strength of weak ties (1978). Indeed, he postulated that weak ties between groups contribute to individual relational bridging social capital. Under the notion of network bridges, Granovetter expected that these weak ties that connect distant groups would transfer a lower volume of information than ties that were formed within the group. Yet this information would be far more valuable to individuals since it contained new information that was not available to the group yet. Although Granovetter highlighted

the importance of the bridging relational dimension of such ties, the influence of individual relational social capital on information diffusion has mainly been explored for bonding and not bridging ties in the social capital context. Luckily, the literature on the relational dimension of Twitter offers insights into the influence of strength of ties on information diffusion between communities: Zhao et al. (2010) found that in an online environment, preferentially selecting weak ties to republish information does not make the information diffuse more quickly, whereas random selection can achieve this goal. However, when weak ties are gradually removed, the coverage of information drops sharply. They conclude that weak ties play a subtle role in the diffusion of information in online social networks: On the one hand, they act as bridges that connect isolated local communities and break the local trapping of information. On the other hand, selecting weak ties to republish information does not help spread the information further throughout the network. Grabowicz et al. (2012) also highlight the importance of weak ties between groups. They note that new information travels preferentially through links that connect different groups (which strengthens the weak ties), or even more so, through links that connect users who belong to several groups and act as brokers. Therefore, hypothesis H4b predicts that actors who accumulate a high relational social capital by holding reciprocal or strong ties to members of other groups will obtain more retweets from members of other groups.

Hypothesis 4b: The more individual relational bridging social capital one realizes, the more one's tweets are retweeted by members of other groups.

Influence of cognitive bridging social capital

The review on individual cognitive social capital showed that its effects are mostly described in a bonding context (Coleman, 1988; Putnam, 1995). In other words the we can expect that members that share the cognitive dimension of a group will receive more retweets from the group. Yet very few studies (Burt, 2010) suggest that members who share many different interests will receive more retweets from other interest groups. Most studies on online cognitive social capital followed the framework of Nahapiet and measured the influence of cognitive social capital *inside* the community (Yoon and Wang, 2011; Chiu et al., 2006; Lu and Yang, 2011; Bauer and Grether, 2005; Xiao et al., 2012; Wasko and Faraj, 2005). From the innovation diffusion perspective, the cognitive dimension is mostly seen as a property of the individuals in the adaption or innovation process inside the group (Rogers, 1995; Menzel and Katz, 1955). Indeed, Rogers suggests for example that depending on the compatibility of the message and the group norm, the diffusion process will be hindered or accelerated. This means that people who can transform a message in such a way that it becomes interesting to another group will obtain retweets from members of that group. From an information diffusion perspective, this can be applied to group norms as well (Rogers and Kincaid, 1981): Individuals that are supposed to transfer information from one group to the other must understand both group norms

and values in order to be perceived as cognitively relevant. In Twitter, there are indications that bloggers (Wu et al., 2011) with a high cognitive bridging social capital are disproportionately responsible for sharing information between different interest groups. Hypothesis H4c echoes hypothesis (H2c) on individual bonding cognitive social capital and postulates that individuals that are perceived as highly diverse in their cognitive dimension —i.e., they have a high number of interests and thus a high cognitive bridging social capital — will also have a higher chance of receiving retweets from members of other groups.

Hypothesis 4c: The more individual cognitive bridging social capital one realizes, the more one's tweets are retweeted by members of other groups.

5.4.2 Measurements

Structural dimension

In order to capture the structural *bridging* social capital of an individual, we postulate that it can only stem from individuals who are not part of the group of the individual, since individuals that are part of his group contribute to the persons *bonding* social capital. Therefore by definition, individuals can only create bridging social capital through ties to members outside of their group. Ties to members inside their group contribute to their individual bonding social capital, not to their bridging social capital. Similarly to the observations on tie directions in individual bonding social capital, the direction of bridging ties here is important since only incoming ties from other group members contribute to the individual's structural bridging social capital. Only people that follow an individual can actually retweet him, and thus allow the individual to obtain retweets. An individual cannot acquire individual bridging social capital by following people from other groups. Indeed, the people he follows might not be aware of him and thus cannot help him diffuse his information throughout the network: they are simply not receiving that person's information in the first place. In order to compute the structural metrics, both the friend and follower network and the interactional network are used and metrics will be computed for both. In the example depicted in figure 5.9, we see that if we want to analyze the ties that contribute to the social capital of actor b5, we have to remove all members of this group. The remaining ties from the other groups (in this case the tie a3 to b5) contribute to this actor's social capital.

The following structural network metrics are used to capture individual bridging social capital.

Measurement by ties incoming from other groups

The most basic measure for each actor representing the actor's structural bridging social capital is the *number of groups* (N_G) (Everett and Borgatti, 1999) that are connected to this actor.

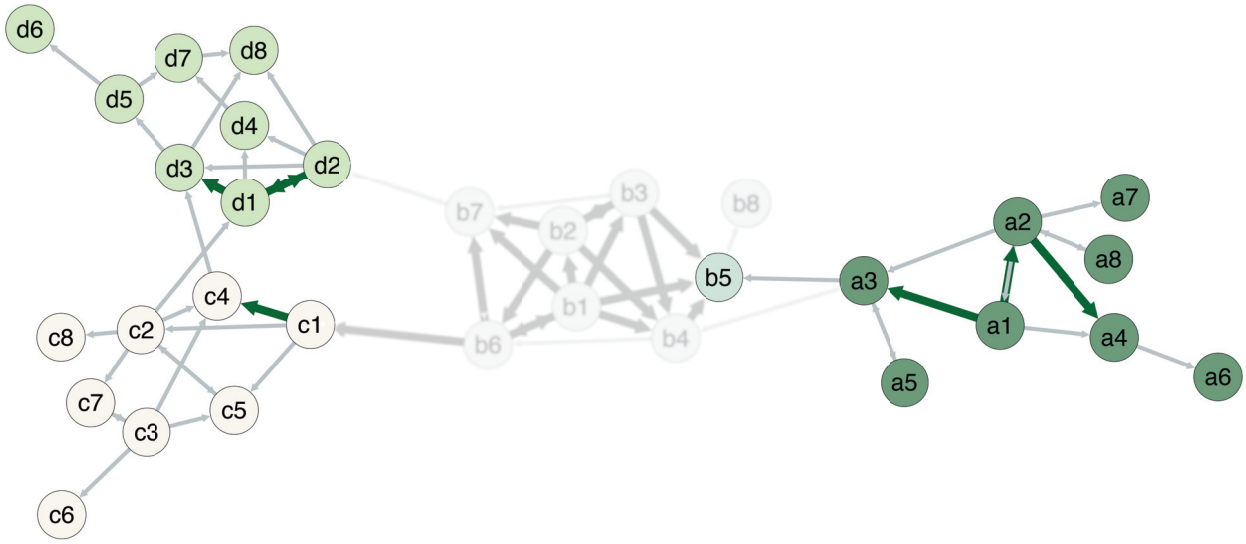


Figure 5.9: The subgraph used to compute the bridging social capital of node b5. Nodes b1-b4 and b6-b8 were removed (marked blurry and grey), since they did not contribute to b5's individual bridging social capital.

The maximum number of groups an actor can be connected to is reached when this actor is followed by at least one member of each group of all existing groups in the system, in our case the Twitterverse. The number of groups has a minimal score of 0 when nobody from the other groups decides to follow this actor. In this case, the actor might only have connections to members within the group and none to members from other groups. In figure 5.9, actor b5 draws the attention of only one member in group A, which means that the number of groups connected to this actor is 1. This metric can be computed both for the attention (FF) network and the interaction (AT) network.

Indegree

Since the incoming groups (N_G) discount the number of incoming ties per group, the *indegree* metric (Wasserman and Faust, 1994)[p.126/127] will be used to capture how much attention and interaction an actor is able to obtain from members of other groups[18] In comparison to the indegree for group bonding social capital, where all members of the same group contribute to the metric, the members are removed in order to capture the bridging social capital. This means that only actors from other groups can contribute to the measure (See figure 5.9). This allows the indegree metric to measure the amount of attention (FF network) and interactions (AT network) an actor obtains from members of other groups. For example in figure 5.9, actor

[18] Borgatti (Borgatti, 1998) showed that the indegree produces very similar results to the original *constraint* and *effective size* metrics that were originally introduced by Burt (Burt, 1999). I therefore decided to use this metric.

b5 has an indegree of one, since he or she is only connected to one other actor from all the other groups.

Measurement by betweenness centrality

A widely-accepted metric that captures brokerage positions is called *betweenness* (C_B) (Freeman, 1977). The betweenness metric intuitively captures how easily an actor can place himself between groups. The betweenness of a node is defined by the number of shortest connections between members of the network that go through a specific actor. This measure describes the potential of the actor to control communication in the network. Nodes that occur on numerous shortest paths between other nodes have a higher betweenness score than those that do not. The corresponding betweenness centrality C_B is computed as follows: For each pair of nodes (s, t), the shortest path between them is computed. Then for each node, the percentage of shortest paths that pass through this node are determined and summed up over all pairs of nodes, resulting in the formula:

$$C_B = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the number of shortest paths from s to t and $\sigma_{st}(v)$ is the number of shortest paths from s to t that pass through a node v. This measure can be normalized by dividing it with the number of pairs of nodes not including v, which is $(n-1)(n-2)$ for directed graphs, and $(n-1)(n-2)/2$ for undirected graphs. The betweenness measure that is computed on the entire network is then able to express the actor's brokerage position in the network. Since the betweenness measure has to take into account the entire structure of the underlying network, actors were not removed like in the case of the indegree metric. To illustrate this measure, actor b1 in figure 5.2 is lying on the greatest number of shortest paths between all members, which results in the highest betweenness metric for this actor, as indicated in table 5.5.

Relational dimension

The same logic that was applied to the measurement of relational individual bonding social capital (see chapter 4) will be applied to the measurement of relational bridging social capital. This means that the relational dimension of such bridging ties will be measured using either the strength of the ties or the reciprocity of the ties. This also means that both the attention and interactional network will be used to measure the relational dimension. Yet whereas in the analysis of bonding ties we examined ties between members of the same group, in the case of bridging ties we will only examine ties that exist between actors belonging to two different groups. In the example network, any tie that connects an individual from group A to members of the other groups (B, C, D) is considered a bridging tie.

Measurement by reciprocity of ties

The reciprocity metric that captures individual bridging social capital is computed in exactly the same way it was used to capture individual bonding social capital (see reciprocal ties in section 5.2.1). The only difference is that the ties that contribute to the metric come from individuals that are outside of the group, rather than from individuals inside the group. For example in figure 5.2 in group B, actor b7 is the only actor that holds reciprocal ties to members of other groups, resulting in a reciprocity of 1. All other actors of group B have a reciprocity of 0 (see table 5.5).

Measurement by average tie strength

The measurement of individual bridging social capital using average tie strength is performed in the same way as for individual bonding social capital (see reciprocal ties in section 5.2.1). The only difference is that the ties that contribute to the metric stem from members that are outside the group. For example, individual b6 has the highest average tie strength (2) since he has only one bridging tie with a weight of two to member c1 of group C.

Cognitive dimension

Measurement by number of interests

The cognitive bridging dimension will be measured by the number of interests the individual has. The conceptual aspect of this metric was introduced in chapter 4. The more interests an individual has, the higher that individual's cognitive bridging social capital is, since the individual is perceived to be expressing multiple interests. In the example in figure 5.10, person P1 has only one interest (cycling), since P1 was not listed on lists for other topics. Person P2 has an interest for running and cycling since he or she was also listed on lists listing cyclists. Person P3 has three interests (running, cycling and swimming) since this person was listed on lists for three different topics. Generally, people can exhibit as many interests as there are topics in the analysis. Details concerning the types and number of interests defined are provided in chapter 6.

Outcome

Measurement by number of received retweets from members of other groups

The outcome variable is measured by the number of retweets the individual is able to obtain from members of other groups. This means that all retweets a tweet might have collected from

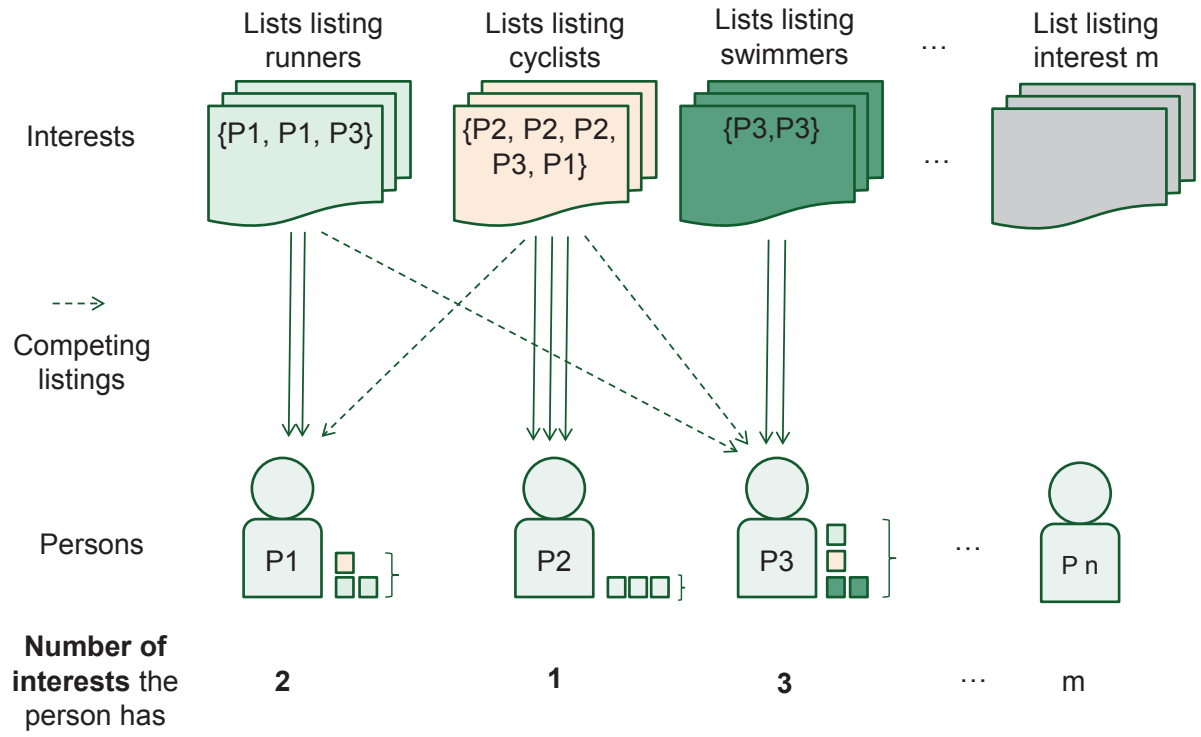


Figure 5.10: Overview of the calculation of the number of interests for each individual.

inside the group are disregarded; instead, only the retweets that resulted from members of other groups are counted. This measure automatically corresponds to the weighted direct *indegree* (Wasserman and Faust, 1994)[p.126/127] in the RT network, in which all actors from the same group as the individual were removed.

Conclusion

This section defined various metrics for the structural, relational and cognitive individual bridging social capital. They are summarized in figure 5.11. The outcome variable is also expressed as the number of retweets each individual was able to obtain from members of other groups. The structural and relational network metrics were computed for the example network and are displayed in table 5.5.

Table 5.5 shows that for most actors the bridging social capital metrics are 0 (not displayed in table 5.5) because the actors only hold ties to members of their own group. The table displays the metrics of the actors in group B. Here, the actors with the highest indegree and highest number of incoming groups are actors b4, b5 and b7 with a value of 1. Additionally, actor b1 has the highest betweenness centrality of 0.61. In the relational dimension, actor b6 has the highest average tie strength of 2 and only actor b7 holds reciprocal ties to members. According to the hypotheses, it can be expected that for group B, actors b1, b4, b5 or b7 will acquire the

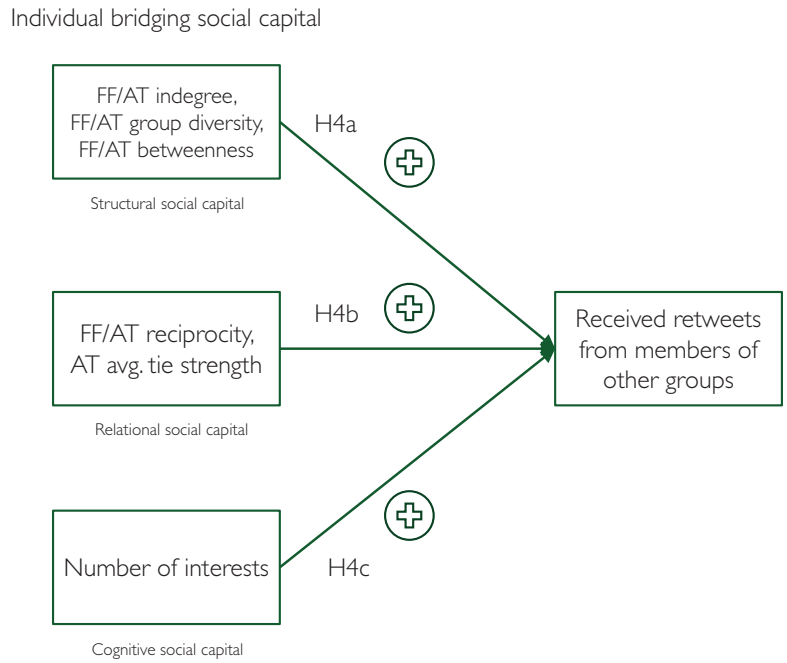


Figure 5.11: Overview of the individual bridging hypotheses

		Structural Metrics			Relational Metrics	
Group	Person ID	Indegree	# incoming groups	Betweenness	Reciprocity	Average tie strength
...	...	...	...	...	...	...
b	b1			0.61		
b	b2			0.36		
b	b3			0.18		
b	b4	1.00	1.00	0.12		1.00
b	b5	1.00	1.00	0.19		1.00
b	b6			0.24		2.00
b	b7	1.00	1.00	0.15	1.00	1.00
b	b8					
...	...	...	...	...	...	...

Table 5.5: Excerpt of the individual network bridging metrics. Zero values not shown.

highest number of retweets from other groups.

5.5 Negative influence of bonding capital on information diffusion

5.5.1 Hypotheses

So far, the hypotheses postulated that social capital positively influences information diffusion. In contrast, the final sets of hypotheses will postulate that individual and group social capital can negatively influence information diffusion. Cultural social capital research has highlighted this "dark side of social capital" (Adler and Kwon, 2002; Gargiulo and Benassi, 1997), which describes how social bonds can be an obstacle to achieving certain outcomes, or can even be responsible for negative outcomes[19]. The path model in figure 5.1 shows the previously postulated hypotheses H1-H4 and introduces the new hypotheses H5-H8, which predict the negative effect of social capital on information diffusion for both groups and individuals.

This set of hypotheses is motivated by the discussion on homophily and social capital. Indeed, Lin (2002) suggested that homophile ties characterize expressively motivated actions and that heterophile ties are necessary for instrumental actions. Translating this observation to twitter we can expect that some groups might have a tendency to form networks with mainly homophile ties. Their group members will be focused on discussing a certain interest e.g., tennis or cycling with other interested group members. Yet other groups might tend to create heterophile ties to other groups with the goal to spread their information beyond their group (e.g., politics or journalists). These two types of ties correspond to closed or dense networks, in contrast to bridging or loose networks. Burt found that the two types are both empirically likely to occur together, although theoretically their mechanisms remain distinct (1999). At the group level, this means that groups with a lower tendency for homophily create more bridging ties to other groups, and that groups with a stronger tendency for homophily create more ties within their groups. Therefore, group bridging social capital is created by the lack of homophily in the group and group bonding social capital is created by the tendency for group homophily.

Among the various studies on the effects of homophily on information diffusion in Twitter (Weng et al., 2010; Kwak et al., 2010; Golder and Yardi, 2010; Bakshy et al., 2011; Takhteyev, Y, Gruzd, A, Wellman, 2010), one study in particular highlights this observation: indeed, Wu et al. (2011) focused on the influence of lack of homophily on information diffusion in Twitter and found that different groups possess different tendencies for homophily. For example, groups in the media, celebrities, organizations and bloggers categories tend to show significant homophily in the tweets exchanged between categories. When they analyzed how much information from each category was retweeted by other categories, they found that retweeting was also strongly homophile among elite categories. However, the group of bloggers was highly responsible for retweeting URLs that came from all categories. They found that this

[19] Such as susceptibility to infection (Cohen et al., 1997) or weak teamwork performance (Hansen, 2002)

result reflected the characterization of bloggers as recyclers and filterers of information. What the study implied is that groups with high tendencies towards homophily — i.e., tendencies towards building bonding group social capital — might receive less retweets from other groups.

In order to explore this effect at the group level, hypothesis H5 postulates that the creation of group bonding social capital has a negative effect on the amount of information the group is able to receive from other groups. The same argument can also be made at the individual level. Indeed, individuals decide for themselves if they want to create individual bonding or bridging social capital. Hypothesis H6 postulates that when individuals create more bonding social capital, this will have a negative outcome on the amount of retweets they can obtain from members of other groups. This hypothesis is interesting because it highlights the limits of topical opinion leaders in their groups. Therefore, the better people are at building up bonding social capital inside their own group, the worse they become at being equally "attractive" to members outside of their group.

Hypothesis 5: The more group bonding social capital the group obtains, the less retweets it obtains from other groups.
Hypothesis 6: The more individual bonding social capital the individual obtains, the less retweets he obtains from members of other groups.

5.5.2 Measurements

To measure these effects, I will use the best indicators from the bonding and bridging social capital models used in hypotheses H1 and H2. This means including the best indicators of structural, relational and cognitive dimensions in the final path model. The path model will assess if these bonding social capital dimensions affect external information diffusion, which will be measured by the number of retweets the group or individual is able to obtain from members of other groups.

5.6 Negative influence of bridging capital on information diffusion

5.6.1 Hypotheses

To test the negative influence of bridging social capital on internal diffusion, a similar set of hypotheses for bridging social capital was created. Although Burt (1999) and Lin (2002) note that bonding and bridging social capital are not exclusive and both might be obtained simultaneously, the hypotheses in this work postulate that the acquisition of bridging social capital might lead to negative effects on the internal information diffusion. The literature that

helped us arrive at these hypotheses is the same as the negative effects of bonding social capital on external information diffusion.

In terms of groups, one might expect that the more a group can pique other groups' interest in its topic — thus building more bridging group social capital — the more this development harms the group's ability to diffuse information within the group. This means that the creation of bridging social capital could have a negative effect on the amount of information the group exchanges within the group (H7). On an individual level, the assumptions leading to the final hypothesis are similar to this, except of course in this case individuals make their own decisions and don't act like a group. Analogously, hypothesis H8 postulates that if individuals invest too much effort into building individual bridging social capital, this might harm their chances of receiving retweets from members of their own group.

Hypothesis 7: The more group bridging social capital the group obtains, the less the group's tweets are retweeted inside the group.
Hypothesis 8: The more individual bridging social capital the individual obtains, the less retweets he obtains from members of his own group.

5.6.2 Measurements

To measure these effects, I will use the best predictors from the individual and group bridging social capital model. The predictors from the models used to examine hypothesis H3 and H4 will include the structural, relational and cognitive dimensions of bridging social capital. By examining their effect on the retweet density and amount of retweets obtained from other group members, we can assess how these dimensions affect the internal information diffusion.

5.7 Conclusion and goals

This chapter presented hypotheses concerning the influence of social capital on information diffusion. It also provided structural, relational and cognitive indicators for social capital in Twitter for each of the four quadrants. Hypotheses H1-H4 postulated that bonding and bridging social capital would have a positive influence on information diffusion, for both individuals and groups. Hypotheses H5-H8 postulated that social capital would have negative cross effects on information diffusion. For each set of hypotheses, the structural, relational and cognitive indicators were proposed. The applicability of the proposed metrics was demonstrated using the example network provided by the author in figure 5.2. A tabular overview of all the hypotheses, indicators and dependent variables can be found in table 7 in the appendix. In order to validate these hypotheses, interest-based groups are needed to form the basis of the empirical analysis. Thus, in chapter 6, a representative set of user interests in Twitter will be collected

and the corresponding groups of individuals that share similar interests will be compiled. The empirical validation of this set of 15 hypotheses will then be presented in chapter 7.

6 Collecting users' interest networks on Twitter

This chapter will describe which methods were used in order to obtain a representative dataset of users' interest networks on Twitter. The first part of this chapter will describe the sampling of users' interests and the second part of this chapter will describe the sampling of users that formed interest networks around those interests (see figure 6.1). The resulting interest-based networks form the database needed to empirically validate or invalidate the hypotheses that were previously postulated in chapter 5.

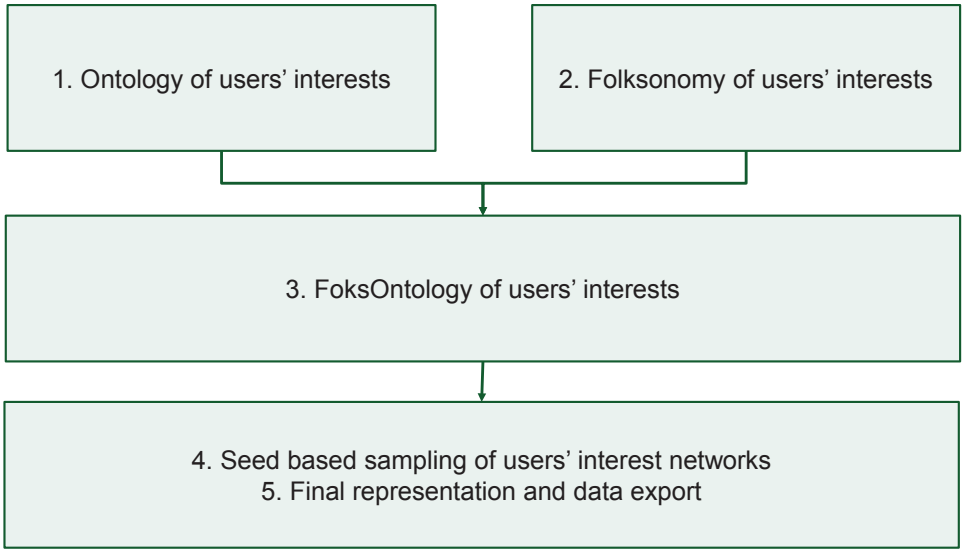


Figure 6.1: An overview of the steps that have been performed in order to collect users' interest networks.

Collection of users' interests

Since the basic assumption in this work is that different topical interests exist in Twitter, the first part of this chapter will describe the process of determining a meaningful and representative set of user interests in Twitter. The overall goal is to extract at least 150 meaningful user interests on Twitter, using keywords that describe these interests with the same level of detail[1]. This goal will be achieved in three steps (see figure 6.2): one top-down data collection step, one bottom-up data collection step, and one merging step.

The first step (see top left in figure 6.1) will review potential providers of users' interests that can be used to extract these interests for Twitter users. Using an already established ordering of concepts of users' interests not only reduces the risk of a certain sampling bias, but also allows integrating additional interests in a meaningful way in an ordered corpus. Here representatives from three potential perspectives will be evaluated: a) a flat (domain knowledge) ordering, b) upper level ontologies of users' interests and finally c) directories and domain specific ontologies of users' interests. The most promising provider will be chosen and a sampling of the most prominent users' interests will be performed. This corresponds to a typology (top down) approach, where one relies on a set of predefined categories.

The second step (see top right in figure 6.1) is a bottom-up taxonomy approach where users' interests will be collected directly from Twitter. This approach will be used to in order to complement the efforts of the first approach. To complete this task, Twitter lists will be used as a data source. The third-party providers of such interest lists will be reviewed in order to find the provider that best represents the variety of different topics that are exhibited on Twitter. The data from this provider will then be collected, and the problems that emerge from the collection of such unstructured data will be discussed.

The third step will combine the interest data collected in the first and second step and organize the obtained users' interests into the ontology resulting from the first collection step. The resulting outcome will therefore combine a (bottom up) Twitter-based folksonomy of users' interests with a (top-down) domain specific ontology of users' interests. The result will be a hierarchical ordering of Twitter interests that offers a representable sample of users' interests that describe users' interests at the same level of detail. This set of user's interests will then be used in the second part of this chapter to collect networks of people exhibiting these interests.

Collection of interest networks of Twitter users

Once a set of potential users' interests has been extracted, the second part of this chapter (see bottom of figure 6.1) will describe how interest-based networks of individuals for each of the interests have been extracted. This will show why a list-based sampling, as a representative of

[1] Obtaining at least 150 meaningful interests is a prerequisite to computing statistically significant results for the group-level hypotheses from chapter 5.

seed-based sampling approaches, is best suited to generate interest-based social networks. The following sections will provide the details of the chosen seed-based method and will discuss its results in terms of its stability on group size and the initial seed contacts. The final group size as well as the selection of seed contacts will both be determined based on the theory and empirical results. The final part of this chapter will describe how the explicit networks (the friend and follower network, the interaction network, and the retweet network) were obtained, and how the collected data was saved, represented and exported for further processing.

6.1 Obtaining ontologies of users' interests

The overall goal of the first part of this chapter is to obtain a representative sample of users' interests and provide an objective perspective on the underlying data. In order to achieve this goal, this chapter will first use a top-down and then a bottom-up data collection approach and then merge both results (see figure 6.2). Regarding the employment of top down data collection approaches researchers (Cucerzan, 2007; Mihalcea and Csomai, 2007; Kulkarni et al., 2009) have shown that ontology providers such as DBPedia[2], Wikipedia or WordNet[3] can be used as knowledge bases that exhibit users' interests. Regarding the employment of bottom up approaches, the chapter 3.4 on information diffusion in Twitter has already highlighted that some of such bottom up approaches have successfully been used to aggregate persons interests (Choudhury et al., 2010; Romero et al., 2011). Additionally there are also examples from bottom up approaches from other domains, where researchers (Cheng et al., 2008; Teng and Chen, 2006; Chen et al., 2010) determined the interests of bloggers by analyzing the content of their blogs.

6.1.1 Comparison of users' interest providers

Since there are different ways to obtain an ontology of keywords representing the users' interests on Twitter the different approaches have been reviewed in this section. The existing interest providers have been reviewed in order to find the best matching provider that is able to a) capture the wide variety of users' interests b) offers enough distinction in different categories c) is operationalizable on the Twitter network and d) can be trusted in regard to the selected terms and ontologies. Table 6.1 provides an overview of the surveyed providers of users' interests. It shows that there are different types of top down interest providers that can be used: Lists, generic ontologies and specific ontologies. The table also highlights a number of distinctive differences about the different types of providers: Firstly, the table shows that the different providers differ on how many keywords are used to distinguish among users' interests in the cognitive top-level category. Secondly, it shows that different strategies exist on how user

[2] http://dbpedia.org/About
[3] http://wordnet.princeton.edu/

interests can be organized: While some providers offer a hierarchical (as in DMOZ or Yahoo) ordering, other providers (as in Appinions or Peerindex) offer no hierarchical ordering between the keyword concepts, and some providers organize the concepts in a network perspective (as in WordNet). Thirdly the overview shows that most of the ontology providers (Yahoo or DMOZ) also contain the same number of top level categories (10-20). Although these categories are different in their exact word definition, they show a high overlap[4] in the semantic top level concepts, indicating that the final number of top concepts in an interest-based ontology should contain 10 to 20 items in order to provide enough distinction on the first cognitive level. Regarding the number of entries that are present in the second and third level of such interest providers, the table shows that the domain specific ontologies (DMOZ or Yahoo) also agree on the number of cognitive subcategories.

6.1.2 Flat lists of users' interest

The easiest form to obtain an overview of users' interests in Twitter, are flat lists of interest areas that have been created by domain experts. The two most prominent providers in this area are:

- Peerindex[5] is a social media analytic service based on footprints from major social media services like Twitter, LinkedIn, Facebook. It divides a user's interest into eight different interest areas (see table 6.1). Peerindex assumes that a user can have a footprint in any of these areas. Beyond the top-level categories Peerindex does not offer subcategories, creating a very shallow and broad differentiation of the interests of a user.
- Appinion is a social media monitoring service originating from academic research[6] at the Cornell University. The platform features an extensive database that includes millions of opinions extracted from blogs, Twitter, Facebook, forums, newspaper and magazine articles, and radio and television transcripts. The interests are differentiated in ten categories, which also map in a general sense to the categories provided by Peerindex (see table 6.1).

Although the available flat lists from these providers could be used to divide Twitter users' interest into 5 to 10 very broad categories, this approach would not be very fruitful because of two reasons: The resulting semantic relationships between the emerging interest-based communities would be very broad and the keywords that result from these providers would neglect a detailed view of user interests. Additionally, these providers do not provide the minimum 150 interests that are needed to compute a statistical significance on a group level. Although these providers are not useful for a direct operationalization, they provide valuable research on how the top levels of an ontology of Twitter interests should look like. The next section will review additional alternatives.

[4] Categories like Arts and Media, Education or Politics and News are prominent throughout all providers.
[5] A service similar to the `http://klout.com` service, which also provides categories of user interest.
[6] `http://www.cornell.edu/nyc/research.html`

Name	Appinions	Peerindex	Wordnet	Yahoo	DMOZ
Website	http://www.appinions.com	http://www.peerindex.com	http://wordnetweb.princeton.edu	http://dir.yahoo.com	http://www.dmoz.org
Type	List	List	Upper level Ontology	Specific ontology	Specific Ontology
Curated by	Individual	Individual	Professionals	Crowd & Professionals	Crowd & Professionals
Subcategories	no subcategories	no subcategories	includes subcategories	includes subcategories	includes subcategories
Related Tags	no related tags	no related tags	related keywords	related keywords	related keywords
Top level categories	10	8	1	12	16
First level categories	n.a.	n.a.	3	214	246
Second level categories	n.a.	n.a.	23	4283	3987
Top-level Keywords	Media	Arts & Media & Entertainment	Entity	Arts & Humanities	Arts
	Technology	Technology & Internet		Computer & Internet	Computers
	Education	Science & Environment		Education & Science	Science
	Health	Health & Medical		Health	Health
	Recreation	Lifestyle & Leisure		Recreation	Recreation
		Sports			Sports
	Politics	News & Politics & Society		Politics	News & Media
	Business	Finance & Business & Economics		Business & Economy	Business
	Travel				
	Entertainment			Entertainment	
	Fashion				
				Government	
				Society & Culture	Society
				Regional	Regional
				Reference	Reference
					Kids & Teens
					Home
					World
					Shopping
					Games

Table 6.1: Overview of the different types of top-down user interest providers.

6.1.3 Upper level ontologies of users' interests

In order to obtain a users' interest corpus with a hierarchical ordering and semantic relations between words, the applicability of upper ontologies[7] (or simply top-level ontologies) has been evaluated. Generally the use of an upper level ontology is promising (Laniado et al., 2007), since a high number of upper level ontologies exist[8] and such ontologies are domain independent, thus potentially applicable in any context. Among such ontologies the most commonly known are Wiki- or DB-pedia[9] or WordNet.

While upper ontologies are very powerful for word disambiguation their ability to derive a meaningful set of users' interests for Twitter is limited: Generally upper level ontologies lack the specific detail on user interests instead they describe very general concepts that are shared across all knowledge domains. This high generality leads to a main disadvantage of upper level ontologies for our task: Since they do not specifically focus on users' interests, but are describing the world on a very generic level, it is by definition impossible to find an upper level ontology that has the right amount of detail and covers enough potential users' interests. For example, in the WordNet ontology the top second level entities are "abstract entity", "physical entity" and "thing", and even on the third of fourth level of abstraction we do not find any related user interests. Thus while one of their main advantages of upper ontologies is that they are very good at describing the semantic relation between entities or concepts, this advantage only holds if such entries can be located in the ontology. Finally the level of detail that describes the relationships between concepts in ontologies also creates a second disadvantage for their applicability as a provider of users' interests: Many existing ontologies can often be too complicated for this approach, since they contain different classes, types and definitions of entities, which can also be connected by different definitions of relations[10]. These two main disadvantages point to the conclusion that the only application of upper ontologies in regard to this goal is that they might rather be used to disambiguate (Michelson and Macskassy, 2010) different existing keywords representing a user's interest, while they are impractical to deduct a predefined set of users' interests on the same cognitive level. In order to find ontologies that are specific enough domain specific ontologies have been reviewed.

6.1.4 Domain specific ontologies of users' interests or internet directories

While domain specific ontologies share the traits of upper level ontologies having a clearly defined set of hierarchies, specific ontologies in forms of internet directories provide a

[7] http://en.wikipedia.org/wiki/Upper_ontology

[8] http://en.wikipedia.org/wiki/Upper_ontology_(information_science)

[9] See http://wikipedia.org or its application interface accessible pendant DBPedia http://dbpedia.org

[10] WordNet is a large lexical database of English nouns, verbs, adjectives and adverbs. Words in WordNet are grouped into sets of cognitive synonyms (so called synsets), each expressing a distinct concept. Synsets are then interlinked by means of conceptual-semantic and lexical relations.

high level of detail of internet specific interests in their domain. Therefore, the final set of reviewed users' interest providers are domain specific ontologies, or in this case web directories. The advantage of the different web directories (e.g., Yahoo Directory, DMOZ Project, StartingPoint[11], WWWVL[12] or AboutUs[13], ...) is their maturity and their specific focus on users' interests. Additionally since such web directories are both curated by individuals, who submit entries and categories and professionals, with expert knowledge who review these submissions, the resulting ontology has both the hierarchical ordering (rigid typology managed by experts), while maintaining a certain flexibility since users are able to suggest new categories. Also web directories offer the possibility to add aliases (symbolic links) between related categories that are part of different sub trees, additionally allowing the creation of non-strictly hierarchical structure. Among this set of web directories two main web directory providers have been reviewed in order to verify their utility towards providing hierarchically ordered semantic relationships between the users' interests:

- Yahoo Directory offers as hierarchical ontology scheme with both a high number of top level concepts and sub-categories. Although entities are ordered in a hierarchical manner, there are also hierarchy transcending categories marked with @-signs that link between different hierarchical concepts. The generation of the categories and the subcategories is guided by a crowd-based process where users submit categories and entries to the directories, which are then successively indexed by experts.
- DMOZ (also called the open directory project) also uses a hierarchical ontology scheme for organizing entities. Entities on a similar topic are grouped into categories which can then include smaller categories. Similar to Yahoo Dir. it also contains a high number of similar top level user interests.

Both providers offer similar insights regarding their actuality, size, number of categories and rigor (see table 6.1). Out of these two directories the Yahoo Dir. was selected over DMOZ because of its better accessibility (the data can be easily obtained through a web scrape) and because the professional staff at Yahoo who are maintaining this directory full time, results in a higher rigor of categories and entries. Yet because of their similarity, both providers potentially provide the same level of detail on the entities and have a similar depth in their ontology. The proposed hierarchical concepts can be considered analogue in both ontologies and are apparent in other web directories, creating similar distinctions between users' interests.

Obtaining the ontology from Yahoo

Since the Yahoo Dir. offers no direct API access for the content of the website, it has been web-scraped in an iterative process for the first three levels of the directory. The results have been

[11] http://www.stpt.com/
[12] http://vlib.org/
[13] http://www.aboutus.org/

saved in a computer and human readable format[14]. The collection process was started from each of the 12 first level categories excluding the categories of "new additions", "subscribe via rss" and "regional", since they are not part of the user's interest representation. Additionally, it has been recorded in the data collection process if a category is a hierarchy transcending link, which links the categories that do not share the same parent node. The final corpus of user's interest contained 4350 concepts and 6703 edges linking those concepts.

The interests in this ontology are ordered in a three level structure, where the first two levels create high level structural entries, which are used to organize the upper levels of the directory, while the third level contains the actual user's interest (see figure 6.3). Some examples demonstrate this ordering: The concepts "skateboarding", "swimming", or "soccer" can all be found under the second level entry "sports" which is a part of the first level category "recreation". The entries "climate change", "sustainability", or "conservation" can all be found under the second level entry "Environment and Nature", which is a child of the first level "Society and Culture". The third level of the ontology therefore corresponds to the actual interests that people exhibit, while the first two categories are used in order to provide a structure among these categories.

Conclusion

The Yahoo ontology of users' interests meets the perquisites for a matching ontology because it a) captures the wide variety of users' interests (by containing over 4350 concepts specifically built to represent the interests of online users), b) offers enough distinction in different categories (contains 12 top categories) and c) can be trusted in regard to the selected terms and ontologies due to the professional staff organizing the entries. The most important benefit is that the keywords extracted from this ontology on the third level have the same level of cognitive granularity. Although this top-down typology approach yielded a high number of potential interests that can be evaluated in the Twitter network, the ontology might have missed some very important users' interests that might potentially only be found in the Twitter network. Therefore, in order to complement this approach the next section will show how the users' interests have been extracted in a bottom up approach.

6.2 Creating a users' interests folksonomy on Twitter

6.2.1 Users' interests lists providers on Twitter

In order complement the users' interests suggested by the Yahoo ontology, a bottom up or folksonomy (Halpin et al., 2007; Robu et al., 2009) approach has been performed, where

[14]The data contains information about the [level of category, name of category, number of sub entries, at_link]

interest categories are constructed bottom up from the available data. To determine users' interests on Twitter, the potential data traces (tweets, hashtags, lists) have been reviewed in section 4.2.3 in chapter 4 and the Twitter list feature has been shown to be the most promising concept to retrieve user interests. Since Twitter does not directly allow to obtain all available lists, third party providers[15] that allow a listing of all available lists in Twitter have been reviewed in this section. In order to select the best-fitting list provider, an overview of such providers has been performed. The results of this overview have been depicted in table 6.2.

The overview in table 6.2 shows that the majority of providers are based on a folksonomy approach aggregating the list data available in Twitter. Users can also use the providers' websites to add themselves to existing lists or create new keywords. Twellow provides the highest number of listed users (2.7 Mio.), yet artificially reduces the number of keywords originating from the lists to 242 keywords. Wefollow lists over 2122 keywords and over 1.71 Mio. persons, related keyword functionality and has no subcategories. Listorious lists 500 keywords, and over 100.000 persons that have been aggregated into these categories. It offers no subcategories and related keywords. Twitter itself offers a controlled representation of 27 keywords of users' interests resulting in a shallow categorization and is showcasing only 2000 selected users[20]. In the final decision for a keyword interest provider Wefollow was chosen over Twellow, Listorious or Twitter because of the higher number of available keywords, its more recent actuality and its higher popularity among visitors (1.8 Mio. users visit the website during the year). The decision for this provider is also supported by the fact that other academic work has also used this provider to for initial data acquisition (An et al., 2011; Zhao et al., 2011).

6.2.2 Data acquisition from Wefollow

The first step of the data collection process from Wefollow, was to collect all available entries that listed different interests. These concept keywords can either be sorted, according to a) the number of people listed for a certain keyword, b) the number of followers that people have that are listed for a certain keyword or c) by the alphabetical order of the keywords. Choosing the first alternative all keywords and the corresponding number of persons that have been listed for this concept have been retrieved[21]. This allows the measurement of the popularity of a certain interest, by the number of persons that are listed for this keyword. The retrieved list of 2099 different keywords, showed the typical lexical problems that result from a non-controlled vocabulary and unstructured data. This means that for the same concept similar terms exist that all refer to the same semantic origin. For example for the keyword music (69 382 members)

[15] Third party providers of lists usually list both users and lists by either a) the number of lists found for a certain keyword, or b) by the number of people that can be found in a certain list.

[20] This might be a rather strategic decision from Twitter, since explicitly very prominent users and the specific showcased categories (NBA, PGA, sports in general) have made this network very popular in the last years.

[21] Due to the lack of an API for Wefollow, a machine readable version of the websites data was web-scraped and saved in a comma separated text based format.

Name	Wefollow	Twellow	Listorious	Twitter
Website	http://wefollow.com	http://www.twellow.com	http://listorious.com	http://twitter.com/who_to_follow/interests
Curated by	Crowd	Crowd	Crowd	Professionals
Subcategories	no subcategories	subcategories	no subcategories	no subcategories
Related Tags	related tags	related tags	related tags	no related tags
Number of Visitors	1.8 Mio[16]	1.4 Mio[17]	700 000[18]	770 Mio[19]
Number of Keywords	2122	242	500	29
Number of Persons	1 764 473	2 730 570	101 151	2030
Top-Keywords (asc.)	Celebrity	Advertising	Socialmedia	Television
	Music	Education	Music	NBA
	Socialmedia	Entertainment	Politics	PGA
	Entrepreneur	Executives	Travel	Music
	News	Family	Food	MLB
	Blogger	Food	Photography	Nascar
	Tech	Government	Writers	Staff Picks
	Tv	Health	Seo	Entertainment
	Actor	Marketing	Arts	Sports
	Comedy	Music	Green	Faith and Religion
	...	...	...	...

Table 6.2: Overview of the different providers of lists on Twitter

numerous entries such as musician (11 799 members), musiclover (6304 members), musicians (158 members) exist or for the concept mother (545 members) other keywords such as mommy (10 000 members), mom (6205 members) exist. Therefore the corpus was cleaned up using a lexical similarity process that is described in the next section.

Clean-up process by lexical similarity

The cleanup process is a crucial part when dealing with unstructured data, due to the high noise that is contained in such corpus. The cleanup process was performed accordingly: All keywords that were shorter than three letters (e.g., C (programming language), IT (information technology)) were excluded from the list because they cannot be reliably searched in the Twitter list feature and are highly ambiguous in nature. The remaining corpus was then grouped according to the lexical similarity of the keywords. This meant a merge of 2122 concepts to 1267 concepts that share a lexicographic root. For each group out of the lexically grouped similar categories the concept with the highest amount of users is kept, while the other keywords representing this process are neglected. The rule to determine if two words were seen as lexicographic similar was:

- Both words start with the same letter (i.e. Travel and travel)
- One word is included in the other (i.e. tech and technology)
- Words are lexicographically "very similar" to each other like (i.e. journalist and journalism)

While the first two constraints can be implemented easily, the lexicographic concept of "very similar" has to be operationalized first. The operationalization of similarly written words is also referred to as morphological similarity and is based on a number of established methods: Different approaches are used to distinguish morphological variants of words are based on edit-distance metrics[22]: Among those, the most common approaches[23] the Levenshtein[24] distance was computed, because of its prevalent use in natural language processing and information retrieval. The Levenshtein distance is small for words that share similar stems[25]: A good example is journalist and journalism, where both words share the same stem "journal", resulting in a small edit-distance of 2 steps. The edit-distance was computed only for longer words (≥ 6) and with a threshold[26] of ≥ 4, meaning that these words can differ on the edit-distance up to

[22] `http://en.wikipedia.org/wiki/Edit_distance`

[23] e.g., also known as longest common sub sequence problem, Hamming distance or Jaro-Winkler-distance. These distance-edit methods show similar results.

[24] `http://en.wikipedia.org/wiki/Levenshtein_distance`

[25] `http://en.wikipedia.org/wiki/Word_stem`

[26] The determination of the threshold was chosen manually: A too small threshold (e.g., 2) leads to the omittance of detection of similarities between journalist and journalism. If the threshold it too large, non-similar words appear to share morphological variant of the word, but are indeed not similar, when classified by a human (e.g., interesting and investor).

four letters (4 insertions or deletions, or 2 substitutions of characters). The selected parameters above have yielded the best results in obtaining groups of morphemes. The morpheme groups revealed that they often contained wrongly spelled words (e.g., “entrepreneur”, “entrepeneur”, “entrepreneurs” or “entreprenuer”) or words that differ only slightly[27] in the singular and plural form (e.g., musician and musicians). The result of the selection and grouping process is a list of keywords, where keywords are grouped according to their lexical similarity and sorted according to the groups highest member count, which is shown in table 6.3 that offers an excerpt of the results. The final list has additionally been manually checked and corrected for positive errors (e.g., spa and space). The final list contains Twitter users’ interests that are sorted according to their prevalence in Twitter, therefore representing the most popular users’ interests on Twitter.

6.2.3 Conclusion

This section has described how the generation of a folksonomy of Twitter users’ interests was successfully performed. In the first step potential list directory providers have been reviewed and the candidate with the most number of concepts and a high number of listed persons was selected. Using a third party list provider successfully allowed to the acquisition of an overview of the exhibited user interests on Twitter. The resulting list keywords have then been collected and cleaned up using a lexical similarity approach. This approach was necessary due to the nature of the unstructured data. The final result is a list (excerpt shown in table 6.3) of Twitter users’ interests that are sorted according to their prevalence in Twitter, therefore representing the most popular user interests on Twitter. The final section will now discuss how the keywords from this list were merged with the Yahoo ontology.

6.3 Merging folksonomies with ontologies into folksontologies

This chapter will detail the process and the advantages resulting from combining an ontology and a folksonomy in order to create common corpus of keywords. This common corpus of keywords is also called a FolksOntology (Damme et al., 2007) - in this case a FolksOntology users’ interests on Twitter. This folksontology was created by merging both data sources against each other and so retaining only concepts that are highly relevant in both data sources. This process combines the strength of flat-tagging schemes and their ability to relate one item to others, while maintaining the ordering nature of ontologies. Both are hierarchical and valid,

[27] On an anecdotal note we find interesting lexical insights: when users refer to their interest in e.g., blogging in lists they like to create lists with the name “blogger” instead of “blog” since lists often refer to people as entities and it seems only natural to name such lists this way. Generally, it was also found, that only slight distinctions in morphological variants of the word often yield to much bigger user groups, as in the example of blog above.

Group	Keyword	Members	Distance
1	***music***	69 382	0
	musician	11 799	3
	musiclover	6304	5
	musicindustry	2354	8
	musicproducer	1819	8
	musica	1238	1
	musical	254	2
	musicaltheatre	192	9
	musicians	158	4
2	*blogger*	49 323	3
	blogs	3827	1
	blogging	3196	4
	blog	2452	0
	bloggerwannabe	584	10
	bloggers	301	4
	bloggermom	222	6
...	...	...	...

Table 6.3: Overview of the stemming process. Stems of a group are formatted in bold and keywords with most members are formatted in cursive.

but also represent the most used current Twitter users' interests, which allows the creation of a meaningful structure that represents the users' interests that describe them on the *same level of detail*. The merging process (see figure 6.2) was performed by first reducing the Yahoo ontology to the most prominent keywords (details see below), second reducing the Wefollow folksonomy to the most important keywords, and third merging both sets of keywords in the final FolksOntology (see figure 6.3).

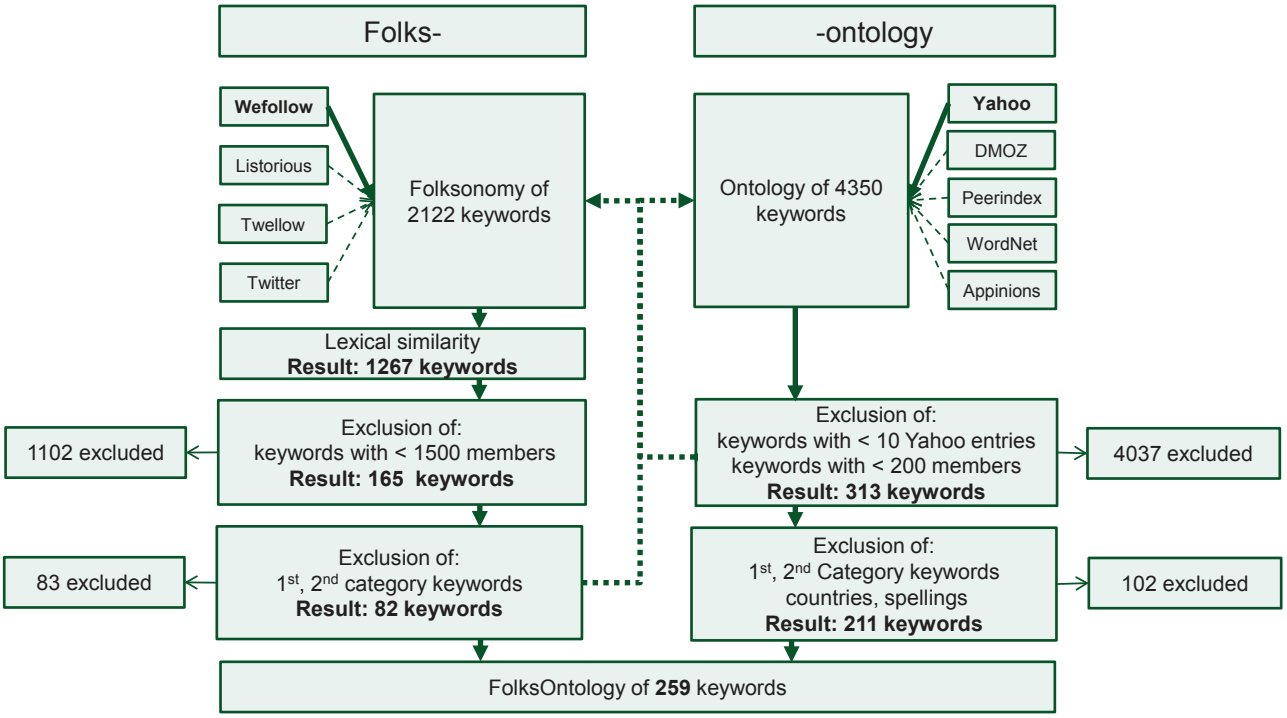


Figure 6.2: An overview of the keyword selection, reduction and merging process.

6.3.1 Reducing the Yahoo folksonomy

The advantages of obtaining the Yahoo ontology of user interests is that it creates a rigid, professionally reviewed structure, of users' interests that possesses a hierarchical ordering of among themselves. The downside of this approach is that the final corpus of entities is too big[28] for our needs (4350 keywords) and not every concept maps to an actual user interest. Therefore this corpus needs to be reduced in such a way that only the most prominent user

[28]For the empirical analysis only a corpus of 100-200 most prominent keywords is needed.

interests are kept. The reduction process contains two steps (see right column in figure 6.2): a filtering step and a merging step.

In the filtering process two operations have been performed: First the 4350 keyword concepts have been reduced to the most prominent ones. Here the categories that contained less than 10 website listings were dismissed from the ontology. Second, in order to find among the remaining concepts the ones that are highly represented in Twitter, every keyword of each Yahoo category was transformed to its lowercase version (e.g., Technology to technology) and for this keyword a lookup on the Wefollow folksonomy was performed. Keywords from the Yahoo folksonomy for which the Wefollow folksonomy reported less than 200 members[29] were then dropped from the initial set. These two procedures reduced the number of concepts to the 313 most important ones.

The second step further helped to reduce the 313 concepts by defining a set of exclusion rules: Firstly, all concepts of the first and second level of the ontology were dismissed as actual keywords (35 keywords). The reason for this is that the first and second level keywords are used as generic structural concepts in the Yahoo ontology. Therefore by excluding those keywords[30] a corpus of keywords is obtained that is able to describe interests at the same cognitive level of granularity, which is the third level of the Yahoo folksonomy. For example, comparing the granularity of the term "Arts" (1st level) vs. "graphicdesign" (3rd level) shows that these are descriptions of user interests at very different cognitive levels. It is important to point out that, without a predefined ontology with three levels of abstraction, ensuring that keywords describe interests at the same level of detail would not have been possible to achieve. Secondly, minor final corrections to the corpus of keywords have been applied: All concepts that described a country (e.g., Germany) or a region (e.g., California) were removed from the list since the goal is not to find geographical similarities between users (47 keywords). Finally, keywords existing in a plural and singular form or as a verb and subject were merged manually (e.g., humor vs. humorous or actor vs. acting (20 keywords)). The final remaining corpus contained 211 keywords.

6.3.2 Reducing the Wefollow folksonomy

The second step deals with the reduction of the number of concepts originating from the folksonomy which will then be merged with the remaining concepts in the Yahoo folksonomy. Section 6.2.2, has already described the lexicographic cleanup process (see left column figure 6.2) of the folksonomy that merged the number of concepts from 2122 keywords to 1267 keywords. For the remaining 1267 keywords two additional steps have been performed: A filtering step and a merging step.

[29] The threshold was manually determined to 200 members, as entries with less members did not provide enough seed data for the network extraction process (see chapter 6.4)

[30] There is one exclusion to this rule due to inconsistencies of the Yahoo ontology: Whenever the second category already included listings the category was then included. This is the case for the top categories "Social Science" and "Science").

The filtering step reduced the number of concepts only to the most prominent users' interests, by retaining only the keywords in the folksonomy that reported more than 1500 members. This led to a reduction of the corpus to 165 keywords. Reducing the corpus to only keywords with more than 1500 members ensured that only the most prominent interests were collected. Due to the immense effort that is needed to extract networks around those concepts (see section 6.4) only the most prominent users' interests can be investigated. Finally, since the remaining 165 keywords might also describe users' interests at different levels of granularity, the second merging step was performed. This step consists of only retaining keywords that are not part of the first and second level of the Yahoo folksonomy. The resulting final corpus of keywords originating from the Wefollow folksonomy was so reduced to 82 keywords representing users' interests on the same cognitive level of granularity.

6.3.3 Creation of the folksontology

The third step of this process was to include the 82 keywords originating from the Wefollow folksonomy into the Yahoo folksonomy in such a way that these keywords would be placed in their appropriate position in the final FolksOntology. This process was performed manually: First, for each Wefollow keyword a lookup in the Yahoo folksonomy was performed: if this keyword could be found in the folksonomy it was inserted at this position in the tree in the final FolksOntology (29 keywords). These 29 keywords are user interests that have held less than 10 entries in the Yahoo folksonomy, yet turned out to be highly important in the Wefollow folksonomy or in Twitter and were therefore included in the FolksOntology. For the remaining 53 keywords that could not be located in the Yahoo folksonomy a manual lookup in the Yahoo directory was performed and each keyword was placed in the category that was suggested by the Yahoo Query process[31] (e.g., the keyword search for "geek" results in an insertion at the branch Society and Culture → Cultures and Groups → Geeks).

The final result of 259 keywords represents a hybrid of the folksonomy of the most frequent Wefollow interests (representing the most prominent users' interests in Twitter) and the most popular entities in the Yahoo folksonomy (representing the most prominent users' interests in a controlled web directory). The result of this merging process is depicted in figure 6.3. It shows the first, second and third level structure of the final FolksOntology. On the third level, keywords that originated from the Yahoo ontology were marked with the prefix "yahoo:". Keywords that originated from Wefollow folksonomy have been marked with the prefix "wefollow:". Keywords originating from the Wefollow folksonomy that could be matched in Yahoo folksonomy are marked with the prefix "wefollow+yahoo:". Figure 6.3 shows that a high number of the keywords that have been found in the Wefollow folksonomy could be merged with the existing Yahoo ontology (29 keywords). It also shows that high number of concepts from the Yahoo ontology could be included in the final result (211 keywords). Finally, the final FolksOntology also highlights that while the Yahoo ontology succeeded in providing

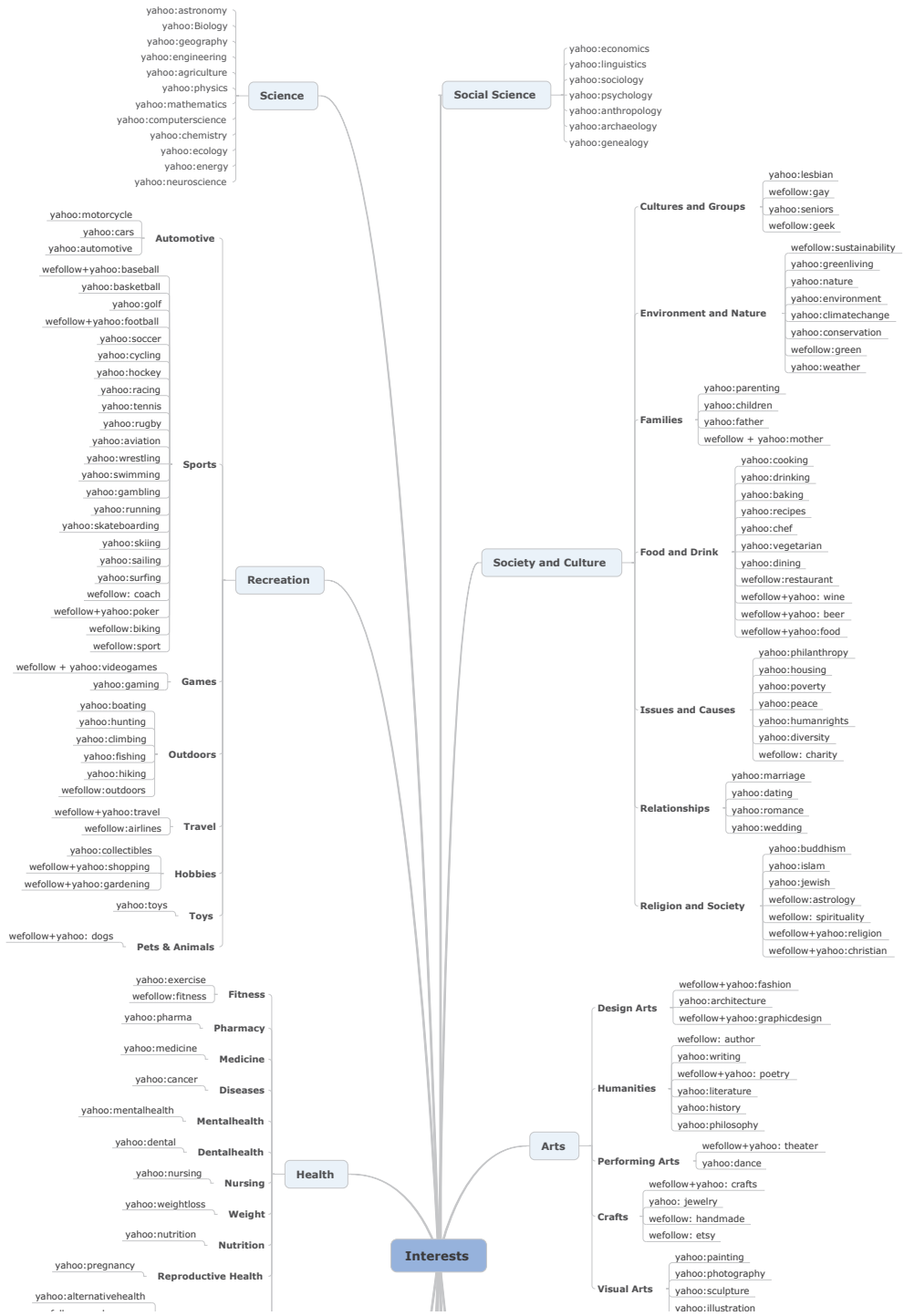
[31] `http://dir.search.yahoo.com/search`

a high order structure, some very frequent concepts of the folksonomy have been missing (53 keywords). This shows that the high effort that has been put into collecting those interests was appropriate, since otherwise those very prevalent user interests would have been missed.

6.3.4 Conclusion

The first part of this chapter described how a FolksOntology of 259 keywords (see figure 6.3) that represent users' interests on Twitter was created. The 259 keywords[32] represent a hybrid of the folksonomy of the most frequent Wefollow interests (representing the most prominent users' interests in Twitter) and the most popular entities in the Yahoo folksonomy (representing the most prominent users' interests in a controlled web directory). The keywords obtained this way, provide a representative sample of users' interests on Twitter, because they have been matched against the actual interests that are expressed on Twitter by using the Wefollow provider. These keywords also express the user's interest on the same cognitive level of detail, since only keywords from the third level of the Yahoo ontology have been used, and only keywords from the Wefollow FolksOntology have been used that map to this level of detail. Additionally keywords originating from the unstructured Wefollow folksonomy have been harmonized using a lexical similarity procedure. This FolksOntology of 259 keywords will be used in the next section to sample the interest-based social networks around these interests in Twitter.

[32] A tabular overview of a sub-sample of these keywords can be found in table 6 in the appendix.


Interests
Science
yahoo:astronomy
yahoo:Biology
yahoo:geography
yahoo:engineering
yahoo:agriculture
yahoo:physics
yahoo:mathematics
yahoo:computerscience
yahoo:chemistry
yahoo:ecology
yahoo:energy
yahoo:neuroscience
Social Science
yahoo:economics
yahoo:linguistics
yahoo:sociology
yahoo:psychology
yahoo:anthropology
yahoo:archaeology
yahoo:genealogy
Recreation
Automotive
yahoo:motorcycle
yahoo:cars
yahoo:automotive
Sports
wefollow+yahoo:baseball
yahoo:basketball
yahoo:golf
wefollow+yahoo:football
yahoo:soccer
yahoo:cycling
yahoo:hockey
yahoo:racing
yahoo:tennis
yahoo:rugby
yahoo:aviation
yahoo:wrestling
yahoo:swimming
yahoo:gambling
yahoo:running
yahoo:skateboarding
yahoo:skiing
yahoo:sailing
yahoo:surfing
wefollow: coach
wefollow+yahoo:poker
wefollow:biking
wefollow:sport
Games
wefollow + yahoo:videogames
yahoo:gaming
Outdoors
yahoo:boating
yahoo:hunting
yahoo:climbing
yahoo:fishing
yahoo:hiking
wefollow:outdoors
Travel
wefollow+yahoo:travel
wefollow:airlines
Hobbies
yahoo:collectibles
wefollow+yahoo:shopping
wefollow+yahoo:gardening
Toys
yahoo:toys
Pets & Animals
wefollow+yahoo: dogs
Health
Fitness
yahoo:exercise
wefollow:fitness
Pharmacy
yahoo:pharma
Medicine
yahoo:medicine
Diseases
yahoo:cancer
Mentalhealth
yahoo:mentalhealth
Dentalhealth
yahoo:dental
Nursing
yahoo:nursing
Weight
yahoo:weightloss
Nutrition
yahoo:nutrition
Reproductive Health
yahoo:pregnancy
yahoo:alternativehealth
Society and Culture
Cultures and Groups
yahoo:lesbian
wefollow:gay
yahoo:seniors
wefollow:geek
Environment and Nature
wefollow:sustainability
yahoo:greenliving
yahoo:nature
yahoo:environment
yahoo:climatechange
yahoo:conservation
wefollow:green
yahoo:weather
Families
yahoo:parenting
yahoo:children
yahoo:father
wefollow + yahoo:mother
Food and Drink
yahoo:cooking
yahoo:drinking
yahoo:baking
yahoo:recipes
yahoo:chef
yahoo:vegetarian
yahoo:dining
wefollow:restaurant
wefollow+yahoo: wine
wefollow+yahoo: beer
wefollow+yahoo:food
Issues and Causes
yahoo:philanthropy
yahoo:housing
yahoo:poverty
yahoo:peace
yahoo:humanrights
yahoo:diversity
wefollow: charity
Relationships
yahoo:marriage
yahoo:dating
yahoo:romance
yahoo:wedding
Religion and Society
yahoo:buddhism
yahoo:islam
yahoo:jewish
wefollow:astrology
wefollow: spirituality
wefollow+yahoo:religion
wefollow+yahoo:christian
Arts
Design Arts
wefollow+yahoo:fashion
yahoo:architecture
wefollow+yahoo:graphicdesign
Humanities
wefollow: author
yahoo:writing
wefollow+yahoo: poetry
yahoo:literature
yahoo:history
yahoo:philosophy
Performing Arts
wefollow+yahoo: theater
yahoo:dance
Crafts
wefollow+yahoo: crafts
yahoo: jewelry
wefollow: handmade
wefollow: etsy
Visual Arts
yahoo:painting
yahoo:photography
yahoo:sculpture
yahoo:illustration

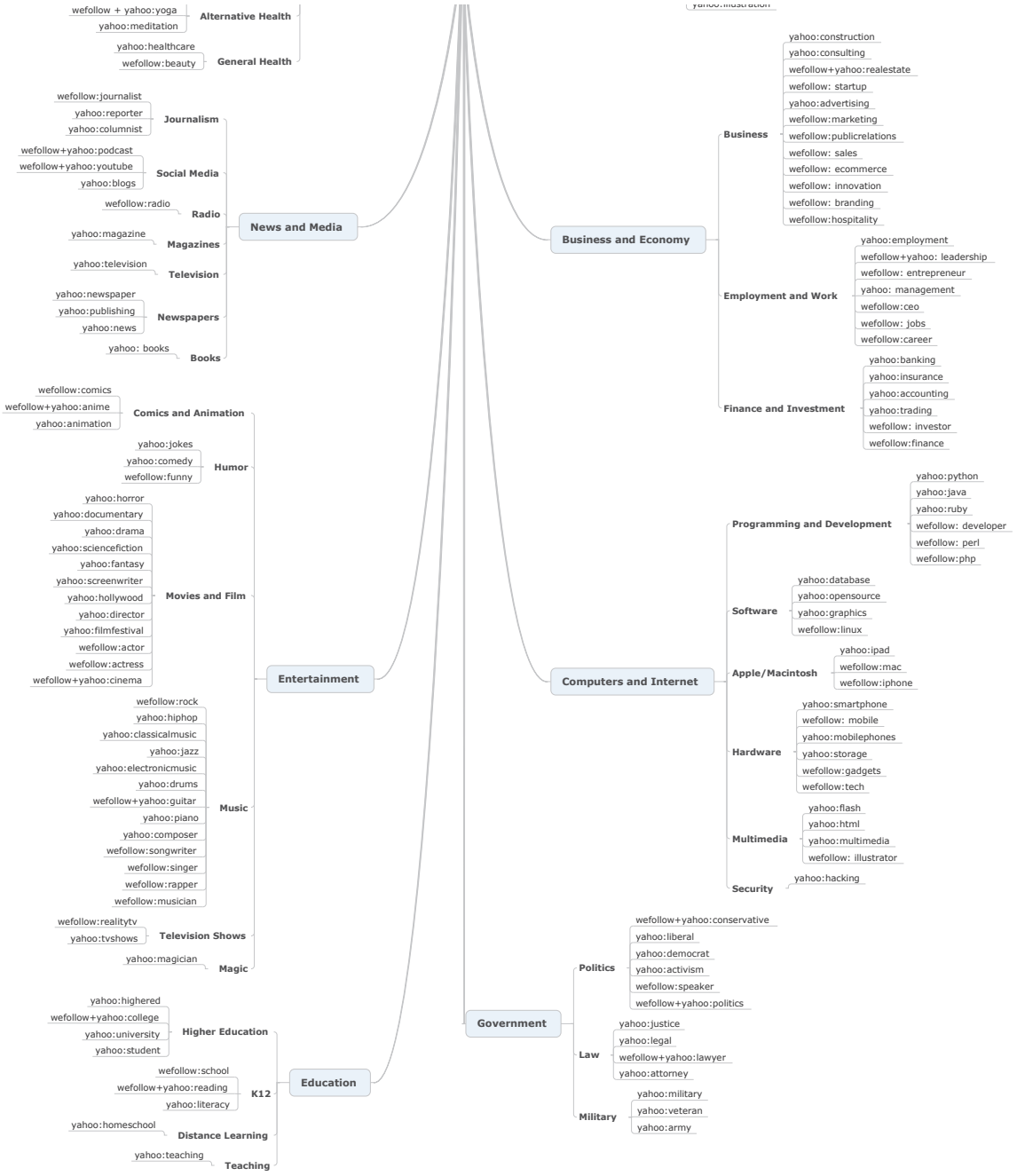


Figure 6.3: A hierarchical display of the resulting FolksOntology. The used keywords representing user interest at the same cognitive level of granularity can all be found the third level of the FolksOntology.

6.4 From users' interests to users' interests networks

After a FolksOntology of 259 potential users' interests was generated, the next sections will describe how networks of individuals for each of the users' interests have been extracted. These sections will describe how the seed based sampling approach was performed for each keyword of the FolksOntology and how it resulted interest networks of users that exhibit a strong cognitive membership in a particular interest. The rest of this chapter will discuss the stability of the performed approach and the choice of group size and finish with the description of the network extraction process for the different types networks and the further data processing.

Seed based network sampling

The idea to sample individuals that exhibit a certain attribute, trait or interest and then analyze their network is in literature generally referred to a seed based sampling approach (Wasserman and Faust, 1994). Seed based approaches all have in common, that they start with a small number of seed users and iteratively widen the sample. The most prominent form of these methods is snowball sampling (Frank, 1979), which starts with one random user and then recursively collects the friend and follower networks of the friends of this user. The problem with this sort of snowball sampling is that it is not sensitive to the user's interest, but simply depends on the explicitly visible edges. This means that when collecting users in regard to a certain interest, the main factor that influences the resulting network is the seed user himself. All other users are simply added, without regard to their actual interests in this topic.

There are yet variations of the seed based sampling approaches that are useful for this work, since they allow collecting networks for multiple groups of users and are sensitive to the users' interests: Weng et al. (2010) propose a seed based way of sampling networks from Twitter which starts with a number of seed users for a certain topic and then only collects friends of those users that are interested in this topic. They determine such users based on the content of their tweets. The users extracted in this way share a strong interest in the topics that they tweet on and so create a topic-specific interest network. Choudhury et al. (2010) propose a similar approach: Corresponding to each value v defined over an attribute, they construct a social graph such that is consists of users with the particular value of the attribute, while an edge exists between two users in this graph, if it has been present in the original graph of Twitter. This approach is similar to the approach of Weng et al. but it also allows to sample users based on different attributes. Finally, Wu et al. (2011) also propose a seed-based method based on Twitter lists. They highlight the importance of the Twitter list feature for the extraction of users that share certain similarities: "Since its launch on Twitter lists have been welcomed by the community as a way to group people and organize one's incoming stream of tweets. [...] List creation therefore effectively exploits the "wisdom of crowds" to the task of classifying users, both in terms of their importance to the community (number of lists on which they appear) and also how they are perceived (e.g., news organization vs. celebrity etc.)"[p.706]. Their method

is also based on selecting a number of predefined seeds for certain predefined categories. Once the seeds and the keywords for each of their categories have been selected, they then perform a modified snowball sample of the graph of users and lists. For each seed they crawl all lists on which that seed appeared. The resulting list of lists is then pruned only to contain the lists whose names matched at least one of the chosen keywords for that category. Additionally to the approach of Wu et al., section 4.2.3 has also highlighted the works of other researchers (Zhang Y, 2012; Greene et al., 2012; Kim et al., 2010) who have followed Wu et al.'s original list based sampling approach and also used lists to collect interest-based communities on Twitter.

Thus, conclusively, this work will also use a seed based list approach to collect groups of people that share the same interest. Interest-based seeds (which are Twitter users that strongly exhibit a certain interest) will allow us to collect multiple groups exhibiting a certain interest without having to collect the whole Twitter network. In comparison to pure snowball sampling, seed based approaches with a high number of seeds do not suffer the deficiency that the choice of seed is very crucial. In fact, regarding the choice of a sampling procedure and the boundaries of a social system Borgatti (2011) gives the following advice: "The naive concern is that we may select nodes "incorrectly" accidentally excluding nodes that should have been there and possibly including nodes that should not have been there. In reality, however, the choice of nodes should not generally be regarded as an empirical question. Rather this should be dictated by the research question and ones explanatory theory."[p.2]. This has explicitly been done in chapter 4, by defining groups by their cognitive similarity in regard to their interest. The next section will describe the details of the performed sampling procedure.

6.5 Description of the seed list-based network-sampling procedure

The performed sampling method in this work is very similar to the method performed by Wu et al. (2011) and Kim et al. (2010). The proposed method consists of two steps: The first step is to collect seed users around a certain attribute, in this case these are users that strongly exhibit an interest in one of the interest areas that have been collected in the last section. The second step is to extend this set of seed users into an interest-based community of users that share the same cognitive interest. This step is performed with the help of Twitter's list feature, which is used to perform a snowball like sampling procedure that is sensitive to the user's interests.

6.5.1 Choice of seed users

When defining a number of seed users that exhibit a certain interest it is important to decide from where to take such seeds. Regarding this task, there are a number of possibilities:

- A first possibility consists of collecting such users as seed users that have written a tweet or mentioned the keyword representing this interest in Twitter. This can be done by

monitoring the live stream of tweets (the Twitter gardenhose). The advantage of this attempt is that one obtains very active seed users on Twitter, yet the downside is that one might either collect a number of miscreants or spam users, who simply post tweets on recent and trending topics, in order to address a bigger audience or miss certain inactive users that are relevant for a keyword but have not mentioned it recently.

- A second possibility consists of finding one Twitter list that lists users that seem to be relevant for a given interest and use these users as the seed users. The upside of this approach is that it is relatively easy to perform, yet the downside is that we have rely on a subjective and arbitrary judgment of one person, to have provided a right set of seed members for a given keyword.
- The third possibility is to use existing third party Twitter directories that list people for a given interest. Such third party dictionaries had been introduced in chapter 6. The advantage of this approach is that those directories are maintained by high number of users and so their results can be expected to provide a relevant set of seed people for a given interest.

In this work the decision has been made to make use of the directory approach since it potentially allows harnessing the computational intelligence existing providers and the social judgment of users that have accepted these directories. Hence the resulting set of seed users can be expected provide relevant results. Additionally, in order to minimize the error of choosing an unrepresentative set of people, a high number of seeds (100) have been chosen. In comparison to snowball sampling that starts with one person, it is safe to assume that with a seed of 100 people, one can account for the error of choosing some seed users that might not fully represent the selected interest. These seed users will serve as the basis to collect more users that exhibit the same interest. Among the possible providers (see comparison in table 6.2) Wefollow was used to collect 100 seed users for each interest.

6.5.2 Extension of seed accounts into groups of users

Once a number of seed users have been collected, the next goal is to extend this seed into a larger group of people that share a certain interest. The whole approach of extending those users into an interest-based community is based on the use of Twitter lists (see figure 6.4). Since it is technically impossible to search for all Twitter lists on Twitter, we will use of the seed accounts to retrieve the appropriate lists: For each seed account all the lists that are listing those users will be collected and then only those lists that list users for the specific interest that we are interested in will be kept. The actual steps performed during the collection phase are displayed in figure 6.4 and can be described as following:

1. Starting with 100 Twitter seed users, the internal list feature from Twitter is used to extract all the lists that those users are listed on. This results in a collection of lists C_l, where each list l is specified by its name, description and the members that it lists.

2. In order to only maintain lists that strongly focus on the selected keyword, the collection of lists C_l is narrowed down to only those lists that contain the desired keyword in the list name or description.
3. This results in a smaller collection of list C_k (830 448 lists).
4. From this collection of lists C_k all the list members are extracted that are listed on each of those lists.
5. This results in a collection of members C_m (1.71 Mio. members).
6. For each member in C_m one counts how often this member was listed among the extracted lists C_k.
7. The more often a member was listed among the extracted lists C_k the higher this person is regarded as a member of the desired group for a given topic.

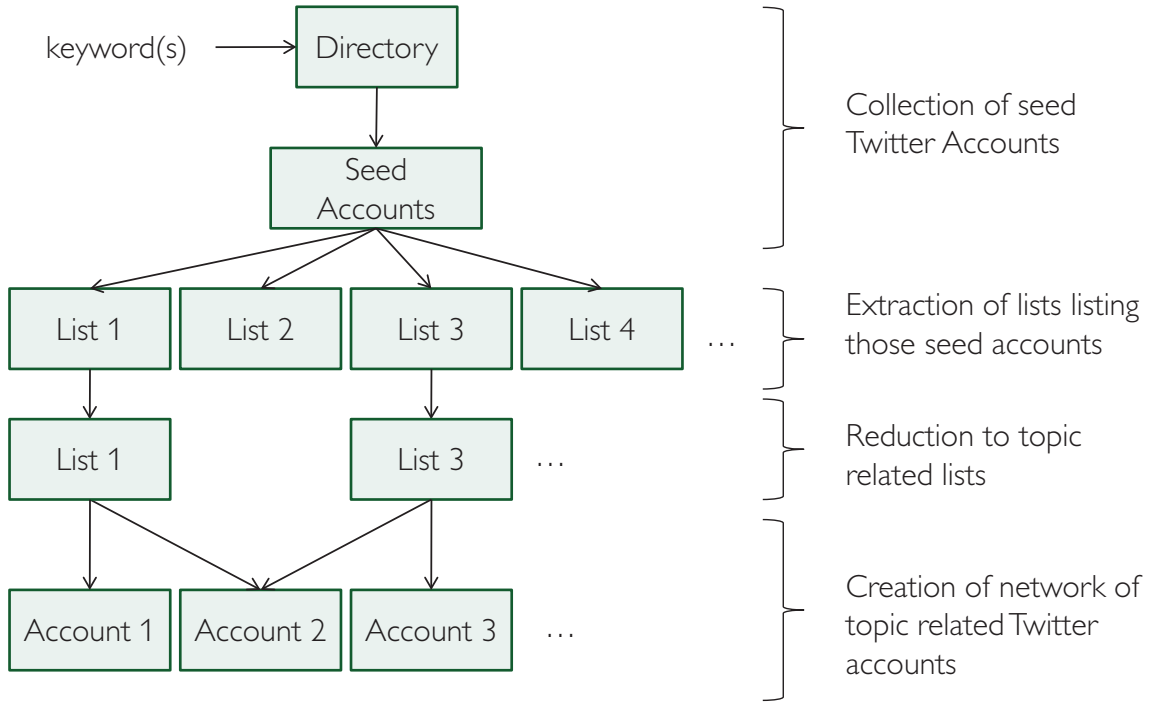


Figure 6.4: Overview of the group collection process in Twitter.

This procedure results in a transparent crowd sourced group nomination sampling process: The more often a person is listed under a certain keyword, the more one can infer that this person is actually relevant for the community existing around this interest. This nomination process also reveals the cognitive social capital dimension among those persons and allows to capture how strongly[33] each person cognitively aligns with the interest group (see section 4.2.3 in chapter 4). The next section will discuss the properties that result from this nomination process under the distribution of attention.

Distribution of attention

Regarding the distribution of listings that persons were able to collect for a given interest category, the first 100-200 group members of all groups - independent of the groups topic - were able to accumulate 80% of all mentions. This property is depicted in the overview of all nomination distributions in the appendix in figure A.8 and figure A.9. Although there were keyword specific differences, which mostly vary in the amount of nominations that the most nominated person in a group can obtain (from 200 up to 1800 nominations), the attention slope in all groups has the same distribution. This means that the overall attention distribution in regard to a certain interest is always the same. People seem to be highly interested in the most prominent persons exhibiting an interest, and then the attention falls off almost exponentially. It seems that lists which have been created by independent persons always seem to agree to a high percentage on the first 50-150 members of a given interest community. After that point, the competition for the highest cognitive overlap with the community is weakening, as if the boundaries of these community are becoming more and more fuzzy with each rank.

This distribution has been used to ensure that the resulting groups are “valid” in terms of that the nomination process has not been gambled by spammers or miscreants that have on purpose copied lists or recreated the same list over and over again under different accounts. Therefore a manual inspection of the distributions of the nomination curves (and the corresponding accounts) has been performed and lists that listed exactly the same persons were excluded. This inspection led to the exclusion of 21 groups (e.g., the group for the interest “weightloss” - which is a typical candidate for spammers trying to sell weight loss products). In such cases the attention curves did not exhibit the typical log-normal distribution, but instead showed equal numbers of votes for all members (which resulted from bots automatically creating lists for those terms and populating them with the same accounts). Additionally, keywords for which not enough lists could be collected (e.g., “sculpture”, “seniors”, “boating”) were also excluded from the data set. Here, such groups were dropped for which the nomination process would be based on less than 10 lists as such group memberships nominations would rather create arbitrary results. This led to the exclusion of 38 groups. This constraint also explains why only

[33] For example if for a given keyword such as “java” (a programming language) a person which is listed over 1000 times and while another person is listed only 3 times, one can infer that the first person has a higher individual cognitive social capital dimension for this group than the later.

popular keywords (that reported at least 200 members on Wefollow) from the Yahoo ontology were included in the final list of interests in section 6.3.1. This constraint ensured that not too many groups would need to be dropped at this point in time.

In order to control for the broadness of the interest category, keywords that describe interests at same level of granularity have been used (see section 6.3.1). To additionally monitor this aspect a variable called "member count" has been introduced. The "member count" is simply the number of unique members that have be found on the lists for a certain category. This variable will be used as a control variable in chapter 7 to control if groups for whom more members have been suggested had different levels of social capital than other groups. An overview of the number of lists that were used for the final interest based communities, including an overview of the minimum, maximum and average number of unique users in the collected lists per category can be found in table 6 in the appendix.

6.5.3 Stability of the performed sampling procedure

One important consideration can be made in regard to the stability of the performed group sampling procedure: Would have other (better) initial seed members have led to other outcomes in terms of other persons being able to collect the highest amount of votes? In order to test for this concern the procedure was repeated with the 100 most listed group members for each keyword as the seed members. This allowed to measure potential improvements that might have occurred if we had started with an even more relevant set of initial seed members for a given keyword. In order to measure a potential improvement, the 100 most nominated members were used as seed members and the sampling procedure was then repeated with those members. This experiment has been performed with 30 random[34] keywords. The overlap of the final sets is expressed as a simple percentage and the correlation between the final rankings as a Kendall's Tau. The results show that not only did 84% of the 100 initial members end up in the top 100 most ranked members, but with a Kendall's tau of .76 on average, there is a high correlation between the initial ranking of the users and the follow up ranking. This result shows that performing the sampling procedure only once is enough to guarantee a high confidence in the results.

6.5.4 Choice of group size

The second main consideration, when studying groups of users that exhibit an interest, lies in determining a meaningful group size. Generally such groups are not limited in size, or have a predefined boundary, as secluded villages in remote areas might have. Yet there are

[34] musician, ruby, java, python, investor, tennis, astrology, automotive, golf, finance, accounting, surfing, basketball, lawyer, healthcare, dating, astronomy, advertising, news, hiking, airlines, religion, sustainability, ceo, hospitality, realestate, radio, sailing, sport, football

theoretical and empirical results that suggest that such groups might have a theoretical limit, which is limited by people's cognitive abilities to deal with social relations. The references from theoretical considerations and the results from empirical research will be presented now.

The theoretical considerations about the cognitive limit in regards to a group size, is also widely known as Dunbar's number (Christopher, 2004; Dunbar, 1992). Dunbar's number is suggested to be a theoretical cognitive limit to the number of people with whom one can maintain stable social relationships. These are relationships in which an individual knows who each person is, and how each person relates to every other person. There is no precise value for Dunbar's number but it has been proposed to lie between 100 and 230 persons, with a commonly used value of 100-150 (Hernando et al., 2009). A similar theoretical consideration is the idea that attention and time are scarce resources. It has also led others (Davenport, 2002) to introduce the concept of an attention economy, coming to similar conclusions that there must be a personal limit on how many ties with a group can be maintained, independent of the medium that those people use to communicate.

These generic theoretical considerations are also confirmed for Twitter by empirical research based on the work of Goncalves et al. (2011) and Grabowicz et al. (2012). Goncalves et al. found validation for the theoretical cognitive limit on the number of stable social relationships in Twitter by analyzing 3 Million users and over 380 million tweets, and 25 million conversations. In the interactional network between users, which represents genuine social interaction, they found that users indeed entertain a maximum of 100-200 stable relationships in support for Dunbar's prediction. They found that the "economy of attention" is limited even in Twitter by cognitive and biological constraints as predicted by Dunbar's theory. The observations of Grabowicz et al. (2012) also affirm these cognitive limits, by finding out that the most emergent group size of relations for groups is up to 150 users. Grabowicz et al. analyzed a sample of the follower network using own clustering algorithms and identified different sets of groups with different sizes around 100-200 users. Their dataset was based on a sample of the Twitter network containing 2.4 Mio. users connected with 4.8 billion follower relations. After applying the clustering algorithms, they found that the fraction of internal links as a function of the group size in number of users is highest around 100 users.

In conclusion, in this work the decision has been made to select the 100 most nominated persons for each interest group as representatives of this interest community. This number is not only in accordance with the theoretical considerations according to Dunbar's number, but also in line with empirical research results in Twitter. Additionally, the overview of the final communities (see table 6 in the appendix) shows that the average number of listed persons per list is 112, which indeed indicates that 100 persons per community is an appropriate group size. Finally, two more procedures have been performed that ensured that those 100 members share a high cognitive overlap in terms of their expressed interest.

Group merging procedure

The next step was to verify if the members of the established interest groups actually represent different sets of persons. This means that it is possible that Twitter users have listed the same persons for e.g., the keyword "sustainability" and "ecology" or e.g., "restaurant" and "dining". In this case, both interest communities would be represented through different keywords but would be perceived[35] under the same concept. Although in the keyword generation process (see section 6.2.2) a lexical and semantic grouping of keywords concepts was performed, it is unclear if the Twitter users actually perceive those concepts as different representations of users' interests.

In order to exclude such cases in which Twitter users perceive keywords as too similar, a threshold has been introduced, and whenever the final groups reported sets that had an overlap of over 80% of members those groups were merged into one group. The membership of a person in this new group was determined by adding up the nominations that members of those groups were able to collect in all the keywords that were merged (e.g., if a member A was listed 5 times for category X and 5 times for category Y, and if category X and Y were merged into category XY, this member would collect 10 votes for this new category). This procedure resulted in the merge of 63 keywords into 29 combined keywords[36]. This resulted in 166 final groups, which are listed in table 6 in the appendix.

Re-ranking procedure

The final step of determining valid representations of interest networks was to allow persons to switch interest groups if the nomination procedure (amount of nominations collected for a keyword) indicates that members are even better suited to represent another interests than their initial one. The logic behind this operation is that whenever a person collected more nominations in a different interest group, it is safe to move this person to a more fitting group (e.g., a person might be listed 115th place for "yoga" but 62nd place for "meditation"). A move of a person is only performed if this person is able to obtain a higher rank in another group than its initial group (e.g., if member A was listed 5 times in category X and has the 3rd rank in the group X, but for a group Y this member was listed also 5 times but has the 1st in this group, then this person is moved from category X to category Y.). The persons considered for

[35] Generally all keywords that share the same second level parent in the Yahoo ontology, are candidates to be perceived as similar concepts. In such a case, this will result in groups that have very similar member sets.

[36] The following keyword concepts have been merged (different keywords are separated by underline): "airlines_aviation", "army_military_veteran", "astronomy_physics", "beauty_fashion_shopping", "career_employment", "charity_philanthropy", "comedy_funny", "etsy_handmade", "exercise_fitness", "finance_economics", "healthcare_medicine", "homeschool_school", "humanrights_activism_justice", "lawyer_legal", "linux_opensource", "mac_iphone", "mobile_smartphone", "musician_singer", "outdoors_hiking", "politics_news", "psychology_mentalhealth", "publishing_literature", "recipes_cooking", "restaurant_dining", "sport_football", "sustainability_ecology", "theatre_theater", "tvshows_drama_actor_hollywood"

this move operation were the first 200 most nominated persons for each group. Out of the 30 385 persons considered in 166 groups such a move operation was performed 1121 times.

6.6 Final representation of networks

After performing the three cleanup-operations (list-pruning, group merging and re-ranking) it can be assumed that the final 166 groups, consisting of the 100 most nominated persons, represent highly prominent users for a given interest and that those persons also cognitively consider themselves as being part of that interest community. In terms of social capital, we can say that those persons hold strong implicit ties with each other or possess a high cognitive social capital dimension. Additionally, chapter 7 will show, that indeed the exhibited explicit ties of those interest-based social networks show properties that classify those groups as social networks.

In order to be able to analyze this corpus the final step was to extract the actual network relations between those people in order to obtain the interest-based networks. For this task, an open-source tool[37] was developed capable of collecting groups based on the approach described above and being able to extract the network data. According to the theoretical considerations on different types of network ties between actors (see chapter 4), the tool was used to create all three types of networks (friend and follower-, interaction- and retweet-networks) for the given groups of users. The construction of those networks was based on the entities related to the users e.g., lists, tweets, mentions and retweets. A short overview of the overall process from the seed based network sampling to the aggregation and extraction of network data is shown in the overview in figure 6.5. The way of constructing the FF, AT and RT network will be described accordingly.

The processed data traces that were used in order to construct the resulting interest networks are the following: As part of the seed based sampling procedure a total of 830 448 lists have been collected. From these lists, over 1.71 Mio. unique Twitter users have been reviewed in order to compile nominations for each keyword category (see table 6 in the appendix). For the final set of 166 groups, 100 persons each have been collected. For each person up to 3 200 tweets have been collected resulting in a total of 46 231 808 collected tweets. For each tweet up to 100 retweets have been collected, resulting in a total 613 404 616 retweets. This data is the basis for the compilation of the friend and follower-, interaction- and retweet-networks.

[37] For more information refer to `https://github.com/plotti/twitterlyzer`. The tool is based on a Ruby on Rails Framework, that allows for a web based user interface and a scalable database architecture. The majority of data is stored in a SQL database, while edges are persisted in a key value store due to performance reasons. The collection tasks are distributed among a pool of workers, able to run on multiple machines. The tool has been developed using a behavioral driven approach (`http://en.wikipedia.org/wiki/Behavior_Driven_Development`), where a number of tests constantly verifies the validity of the processes performed by the tool.

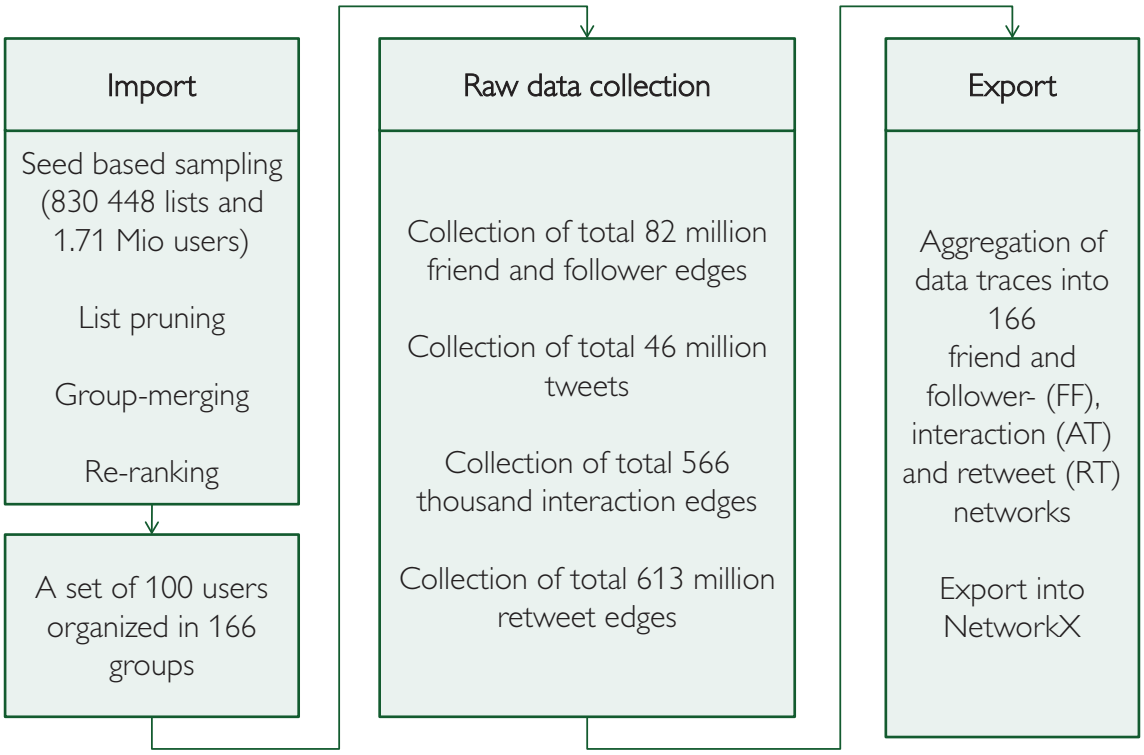


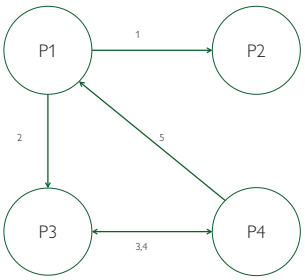
Figure 6.5: Schematic overview of the data collection process

Construction of the FF network

The attention network, or simply the FF network, among the set of 100 users (of the 166 groups) representing a shared interest is constructed by:

- Collecting the friend ties (outgoing ties) of each member (e.g., *P1*) of the group
- For each friend tie *T* the following task is performed:
  - if that tie T is pointing to one of the other members (e.g., *P2*) of the group, then a triple of the form P1,P2,1 is constructed
  - the tie strength in this triple is always set one, since friend ties are binary and non-valued

The group of all such triplets is an adjacency list[38] based representation of the friend and follower network. The total number of analyzed friends and follower relations was 82 808 978 edges. The resulting non-partitioned network (including ties inside and between groups) had a size of 16 370 nodes and 2 281 738 edges. An example of a graph and its corresponding adjacency list is shown in figure 6.6.



Tie ID	From	To	Tie Strength
1	P1	P2	1
2	P1	P3	1
3	P3	P4	1
4	P4	P3	1
5	P4	P1	1

Figure 6.6: Notation of a simple FF network

Construction of the AT network

The interactional network, or simply the AT network, among the set of 100 users (of the 166 groups) representing a certain interest group is constructed by:

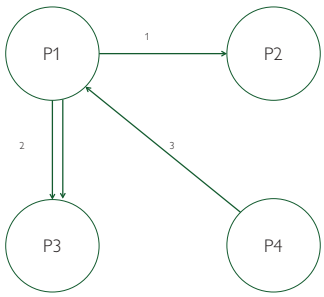
- Collecting all up to 3200 tweets of each member of the group
- For each tweet *TW* of each member (e.g., *P1*) the following task is performed:

[38] http://en.wikipedia.org/wiki/Adjacency_list

  - if in the tweet *TW*, the member *P1* mentions another member *P2* of the group,
  - and if this mention is not a retweet,
  - then a triple of the form P1,P2,1 is created.
- By aggregating multiple mentions (e.g., P1,P2,1 and P1,P2,1) between the same users, the resulting interactional valued tie between the users is represented in the triplet P1,P2,2.

The group of all such triplets is an adjacency list based representation of the interactional network among the users. Interactions in the form of:"@drewconway Do you know any good resources on network panel analysis?" are considered to be interactions. However tweets in the form of "RT @drewconway A great collection of network analysis data http://bit.ly/ItI1uo" are considered to be retweets. Retweets were additionally distinguished from interactional actions by using the Twitter API, which reveals meta-information on if a tweet has been generated by using the built in retweet function or not.

In order to generate the interactional networks, the 46 231 808 collected tweets have been searched for the occurrence of the 16 370 Twitter users constituting the 166 groups[39]. The resulting non-partitioned interactional network had a total size of 16 370 nodes and 566 884 edges. An example of an interaction graph and its corresponding adjacency list is shown in figure 6.7.



Tie ID	From	To	Tie Strength
1	P1	P2	1
2	P3	P4	2
3	P4	P3	1

Figure 6.7: Notation of a simple AT network

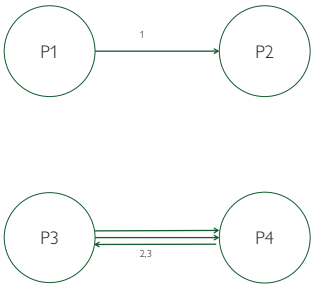
Construction of the RT network

The information diffusion network, or simply the RT network, among the set of 100 users (of the 166 groups) representing a certain interest group is constructed by:

[39]This task creates a significant run time problem as the search for so many users in so many tweets would take approximately 70-80 days on a regular desktop computer. In order to reduce this lookup time, a full-fledged back end search engine (`http://de.wikipedia.org/wiki/Lucene`) was used that computed a full index over the text corpus. This lead to a reduction of computational time of a factor of 93x.

- Collecting all up to 3200 tweets of each member of the group
- Collecting all up to 100 retweets for each tweet
  - For each retweet, the following additional information is saved: retweeters user name, follower count, friends count, date of the retweet.
- For each tweet *TW* of each member *P1* the following task is performed:
  - if a retweet is made by members *P2* of the group,
  - and if the retweet is not an interaction,
  - then a triple of the form P1,P2,1 is created
- If retweets happen multiple times over a number of tweets between the same members the tie strength is added up by aggregating multiple retweets (e.g., P1,P2,1 and P1,P2,1) between the same users. The resulting information flow between the users is represented in the triplet P1,P2,2.

The group of all such triplets is an adjacency list based representation of the information diffusion network among the users. For the generation of the retweet networks the total of 613 404 616 retweets have been processed and whenever a retweet took place a retweet edge was constructed between the members of the final set. The resulting non-partitioned network had a size of 16 370 nodes and 205 787 edges. An example of a graph and an example of the corresponding adjacency list is shown in figure 6.8.



Tie ID	From	To	Tie Strength
1	P1	P2	1
2	P3	P4	2
3	P4	P3	1

Figure 6.8: Notation of a simple RT network

Attribute data for each actor

The sampling steps that have been described in the last sections led to the collection of additional attribute information for every actor. This data is stored in a table form. For each actor that has been nominated as part of the sampling process into the final set of 100 actors per

group, the Twitter ID and the group name have been stored. Additionally, the actor's interest, as expressed by the number of listings that the member was able to receive from the lists, was stored in form of a ranking in interest amongst members of the same group. The more votes the actor was able to collect, the higher his position on the rank of actors that represent this interest group. This variable (Ranking in interest) captures the actors cognitive bonding social capital. Finally, the number of the actors' interests was stored, which indicates how often the actor also appeared on other lists for different topics. This variable (Number of interests) captures the actors cognitive bridging social capital. An example is shown in figure 6.9 for four persons that have been nominated for three interest groups (running, swimming and cycling). Actor P1 has a higher ranking in interest than actor P2, because he was able to collect more listings. Actor P4 has the highest number of interests because he has been nominated in lists for three different topics.

Export

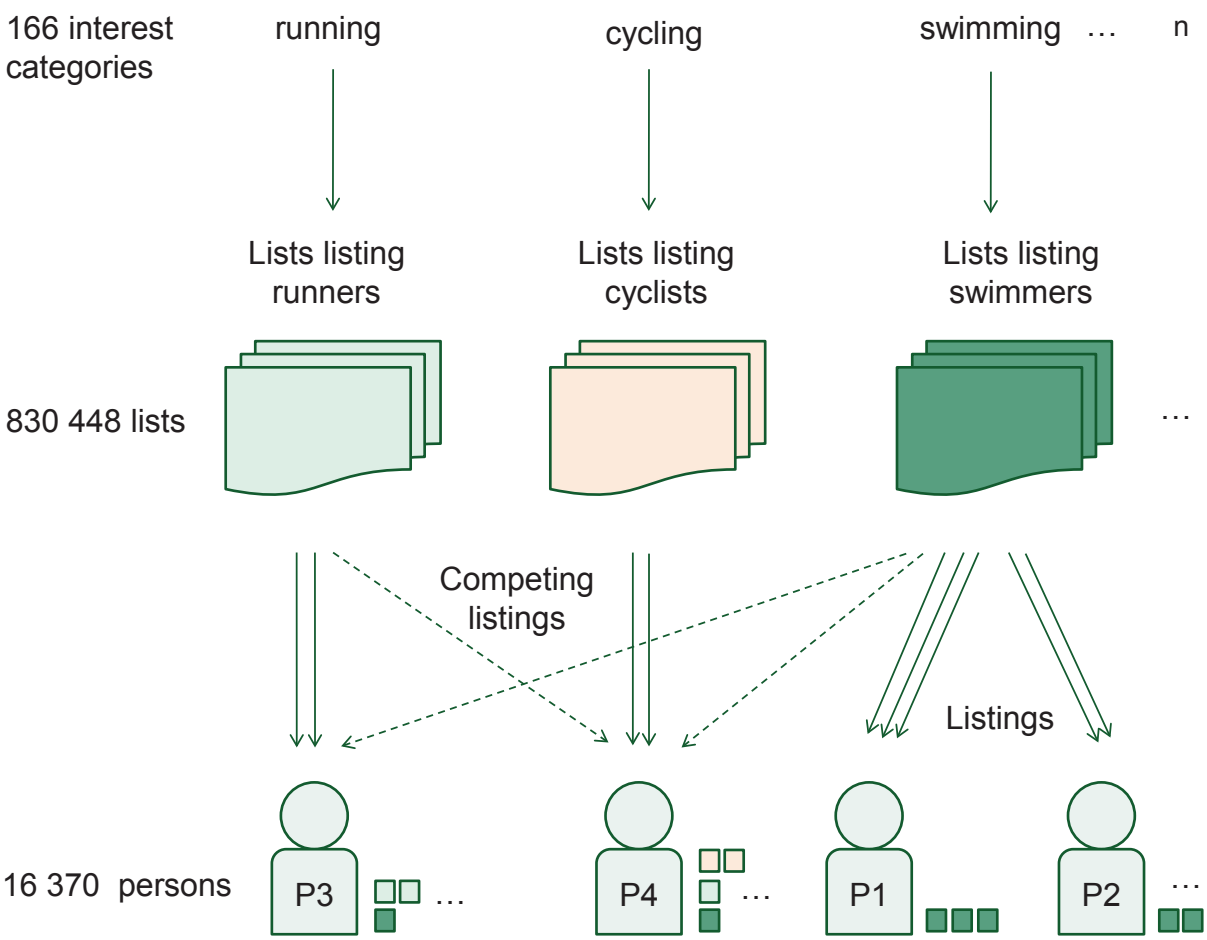
The last step of the data processing was to export the resulting network data for each of the 166 groups for further processing by network analysis programs. All further analysis is performed by the standard network analysis software NetworkX[40]. Finally in order to highlight certain aspects of the networks, network visualizations have been used. For this purpose the standard open source visualization software Gephi (Bastian et al., 2009) was employed.

6.7 Conclusion

In the first part of this chapter, a FolksOntology of 259 potential users' interests was generated. It provided a representative sample of users' interests in Twitter at the same level of granularity. The second part of this chapter showed how the implementation of the Twitter-list seed based sampling method was performed in order to generate interest-based networks for these interests: The process started with the collection of 100 initial seed users which were sampled by using third party providers. Using the list feature of Twitter for each of the seed users, all lists have been retrieved that listed more users for the appropriate keyword. These lists have been used to extend the initial seed of users for each interest, into larger groups of users that were considered for each category. This resulted in a collection of 830 448 lists, and a review of 1.71 Mio. potential candidates that have been reviewed for each interest category. For each of the keywords the 100 most listed users that were able to generate the most nominations for the keyword were collected.

The remaining challenges where potentially multiple keywords might represent the same semantic concept and the consideration that users might be able to acquire memberships in

[40] NetworkX can be found under `http://networkx.lanl.gov`. It is a network toolkit similar to the popular software like UCINET (Borgatti et al., 2002) or Pajek (Batagelj and Mrvar, 2008).



Twitter ID	Group name	Ranking in interest	Number of interests
P1	swimming	1	1
P2	swimming	2	1
P3	running	1	2
P4	cycling	1	3
...	...	...	...

Figure 6.9: Overview over the attribute data for each sampled Twitter user on the example of three interests and four users.

multiple groups were approached by this work. This chapter also highlighted the stability of the performed procedure, and presented theoretical and empirical evidence supporting the choice of 100 users as the group size for each category. The final set of 166 interest-based networks (see table 6 in the appendix) consists of 100 persons and their connections. For each of these persons up to 3 200 tweets have been collected resulting in a total of 46 231 808 collected tweets. For each tweet up to 100 retweets have been collected, resulting in a total 613 404 616 retweets. This data has been used to aggregate the underlying data traces into the friend and follower-, interaction- and retweet-networks for each category, which are the basis for the empirical analysis of the research questions, that will now follow in chapter 7.

7 Results & discussion

Selection of data analysis method

This chapter will empirically answer the hypotheses postulated in chapter 5, by applying a statistical method of **multiple regression analysis**. This method is applicable since the hypotheses always imply a **directional influence** of social capital on information diffusion. The social capital metrics resulting from the attention (FF) and interaction network (AT) are used as indicators for structural, relational and cognitive social capital at the group level and at the individual level, as described in chapter 5. These independent variables (IV) are regressed on the metrics from the retweet network (RT), which serve as the dependent variables (DV). At the individual level, the Twitter accounts are the unit of analysis, whereas at the group level, the entire group is the unit of analysis. Since multiple IVs are used to explain the DV, this empirical method is composed of different multi-regression analyses at the group and individual level. This chapter will be divided into five sections. Section one to four will answer hypotheses 1-4 and section five will answer hypotheses 5-8. Each of the following five sections of this chapter will be structured accordingly: First, the descriptive statistics of the distribution of variables, the regression method applied and the underlying assumptions of the analysis will be presented. Then, the model of the corresponding type of social capital will be constructed and interpreted. Each section will end on a discussion of how the results led to the acceptance or rejection of the underlying hypotheses. Table 7 in the appendix contains a short overview of all the hypotheses, indicators and outcome variables.

7.1 Influence of group bonding social capital on information diffusion

Hypothesis 1a: The more structural bonding group social capital the group realizes, the more retweets are exchanged in the group network.

Hypothesis 1b: The more relational bonding group social capital the group realizes, the more retweets are exchanged in the group network.

7.1.1 Plausibility assumptions

Before beginning to empirically verify the postulated hypotheses, a first plausibility assumption has to be met: This basic assumption is that each interest-based community analyzed

corresponds to a social network with similar properties to other social networks. The fulfillment of this assumption is necessary since groups of people that are not creating social ties cannot be analyzed using theories that describe information diffusion in social networks. In order to show that the collected groups express properties that are also expressed by other social networks, the characteristic network measures of each group were computed and the minimum, maximum, mean and standard deviation values were calculated. These values can be seen in table 7.1. The mean values were compared to similar communication networks that are a) social b) directed and c) weighted. They were also compared to research results in Twitter networks. Indeed, if the means and ranges of the network metrics obtained for the analyzed groups are the same as the ranges in the reference networks, we can safely assume that the social structure in these interest networks creates an authentic underlying social network. The following weighted directed communications networks, which are widely known and used in the literature, were used to provide comparable metrics:

- Opsahl Network (Opsahl and Panzarasa, 2009): The Facebook-like social network originated from an online community of students at the University of California, Irvine. The dataset included 1 899 users that sent or received at least one message. A total number of 59 835 online messages were sent over 20 296 directed ties that connected these users.
- Freeman Network (Freeman, 1979): The second dataset is Freeman's EIES networks, which was also used in the works of Wasserman and Faust (Wasserman and Faust, 1994). This dataset was collected in 1978 and contains a network of researchers working on social network analysis. It represents the number of messages sent between 32 of the researchers using an electronic communication tool.
- Cross-Parker Network (Cross and Parker, 2004): This dataset contains an intra-organizational communication network and was collected by Cross and Parker in 2004. The network was collected at a consulting company with 46 employees. The participants indicated on a scale from 0 to 5 the frequency at which they exchanged information with each other or requested advice.

Additionally to these three reference networks, the network properties of a number of Twitter networks were compared to the network metrics of the 166 interest networks that were sampled in the last chapter. The Twitter networks collected were much larger in size than the reference networks and the 166 interest networks of this study. The results of this comparison are displayed in table 7.1 and lead to the following conclusions:

Average path length: The average path length in the reference network lies between 1.49 and 2.20 hops. In Twitter, Kwak et al. (2010) reported an average path length of 4.12 in their dataset and Galuba et al. (2010) reported a shortest path of 3.61. In the collected dataset, the average path length is 1.71 in the interaction (AT) network and 1.94 in the attention (FF) network respectively. This means that on one hand the average path length computed is highly comparable to the average path length of the reference networks. On the other hand, it the

Measure	Min	Max	**Mean**	Std. Deviation	Opsahl Network	Freeman Network	Cross Parker Network
Nodes	101	101	**101**	0	1899	32	46
FF_edges	477	7064	**3242.11**	1425.51	20296	460	879
AT_edges	94	3729	**1324.94**	808.96	-	-	-
RT_edges	48	1714	**606.97**	343.89	-	-	-
FF_bin_density	.05	.70	**.32**	.14	.01	.46	.42
AT_density	.02	4.78	**1.07**	.89	.02	15.64	.95
RT_density	.01	.74	**.25**	.16	-	-	-
FF_avg_degree	4.72	69.94	**32.10**	14.07	21.37	28.75	38.21
AT_avg_degree	.93	36.92	**13.12**	7.98	-	-	-
RT_avg_degree	.47	16.97	**6.01**	3.39	-	-	-
FF_bin_avg_path_length	1.29	2.93	**1.71**	.28	2.20	1.56	1.49
AT_bin_avg_path_length	.23	3.00	**1.94**	.38	-	-	-
FF_bin_clustering	.39	.86	**.67**	.09	.18	.80	.80
AT_bin_clustering	.16	.72	**.46**	.11	-	-	-
FF_reciprocity	.22	.80	**.49**	.13	.47	.68	.60
AT_reciprocity	.09	.63	**.33**	.10	-	-	-

Table 7.1: Comparison of the network measures of the 166 groups with reference networks from literature.

average path length is smaller than the one reported in the researched Twitter networks. This can be explained due to the smaller size of the interest-based networks, which contained only 100 members each, whereas the Twitter networks contained thousands of members.
Reciprocity: The reference networks report reciprocity values that vary between 47% and 60%. In Twitter, Kwak et al. reported a reciprocity of 22.1%; Cha et al. (2010) reported 10% in their dataset and Java et al. (2009) found a reciprocity of 58%. In the collected dataset, we found a mean reciprocity of 33% in the attention network, and 49% in the interaction network — values which fall into the reported ranges.
Clustering coefficient: The clustering coefficient in the reference networks ranges from .18 in the Opsahl network to .80 in the Freeman and Cross-Parker networks. In Twitter, Java et al. found it to be 0.10 (in the FF network), and Choudhury (2010) reported it to be 0.022 in the FF network and 0.0604 in the AT network. In the collected dataset, the clustering coefficient ranged from .46 in the attention network to .67 in the interaction network. On one hand, this clustering coefficient ranges within the values found in the reference networks; on the other hand, it is rather high in comparison to the values found in the Twitter networks. Such a difference can be explained by the larger size of the Twitter networks, and because members cluster less in these networks: indeed, they were not sampled according to homophile interests.
Average degree: The reference networks presented an average degree of 21.37 in the Opsahl network, 28.75 in the Freeman network, and finally 38.21 in the Cross-Parker network. In Twitter, Java et al. reported an average degree of 18.86. Choudhury reported 3.59 in the attention network and 8.82 in the interaction network, while Kwak et al. reported a much higher average degree of 70.51 edges/node. In the collected dataset, the mean average degrees were between 6.01 and 32.10, which lie well in between the reported values. The highest mean average degree was found in the FF network (32.10), followed by the AT network (13.12) and the RT network (6.01).

Regarding the comparison of the previously discussed network metrics, it can safely be assumed that the generated interest networks exhibit the same network structure as that of other social networks. For this reason, the generated interest networks represent well social networks in general. The dataset contains no missing samples, since all of the group network measures were computed for all of the groups. The measures that were computed on a dichotomized version of the network were marked with a "_bin" in the variable name. Measures that were transformed were marked with the corresponding transformations in front of the variable (e.g., LN_AT_density). A descriptive overview of the variables can be found in table 7.1. These measures have already been discussed under the plausibility assumption section and indicate that the underlying networks are highly clustered with a high average degree. These measures can be interpreted as a result of homophily, where community members have a high tendency to create ties with members that share the same interest. This is highlighted in the high assortativity (Newman, 2002a) of the FF network (0.22), AT network (0.49) and RT network (0.39).

Distributions of variables

The IVs and DVs in the dataset of 166 groups possess different distributions. The K-S and Shapiro-Wilk tests of normality (see table 2 in the appendix) indicate that some of the IVs are not normally distributed. The IVs that are normally distributed according to the K-S statistic are: FF_bin_density, AT_bin_clustering, FF_reciprocity, AT_reciprocity. The variables that have a non-normal distribution are: Member_count, AT_bin_density, FF_bin_avg_path_length, AT_bin_avg_path_length and the FF_bin_clustering. The DV RT_density is also not normally distributed. Non-normally distributed variables were transformed using the LN transformation, which is indicated in the variable name. These transformations have commonly been used in other studies (Wasko and Faraj, 2005)[1].

Regression Method: Stepwise Enter

For all subsequent statistical analyses, the method of multiple regression with a stepwise approach was used. This method allows us to determine to which extent indicators of the structural and relational dimensions of group bonding social capital actually predict information diffusion. The diffusion itself was measured by the retweet network density in the group. The regression is performed by placing theoretically-determined indicators for each dimension into the model, and then calculating the contribution of each one by looking at the significance value of the t-test for each predictor. This significance value is compared to an addition criterion (which can either be an absolute value of the test statistic, or a probability value for that test statistic). If a predictor meets the addition criterion (i.e., if it is making a statistically significant contribution to the accuracy of the predicted outcome variable), it is added to the model and the model is re-estimated for the remaining predictors. The contribution of the remaining predictors is then reassessed. This step was performed for models 1.1 through 1.5., model 1.6 represents the best model of group bonding social capital, since it includes the most significant predictors for each dimension of social capital.

Regression assumptions

The Durbin-Watson statistic (Durbin and Watson, 1957) for models 1.1 to 1.5 is 2.039; for model 1.6 it is 2.045. These values are close to 2, thus revealing that the assumption of independent errors is met. Addressing the assumption that multicollinearity is absent, the VIF tolerance statistics (Bowerman, B.L. , O'Connell R, 1990) show that for models 1.1 through 1.5, the VIF is not greater than 10, while in model 1.6 the VIF is less than 4. The plot of *ZRESID against *ZPRED (see figure A.1 in the appendix) displays a random array of dots

[1] Wasko et al. reported that "Centrality was calculated using the UCINET 6 program (Borgatti et al. 1999). There were 3 000 messages posted by 604 participants in the network during this time frame, indicating a vibrant, active network. To reduce skewness, the variable was transformed using a log transformation."

evenly dispersed around zero. The histogram of the regression standardized residuals and the P-P plot (see figure A.1 in the appendix) show that the points mostly lie on a straight line. The probability plot is a little skewed to the left, but does not raise serious concerns.

7.1.2 Models of group bonding social capital

Interpretation of the models coefficients

In this regression, six different models were introduced, each building on top of the baseline model 1.1. Each model extended the number of group bonding social capital indicators successively. Models 1.1 to 1.5 explore the structural and relational dimensions of social capital. Model 1.6 contains only the most significant predictors. The R^2 of model 1.1 indicates that the IV Member_count is not significant at explaining the DV, whereas the two following models make significant contributions to explaining the outcome variable. Regarding the R^2, the IVs FF_bin_avg_path_length and AT_bin_avg_path_length in models 1.4 add only limited information to model 1.3, resulting in insignificant R^2 change statistics. However, model 1.5, which contains the metrics of the relational dimension, results in a significant R^2 change statistic.

In models 1.1 through 1.5, the standardized Beta values of the coefficients reveal that the control variable Member_count - which corresponds to the number of people considered per category, and is used to determine the general breadth of the interest category - is unable to predict the information diffusion within the group. This means that interest groups which have been labeled under a more generic term (e.g., liberals) exchange the same amount of information as more specific interest groups (e.g., php programming).

Model 1.2 reveals that the IV LN_AT_density (with a Beta of .60) is the strongest indicator of information diffusion inside the group. This means that groups with a lively interaction between members also tend to have better information diffusion within the group. The models also reveal that when both the LN_AT_density and the FF_bin_density are taken into account, the FF_bin_density metric is often rendered insignificant ($p \geq .05$). This observation can also be found in the other IVs from both FF/AT networks in models 1.2 through 1.5: Whenever one of the network metrics is significant at predicting the DV (e.g., in model 1.3 AT clustering), it is usually the metric derived from the AT network, not the FF network. In this case, it seems that the data derived from the computationally[2] more expensive AT network is better at explaining the information diffusion than the metrics derived from the simple friend and follower network. This observation is in line with the research found in the literature review regarding the influence of @mentions on information diffusion at the individual level. It is also in line with the network flow model, which indicates that interpersonal interactions are the last of three steps[3] that lead to information diffusion.

[2] In order to compute the AT network all interactions between Twitter users have to be analyzed.

[3] Similarities, social relations and interactions

	Model 1.1		Model 1.2		Model 1.3		Model 1.4		Model 1.5		Model 1.6	
	Beta	VIF	Beta	VIF	Beta	VIF	Beta	VIF	Beta	VIF	Beta	VIF
Member_-count	-.08	1.00	-.09	1.00	-.07	1.02	-.06	1.02	-.09	1.05		
LN_AT_density			.60 ***	1.49	.38 ***	3.96	.37 **	4.01	.40 ***	4.61	.33 ***	3.29
FF_bin_density			.05	1.49	-.19	5.31	-.33 *	6.69	-.07	9.47		
FF_bin_clustering					.23	5.83	.16	6.24	.05	6.69		
AT_bin_clustering					.30 *	4.56	.33 **	4.68	.37 **	6.29	.35 ***	3.32
FF_bin_avg_-path_length							-.22 *	3.13	-.17	3.26	-.18 ***	1.47
AT_bin_avg_-path_length							-.01	1.06	.00	1.07		
FF_reciprocity									-.22 **	1.96	-.24 ***	1.25
AT_reciprocity									-.09	3.49		
R^2	.00		.40		.46		.48		.51		.50	
Adj. R^2	.00		.39		.44		.45		.48		.49	
R^2 change	.00		.40 ***		.05 ***		.02		.03 ***		.50 ***	

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Table 7.2: Regression results of the influence of group bonding social capital on information diffusion.

In model 1.4, which describes the influence of the average path length on the DV, we found that the average path length in the AT network was insignificant at describing the information diffusion, whereas the average path length in the FF network is significant. This means that the shorter the path length is in the attention network, the more information gets diffused within the group. This observation is rather surprising because here the metrics from the attention network are better able to predict the information diffusion.

Model 1.5 explores the additional inclusion of the relational social capital dimension, which is measured by the overall reciprocity and average tie strength in the networks. Interestingly, the FF reciprocity metric contributes significantly to explaining the DV: indeed, it seems to have a negative effect on the amount of retweets exchanged in the group. This indicates that groups with less reciprocity in the FF network seem to be better at diffusing information inside the group[4], which seems counter intuitive at first. On second thought this finding is not so surprising as we are going to expect option leadership inside the group (see hypothesis H2a) which can only exist when the group is not totally free of hierarchy. And hierarchy is created by non-reciprocal relationships, or in other words people that are followed by many but do not follow back.

Finally, model 1.6 shows how the best indicators from the structural and relational dimensions of social capital are able to explain 50% of the variance in the DV. The model indicates that the best way to describe information diffusion inside the group is by using a combination of the structural and relational group social metrics from the FF and AT network. The final set of indicators in model 1.6 showed that these bonding social capital metrics are all highly significant. The IVs LN_AT_density and AT_bin_clustering both explain most of the variance, and are, as earlier, positively correlated with the outcome variable. In contrast, FF_bin_avg_path_length and FF_reciprocity are both negatively correlated with the DV. While the negative correlation of FF_bin_avg_path_length with the DV is not surprising, as shorter path lengths lead to more diffusion; the negative influence of FF_reciprocity on information diffusion is interesting. Indeed, it shows that groups with less reciprocity tend to be better at disseminating information inside the group than groups with more reciprocity. This finding indicates that such groups must follow some sort of hierarchical structure where opinion leaders disseminate information inside the group. This will be explored in greater detail in the next set of hypotheses concerning the influence of individual bonding social capital on information diffusion.

Discussion

After studying the groups' social structures as described in the plausibility assumption, it can safely be assumed that networks formed around users' interests, i.e., interest-based networks, possess the same properties as social networks. The results of this section showed that

[4]The overall fit of the model is improved only slightly (R^2 change of 0.03)

hypothesis 1a can be accepted, proving that the more a group creates structural group bonding social capital the group, the more information diffuses through the group network. The most predictive metrics here are the group's density in the AT network and the clustering in the AT network, followed by the average path length in the FF network.

However, hypothesis 1b, which states that the more the group creates relational social capital the group, the more information diffuses through the group network, had to be rejected. Indeed, the IV FF_reciprocity showed a negative effect on information diffusion, and the influence of AT_reciprocity on information diffusion was found to be insignificant.

7.2 Influence of individual bonding social capital on information diffusion

Hypothesis 2a: The more structural individual bonding social capital one realizes in his own group, the more one's tweets are retweeted inside the group.

Hypothesis 2b: The more relational individual bonding social capital one realizes in his own group, the more one's tweets are retweeted inside the group.

Hypothesis 2c: The more cognitive individual bonding social capital one realizes in his own group, the more one's tweets are retweeted inside the group.

7.2.1 Descriptive statistics

Descriptive statistics

The dataset contains no missing samples, since all of the network measures were computed for all of the individuals. The measures that were computed on a dichotomized version of the network were marked with a "_bin" in the variable name. Measures that were transformed were marked with the corresponding transformation in front of the variable (e.g., LN_AT_bin_in_deg). A descriptive overview of the variables can be found in table 7.3.

Distributions of variables

The descriptive statistics of the IVs and DVs show that the IV's in dataset of 15 413 individuals are non-normally distributed due to the significant K-S tests of normality (see table 3 in the appendix). However, the individual Q-Q plots indicated that a LN transformation of the FF_bin_deg, FF_bin_close, AT_bin_close metrics and the AT_rec and FF_rec reciprocity metrics was not necessary since the transformed versions did not improve the results. The remaining

Measure	Min	Max	**Mean**	Std. Deviation
RT_vol_in	0	1009	**26.84**	48.54
Ranking_in_interest	1.00	101.00	**49.97**	29.19
FF_bin_in_deg	.00	.93	**.33**	.19
AT_vol_in	.00	27385.00	**115.13**	308.13
FF_bin_close	.00	1.00	**.56**	.19
AT_bin_close	.00	.85	**.37**	.19
FF_bin_page	.00	.08	**.01**	.01
AT_bin_page	.00	.34	**.01**	.01
FF_rec	.00	1.00	**.46**	.25
AT_rec	.00	1.00	**.28**	.22
AT_avg	.00	537.00	**6.29**	11.30

Table 7.3: Descriptive overview of the individual social capital measures and outcome

variables were transformed using the LN transformation, which was then indicated in the variable name.

Regression Method: Stepwise Enter

The regression method stepwise enter was used and resulted in six different models. Model 2.1 captures only the influence of the cognitive dimension of social capital on information diffusion. Models 2.2 - 2.4 assess the influence of the cognitive and structural dimensions of social capital on information diffusion, and finally, model 2.5 captures the cognitive, structural and relational dimensions of social capital. Model 2.6 contains the best predictors of individual bonding social capital for each dimension.

Regression assumptions

The Durbin-Watson statistic of model 2.5 had a value of 1.69; for model 2.6, the value was 1.68. These values are close to 2, thus revealing that the assumption of independent errors is met. The VIF tolerance statistics for models 2.1 through 2.5 showed that the VIF is lower than 10, which confirms the assumption that multicollinearity is absent. In model 2.6, the VIF is even well below 2. The plot of *ZRESID against *ZPRED (See figure A.2 in the appendix) displays a random array of dots evenly dispersed around zero. On the histogram of the regression standardized residuals, the points lie on a straight line, and the P-P (probability) plot raises no concerns (See figure A.2 in the appendix).

7.2.2 Models of individual bonding social capital

Interpretation & discussion of the models' coefficients

Model 2.1 shows that individual cognitive social capital positively influences information diffusion: The lower a person is ranked for a particular interest, and thus the more he or she is perceived to express this interest, the more retweets he or she receives from group members. The influence of this IV is significant for models 2.1 through 2.6.

Models 2.2 to 2.4 explore the structural dimensions of social capital: The amount of @ mentions each individual receives from members of his own group, as measured by the IV LN_AT_vol_in, has the highest influence on the amount of retweets that person obtains from group members. This means that opinion leaders who are often solicited by members of their community also succeed at diffusing information inside the group. However, the attention that an individual receives, as measured by FF_bin_in_deg, has a much smaller impact on the retweet volume the individual obtains. This is an indication that receiving attention from members of the community does not play an important role in receiving retweets, whereas receiving @ mentions does. This finding has also been confirmed in prior studies of Twitter. It is important to note that the very simple model 2.2 already explains 42% of the variance in the DV. Model 2.3 explores the additional inclusion of the closeness metric of each individual on the DV. The model shows that the closeness metric from the AT network has a much higher influence on the DV (Beta of .22) than the FF_bin_close metric[5] (Beta of -.08). Indeed, individuals that play a central role in the groups' interaction do tend to receive more retweets. Moreover, Model 2.4 shows that by taking into account the PageRank of each individual in the network, only small improvements can be made to the exploratory value of the model. The R^2 increased only by 0.01 between model 2.3 and model 2.4. This reveals that indirect social capital is not an important driver of retweets obtained.

Model 2.5 tests the influence of the individual relational social capital dimension on the amount of retweets an individual is able to obtain. We found that the more an individual's interactions displays reciprocity, the more that person's information is diffused inside the group. This finding is also mirrored in the positive coefficients of the average tie strength of an individual, which has a positive effect on the number of retweets the individual obtains. Yet both effects have a Beta smaller than .10, indicating that from a statistical point of view, this is a negligible effect. At the same time, this model also shows that reciprocity in the friend and follower relations can have a negative influence on the amount of retweets obtained. This indicates that individuals whose messages are highly retweeted do not hold reciprocal relationships with their followers. The model thus highlights the fact that people who (automatically) follow back actually have less chances of receiving retweets from group members. Combined with the insights given by the structural dimension of social capital, this confirms the finding that groups

[5] Notice that the metrics from the AT and FF network have a different effect direction on the DV. Since in models 2.4 and 2.5, the effect of the FF_bin_close IV is diminished through the addition of other variables, and finally rendered insignificant in model 2.5, this effect is likely an artifact of multicollinearity of those two metrics.

	Model 2.1		Model 2.2		Model 2.3		Model 2.4		Model 2.5		Model 2.6	
	Beta	VIF	Beta	VIF	Beta	VIF	Beta	VIF	Beta	VIF	Beta	VIF
Ranking_in_-interest	-.27***	1.00	-.08***	1.15	-.11***	1.24	-.09***	1.34	-.09***	1.35	-.10***	1.27
LN_AT_vol_-in			.53***	1.39	.43***	1.93	.43***	2.75	.34***	4.71	.42***	1.94
FF_bin_in_-deg			.14***	1.49	.12***	2.41	.08***	2.78	.12***	3.11		
AT_bin_close					.22***	2.17	.22***	2.32	.18***	2.67	.24***	1.74
FF_bin_close					-.08***	2.07	-.06***	2.13	.00	2.64		
LN_AT_bin_-page							-.03***	1.76	-.04***	1.79		
LN_FF_bin_-page							.10***	1.71	.11***	1.72	.12***	1.38
AT_rec									.06***	2.02		
LN_AT_avg									.08***	2.64		
FF_rec									-.12***	1.72	-.07***	1.08
R^2	.07		.42		.44		.45		.46		.45	
Adj. R^2	.07		.42		.44		.45		.46		.45	
R^2 change	.07 ***		.35 ***		.02 ***		.01 ***		.01 ***		.45 ***	

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Table 7.4: Regression results of the influence of individual bonding social capital on information diffusion.

tend to have a highly hierarchical structure. Since the addition of the relational dimension significantly improved the explanatory value of the model, the IV FF_rec was included in the final model. The final model 2.6 shows that 45% of the variance in the DV can be explained by using only 5 of the original 10 IVs. In this model, the indegree and closeness metrics computed on the AT network have a higher Beta value than their FF counterparts. Moreover, the PageRank and the reciprocity metrics derived from the FF network have a higher influence on the DV than the metrics from the AT network. The relational and cognitive dimensions of social capital thus appear to be significant predictors of internal information diffusion.

Discussion

Hypothesis 2a postulated that the person's structural bonding social capital could positively influence the amount of retweets that the individuals receives from group members. According to the results, this hypothesis was accepted. In particular, the centrality metrics that emerged from the AT network were able to explain a high proportion of variance in the DV. Thus whenever data on the amount of interactions of an individual is available it should be preferred to the only structural data of the friend and follower network, since it is better able to explain information diffusion.

Hypothesis 2b stated that "the more relational individual bonding social capital one realizes in his own group, the more one's tweets are retweeted inside the group". The results showed evidence leading to the rejection of this hypothesis. Indeed, the results from the regression model showed that higher average tie strengths, as measured by the IV LN_AT_avg, had too small of an effect ($\leq$.10) on the DV for it to be considered significant (Cohen, 1988). The effect of reciprocity in the AT network (AT_rec) was also too small (.06) to be considered significant. Moreover, the reciprocity in the FF network, as measured by the IV FF_rec, had a negative effect on the DV. Therefore, hypothesis 2b was rejected.

Hypothesis 2c was accepted since the regression models showed that the more cognitive social capital the individual collects, the more retweets he is able to obtain from group members. This IV had a significant and moderate effect in each model of the regression. This means that indeed individuals that are able to well serve the interests of the interest group are retweeted more often.

7.3 Influence of group bridging social capital on information diffusion

Hypothesis 3a: The more structural group bridging social capital the group realizes, the more its tweets are retweeted by other groups.

Hypothesis 3b: The more relational group bridging social capital the group realizes, the more its tweets are retweeted by other groups.

Hypothesis 3c: The more cognitive group bridging social capital the group realizes, the more its tweets are retweeted by other groups.

7.3.1 Descriptive statistics

Visual description of the blockmodel network

As mentioned in chapter 5, the basis for the empirical analysis of the previous hypotheses is the blockmodel network. This blockmodel network was constructed by including all group members of a group into a single node (see section 5.3.2 in chapter 5). The blockmodel network of nodes is the basis for this analysis and can be seen in figure 7.1. This figure shows the interaction patterns in the blockmodel of the 166 groups studied. The attention (FF) and information diffusion (FF) blockmodel networks can be found in the appendix in figure A.5 and A.6. The displayed interaction network was highly trimmed[6] (all statistical computations were computed on the unmodified network). Indeed, edges with a weight less than 250 were removed[7]. Studying the structure of the emerging network of interests groups leads to a number of interesting descriptive statements: First, the network revealed that most topic areas are densely connected to other topic areas. This lead to the emergence of a semantic map (Doerfel and Barnett, 1999; Doerfel, 1998) of the 166 selected interest groups. The spring embedding layout (Wasserman and Faust, 1994) of this semantic network generally places interests that are cognitively related next to each other. This is a direct result of the network data where cognitively related groups also exchange a high number of ties. This leads to higher tie strengths between groups that have similar interests, which in the layout leads to the mutual attraction of these groups.

In figure 7.1, the blockmodel presents a core-periphery structure and the interaction between groups seems to favor some groups. Indeed, it seems that some groups were being interacted with more than others. Such groups are found in the core of the network and cover interests such as “politics_news”, “humanrights_activism_justice”, “tech”, “newspaper”. Interestingly, such groups are generally seen as highly cognitively diverse. These central groups tend to interact with all the other groups in the sample, therefore resulting in such a high centrality in the AT network. Around this core is a periphery of multiple “special-interest” clusters. In the middle upper left of the network, there are clusters focused on career, work and finances, such as “ceo”, “career_employment”, “banking”, “investor”, “accounting”, or clusters focused on technology such as “tech”, “mobile_smartphone”, “developer”, “mac_iphone”. On the lower left, there are clusters focused on entertainment, represented by keywords such as “tvshows_-drama_actor_hollywood”, “comedy_funny”, “director”, “youtube” and so on. In the lower right

[6] For enhanced visibility, the reader is advised to study this network in the electronic version of this thesis, which allows the reader to seamlessly zoom into the high-dimensional network structure.

[7] This corresponds to removing edges that correspond to less than 250 interactions between groups. Such trimming is performed for visualization reasons, in order to better reveal the network structure and hide clutter that emerges from too many network ties.

corner, there are topics related to health, food and the outdoor, as represented by labels such as “exercise_fitness”, “recipes_cooking”, “nature” or “gardening”. Finally, in the upper right corner, there are topics related to healthcare, social and family issues, such as “parenting”, “psychology_mentalhealth”, “healthcare_medicine” or “pregnancy”. The transitions between these 166 interest groups form a very detailed, highly accurate interest network. This network of user interests is in stark contrast to the artificially-created ontology, used in chapter 6 to collect the interest categories. The ontology used to structure the unstructured set of user interests was useful for providing users’ interests at the same cognitive level of granularity, but the ecosystem of interconnected user interest groups in Twitter can be better studied by looking at the aggregated network data.

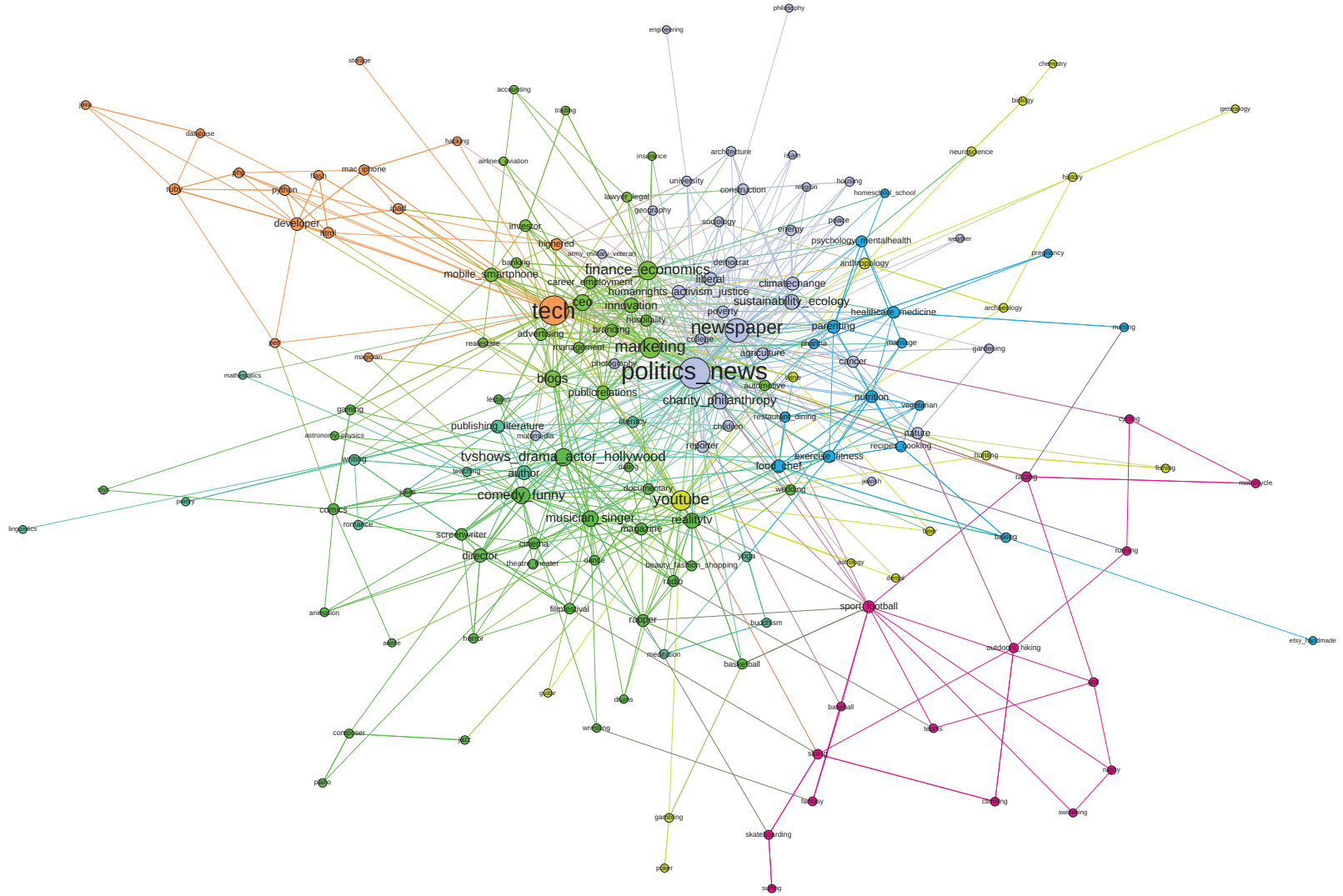


Figure 7.1: Reduced version of the blockmodel of the AT network. Each node represents the network of 100 Twitter users for this group. For enhanced visibility edges with weight of < 250 have been removed; coloring chosen according to modularity detection; node size chosen according to AT degree; Yifan Hu Layout.

Descriptive statistics

The dataset contains no missing samples, since all of the group network measures were computed for all of the groups. The measures that were computed on a dichotomized version of the network were marked with a "_bin" in the variable name. Measures that were transformed were marked with the corresponding transformation in front of the variable (e.g., LN_FF_-volume_in). A descriptive overview of the variables can be found in table 7.5.

In order to describe the group's cognitive social capital, the metric Agg_number_of_interests was used. The metric Agg_number_of_interests is an aggregate measure that corresponds to the total number of interests in the group. Indeed, it adds up all the interests of all individuals in the group. If each member of the group has only one interest, this results in a total of 101 nominations for the group, which means that the group overall has a low cognitive bridging capital, since its members don't seem interested in other topics (e.g., "golf" or "mathematics"). On the other end of the spectrum, the group with the highest value of Agg_number_of_interests is the group "tech", whose members collected 184 interest nominations, which is almost 2 interests per member. Since this group presents a high cognitive diversity, it is likely to pique other groups' interest and build conversations with other groups. The most important structural indicators of social capital are FF_volume_in and AT_volume_in. These two indicators measure the number of friendship ties, or @-mentions, that a group managed to collect as a whole. The table 7.5 reveals a wide standard deviation for these indicators: While some groups did not obtain any incoming ties from other groups, the group with the interest labeled as "marketing" managed to collect 67 244 friend and follower ties, and the group "politics_news" collected 66 671 @-mentions. Due to the structure of the resulting blockmodel networks and their high clustering (to see the blockmodel network, see figure 7.1; for the FF- and RT-network, see figure A.6 and figure A.5 in the appendix), the betweenness metrics resulted in rather small values. This is due to the numerous connections between each pair of groups, and to the fact that no group dominates the shortest paths between other groups. The closeness metrics are fairly normally distributed (see distributions of IVs below) among the different groups, with only a small standard deviation. Finally, the PageRank metric suffers from the same high clustering of the overall network, resulting in rather low values since no central group is able to dominate this measure. The indicators of the relational social (FF_rec, AT_rec, FF_avg, AT_avg) capital show that groups exhibit reciprocal relationships in the friend and follower network, with a large variance in tie strength.

Distributions of variables

The descriptive statistics of the IVs and DVs show that the IV's in the dataset of 166 groups have rather log-normal distributions. The K-S and Shapiro-Wilk tests of normality (see figure 4 in the appendix) indicate that the majority of the independent variables (IVs) are not normally distributed. The only IV that is normally distributed according to the K-S

Measure	Min	Max	**Mean**	Std. Deviation
RT_volume_in	27	9074	**1446.06**	1532.33
Agg_number_of_interests	101	184	**116.82**	16.42
Member_count	598	220 652	**13880.66**	26713.50
FF_volume_in	0	67 244	**6864.05**	9796.64
AT_volume_in	0	66 671	**4535.46**	8463.95
FF_bin_betweenness	0.00	.13	**.01**	.02
AT_bin_betweenness	0.00	.11	**.01**	.02
FF_bin_closeness	.26	.76	**.46**	.07
AT_bin_closeness	.23	.43	**.34**	.04
FF_bin_pagerank	.00	.05	**.01**	.01
AT_bin_pagerank	.00	.09	**.01**	.01
FF_rec	0.00	1.00	**.47**	.23
AT_rec	0.00	1.00	**.29**	.18
FF_avg	140.50	578.75	**278.45**	73.49
AT_avg	157.25	903.50	**371.26**	139.39

Table 7.5: Descriptive overview of group bridging social capital measures and outcome

statistic is the AT_closeness metric. A manual inspection of the Q-Q plots indicates that the variables Agg_number_of_interests, FF_closeness, FF_rec and AT_rec metrics do not need a transformation. The rest of the IVs follows a log-normal distribution and was transformed using a LN transformation. The DV RT_volume also follows a log-normal distribution, and was transformed using the LN transformation. All transformations were subsequently indicated in the variable name with an _LN underscore.

Regression Method: Stepwise Enter

The regression method used was stepwise enter and resulted in seven different models. Model 3.1 captures only the influence of the cognitive dimension of social capital on information diffusion. Models 3.2 through 3.5 assess the influence of the cognitive and structural dimensions of social capital on information diffusion and model 3.6 captures the cognitive, structural and relational dimensions of social capital. Model 3.7 contains the best predictors of group bridging social capital for each dimension.

Regression assumptions

The Durbin-Watson statistic for model 3.6 is 1.97 and for model 3.7, it is 1.81. Since these values are close to 2, the assumption of independent errors was met. In terms of the assumption of multicollinearity absence, the VIF and tolerance statistics (Bowerman, B.L. , O'Connell R, 1990) show that the models 3.1 through 3.6 have a VIF score that is not greater than 10. In model 3.7 the VIF is well below 2. The plot of *ZRESID against *ZPRED (See figure A.3 in the appendix) displays a random array of dots evenly dispersed around zero. The histogram of the regression standardized residuals reveal that the points lie on a straight line and also that the P-P (probability) plot does not raise concerns (See figure A.3 in the appendix).

7.3.2 Models of group bridging social capital

Interpretation & discussion of the models' coefficients

Model 3.1. explores the influence of the cognitive dimension of social capital on the amount of retweets a group is able to obtain. With a Beta value of .67 in this model, the IV Agg_-number_of_interests explains almost 50% percent of the variance (R^2 of .49) of the DV. This means that information diffusion at the group level is highly heterophile. Groups that tend to exhibit multiple interests are able receive the greatest number of retweets. Additionally, in model 3.1 through 3.5, the high significance and high positive influence of the cognitive variable Agg_number_of_interests on the outcome variable can be observed. However, the control variable Member_count is insignificant in all models. This means that the breadth of a group has no effect on the DV when the cognitive dimension is taken into account.

Model 3.2 reveals that in the structural dimension, the LN_AT_volume_in (i.e., the number of interactions received from other groups) is a strong and significant predictor at explaining the number of retweets the group obtains from other groups. Groups that are able to spark a high number of interactions with other groups are also successful at obtaining retweets from other groups. The @ mentions a group can obtain from other groups (LN_AT_volume_in) supersedes the attention the group is able to garner from other groups (LN_FF_volume_in) - a pattern that was observed at the individual level in the previous section 7.2. This means that once groups succeed at being mentioned by other groups, it becomes less important to get a high number of followers from other groups.

Model 3.3 shows that the groups betweenness metric is insignificant at explaining how many retweets the group can obtain from other groups. This can be explained by the highly clustered structure of the network, and the high amount of ties between groups, where no group is able to position itself in between different groups. This means that at the group level, the underlying model of the betweenness metric does not apply, because information does not have to travel from group A to group C through group B. Instead, information originates in group A and is then directly consumed by the corresponding groups without intermediaries. Therefore, only

	Model 3.1		Model 3.2		Model 3.3		Model 3.4		Model 3.5		Model 3.6		Model 3.7	
	Beta	VIF	Beta	VIF	Beta	VIF	Beta	VIF	Beta	VIF	Beta	VIF	Beta	VIF
Agg_number_of_interests	.67 ***	1.28	.42 ***	1.87	.39 ***	2.10	.33 ***	2.16	.35 ***	2.32	.35 ***	2.56	0.39 ***	1.73
Member_count	.06	1.28	.00	1.32	-.02	1.88	.02	1.96	.02	2.19	.03	2.20		
LN_FF_volume_in			.11	1.61	.12	1.72	.16*	3.15	.18*	3.54	.20*	3.94		
LN_AT_volume_in			.39 ***	1.62	.36 ***	1.74	.33 ***	1.80	.32 ***	1.87	.29 ***	2.54	.40 ***	1.55
LN_FF_bin_betweenness					-.02	2.14	.08	2.96	.13	3.97	.10	4.04		
LN_AT_bin_betweenness					.10	1.99	-.07	2.55	-.08	3.28	-.05	3.61		
FF_bin_closeness							-.29 ***	3.82	-.29 ***	4.09	-.28 **	5.00		
AT_bin_closeness							.37 ***	2.30	.38 ***	2.39	.41 ***	2.53	.30 ***	1.48
LN_AT_bin_pagerank									.07	3.60	.06	4.12		
LN_FF_bin_pagerank									-.12	5.19	-.11	5.36		
FF_rec											.04	2.92		
AT_rec											.03	1.73		
LN_FF_avg											-.19 ***	2.24	-.14 ***	1.40
LN_AT_avg											.11 *	2.38		
R^2	.49		.61		.62		.68		.68		.70		.67	
Adj. R^2	.48		.60		.60		.66		.66		.68		.66	
R^2 change	.49 ***		.12 ***		.00		.06 ***		.00		.02 *		.67 **	

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Table 7.6: Regression results of the influence of group bridging social capital on information diffusion.

groups that can achieve a high indegree in the FF and AT network will be able to disseminate their information successfully to other groups.

Model 3.4 shows how the closeness centrality metric successfully explains a significant part of the variance in the DV, and also leads to a significant improvement of the R^2. In combination with the observation that highly cognitively diverse groups are able to receive a high number of retweets, this means in other words that groups with members who have a highly diverse cognitive bridging social capital, seem to emerge in the center of the interaction network. This is indicated by the positive influence of closeness centrality metric (AT_bin_closeness) on the DV. As depicted in figure 7.1, groups that are in the center of interactions, with diverse topics such as "politics", "news" or "tech", seem to be successful at spreading their information to the other groups. In contrast, highly specialized groups such as "programming" or "anthropology", with lower closeness values in the AT_bin_closeness metric, are less at the center of interactions and therefore, are also less retweeted. The closeness metric from the AT network has a positive influence on the DV, whereas the metric from the FF network has a negative influence on the DV[8].

Model 3.5 explores if a certain group has a "hidden influence" on other groups, measured by the PageRank metric. The regression shows that the PageRank metric is not significant at the group level. Interestingly, this IV was a significant metric for individual bonding social capital (see table 7.4). One explanation for the insignificance of this metric for group bonding social capital could be that no indirect social capital can be exhibited at the group level. Groups simply openly compete for the direct attention of other groups.

Model 3.6 explores how the relational measures of social capital are able to predict information diffusion. On one hand, a very small effect shows that the higher the average number of interactions between groups (LN_AT_avg), the more retweets the group is able to obtain. On the other hand, there is a stronger opposite effect that uses the LN_FF_avg metric: The lower the average number of attention ties with other groups, the more retweets the group is able to obtain. This finding confirms that, even at the group level, there are hierarchical tendencies between groups, where certain groups get a lot of attention, yet they do not follow other groups. This seems plausible since many central groups contain accounts that serve only as news outlets for news corporations, governmental organizations, or political organizations. Additionally, reciprocity at the group level is not a significant indicator for obtaining retweets from other groups.

Finally, model 3.7 shows the combination of the IVs with the most predictive power. These are used to best capture the group's bridging social capital. These IVs are able to explain 66% of the retweets the group obtains from other groups. Model 3.7 shows that the overall cognitive diversity of the group's members (Agg_number_of_interests), the interaction with the group at an aggregate level (LN_AT_volume_in), and its interaction centrality (as measured by the

[8]This can be explained by the multicollinearity of these two IVs, as indicated in the highest VIF scores of the significant predictors. When only the FF closeness metric is put into the regression (this is not shown), its negative effect disappears.

groups closeness (LN_AT_bin_closeness) and the negative influence of the average tie strength to other groups (LN_FF_avg)) best explain how well the group's information is diffused by other groups.

Discussion

Hypothesis 3a postulated that the more a group created structural group bridging social capital, the more its information would be disseminated by other groups. According to the results, this hypothesis can be accepted. Indeed, groups at the centers of attention and interaction received the highest amount of retweets.

Hypothesis 3b stated that "the more relational group bridging social capital the group realizes, the more its tweets are retweeted by other groups". The results of this analysis provided evidence that disproves this hypothesis. Indeed, the regression showed that reciprocal relations had no significant effect on the outcome variable. Moreover, the average tie strength between groups, as measured by the IVs LN_AT_avg and LN_FF_avg, was either almost non-significant, or had a significant negative influence on the DV. This means that groups obtain a large number of retweets thanks to their central position within the network (AT_bin_closeness, LN_AT_volume_in) and not because of the efforts they make to manage the relational dimension of social capital. Therefore, Hypothesis 3b was rejected.

Finally, regarding the cognitive dimension of social capital, Hypothesis 3c was accepted. Indeed, the total number of user interests within a group, which measures the groups' overall cognitive bridging social capital, has a strong positive influence on the number of retweets the group can obtain.

7.4 Influence of individual bridging social capital on information diffusion

Hypothesis 4a: The more individual cognitive bridging social capital one realizes, the more one's tweets are retweeted by members of other groups.

Hypothesis 4b: The more individual relational bridging social capital one realizes, the more one's tweets are retweeted by members of other groups.

Hypothesis 4c: The more individual cognitive bridging social capital one realizes, the more one's tweets are retweeted by members of other groups.

Measure	Min	Max	**Mean**	Std. Deviation
RT_vol_in	0	1086	**15.59**	42.23
Number_of_interests	1	29	**1.17**	.58
FF_vol_in	0	3950	**108.02**	199.41
AT_vol_in	0	28 954	**68.25**	382.12
FF_groups_in	0	165	**32.27**	34.12
AT_groups_in	0	165	**9.66**	17.06
FF_bin_betweenness	0.00	.08	**.00**	.00
AT_bin_betweenness	0.00	.05	**.00**	.00
FF_rec	0.00	1.00	**.24**	.25
AT_rec	0.00	1.00	**.08**	.11
AT_avg_tie_strength	0.00	199.00	**2.67**	4.42
AT_strength_centrality_in	0.00	9933.19	**23.92**	136.21

Table 7.7: Descriptive overview of the individual bridging social capital measures and outcome

7.4.1 Descriptive statistics

Descriptive statistics

The descriptive statistics show that people are listed as having, on average, 1.17 interests. This number is close to one, which indicates that grouping according to topics is a viable assessment (this was performed in chapter 6.4). The IVs FF_vol_in and AT_vol_in show a high standard deviation with a mean of 108 and 68 respectively, suggesting there is a high difference in popularity among actors in the network. The betweenness metrics have very low mean values and suffer from a high clustering of the networks. The FF and AT reciprocity metric have respective means of .24 and 0.08, showing, that the majority of edges between members of different groups are not reciprocated. The mean average tie strength in the AT network is 2.67, indicating that members hold on average between two or three interactions with members of different groups. The outcome variable RT_vol_in has a mean of 15.15, indicating that, on average, a member receives 15 retweets in the sample. In contrast, some members[9] were able to collect over 1000 retweets from people that did not belong to their own group.

Distributions of variables

The Kolmogorov-Smirnov test of normality (see table 5 in the appendix) indicated that the majority of the independent variables were not normally distributed. The only IVs that

[9] e.g., a democrat with the Twitter handle "@HarryReid"

came close to being normally distributed, according to their individual Q-Q-plot, were the FF_groups_in, AT_groups_in, FF_rec and AT_rec metric. However, they still did not fulfill the K-S criterion. The rest of the IVs (FF_vol_in, AT_vol_in, FF_bin_betweenness, AT_bin_-betweenness, AT_avg_tie_strength, Number_of_interests) follow a log-normal distribution and were transformed using an LN transformation. The DV RT_volume_in also followed a log-normal distribution, and was transformed using the LN transformation. All transformations were subsequently indicated in the variable name with an _LN underscore.

Regression Method: Stepwise Enter

The stepwise enter regression method was used and resulted in six different models. Model 4.1 captures only the influence of the cognitive dimension of social capital on information diffusion. Models 4.2 through 4.4 assess the influence of the cognitive and structural dimensions of social capital on information diffusion. Model 4.5 captures the cognitive, structural and relational dimensions of social capital. Model 4.6 contains the best predictors of individual bonding social capital for each dimension.

Regression assumptions

The Durbin-Watson statistic reported the value of 1.71 for model 4.5 and 1.86 for model 4.6. This value is close to 2, thus revealing that the assumption of independent errors is met. The VIF tolerance statistics for models 4.1 to 4.5 showed that the VIF score is not greater than 10, which addresses the assumption that multicollinearity is absent. In model 4.6, the VIF is well below 3.5. The plot of *ZRESID against *ZPRED (See figure A.4 in the appendix) displays a random array of dots evenly dispersed around zero. On the histogram of the regression standardized residuals, the points lie on a straight line, and the P-P (probability) plot raises no concerns (See figure A.4 in the appendix).

7.4.2 Models of individual bridging social capital

Interpretation & discussion of the model's coefficients

In this regression, six different models were introduced. Models 4.2 to 4.5 build on the baseline model 4.1 by successively extending the number of IVs. The R^2 of all models and its change indicate that the baseline model 4.1 and the four successive models 4.2 through 4.5 make significant contributions to explaining the DV. However, models 4.3 and 4.4 presented only small improvements to previous models. For this reason, the decision was made to not include IVs from these models in the final model 4.6.

	Model 4.1		Model 4.2		Model 4.3		Model 4.4		Model 4.5		Model 4.6	
	Beta	VIF	Beta	VIF	Beta	VIF	Beta	VIF	Beta	VIF	Beta	VIF
LN_Number_-of_interests	.45 ***	1.00	.16 ***	1.24	.15 ***	1.34	.15	1.35	.14 ***	1.36	.13 ***	1.26
LN_AT_vol_in			.56 ***	1.91	.50 ***	3.07	.50 ***	3.09	.45 ***	4.85	.48 ***	2.88
LN_FF_vol_in			.16 ***	1.85	.26 ***	6.01	.26 ***	6.01	.37 ***	6.82	.29 ***	3.15
FF_groups_in					-.13 ***	6.07	-.13 ***	6.12	-.09 ***	6.25		
AT_groups_in					.10 ***	3.12	.10 ***	3.49	.03 ***	4.13		
LN_FF_bin_betweenness							-.01	1.08	.00	1.09		
LN_AT_bin_betweenness							-.01	1.29	.00	1.31		
FF_rec									-.17 ***	1.79	-.17 ***	1.64
AT_rec									.02 ***	1.36		
LN_AT_avg_-tie_strength									-.01 ***	1.47		
R^2	.20		.58		.58		.58		.60		.59	
Adj. R^2	.20		.58		.58		.58		.60		.59	
R^2 change	.20 ***		.37 ***		.00 ***		.00 ***		.02 ***		.59 ***	

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Table 7.8: Regression results of the influence of individual bridging social capital on information diffusion.

In terms of the cognitive dimension of social capital, we found that in models 4.1 to 4.6, IV LN_Number_of_interests was a significant and moderate predictor of the number of retweets obtained from members of other groups. This demonstrates that brokers who exhibit a variety of interests are indeed successful at obtaining retweets from members of other groups. In terms of the structural dimension of individual bridging social capital, the standardized Beta coefficients in model 4.2 revealed that the IV LN_AT_vol_in is the strongest predictor of retweets from members of other groups. It was found to be significant for models 4.2 through 4.6. Yet its effects on the DV seemed to vary: On one hand, the IV LN_FF_vol_in in model 4.2 was the weakest predictor (.16) of the DV, yet it became more important (.37) in models 4.2 to 4.5. On the other hand, the strong effect (.56) of the IV LN_AT_vol_in slowly decreased (.45) when more IVs were added to the model. This means that parts of the effects of structural social capital were compensated by the addition of relational social capital into the model.

Model 4.3 explored the group diversity metric, which predicts that getting attention from, or interacting with, at least one member of another group also increases the chance of being retweeted by members of other groups. The influence of the group diversity metrics FF_groups_in and AT_groups_in were found to significant but very small[10] (with Beta values of -.13 and .1 respectively), which led to the decision to not include these effects in the final model 4.6.

Model 4.4 explored how the betweenness of actors in the overall network affected their potential to being retweeted. In both models 4.4 and 4.5, the effect of this IV was insignificant. This might be explained by the high amount of ties between members of different groups: indeed, in this case, no actor can actually benefit from brokering information between members. This means that bridging social capital results from the central position of the users in the global Twitter network, rather than their brokerage position. Members that are able to successfully capture the attention of members from different groups are also successful at receiving information from those members.

The analysis of the relational dimension of individual bridging social capital in model 4.5 paints a coherent picture: First, we find a negative effect between reciprocity (FF_rec) and the DV (Beta of -.17), suggesting that certain hubs do not follow back other members, yet their information is highly retweeted. Secondly, we find only very small positive effects of reciprocal interaction with members on the DV (Beta of 0.02). Finally, we see that the average tie strength (LN_AT_avg_tie_strength) also showed a marginal negative effect (-.01) on the DV. Thus, the relational dimension of bridging social capital seems to be insignificant or even have a negative effect (FF_rec) on information diffusion between groups.

The final model 4.6 that contained the strongest significant predictors from each dimension of social capital and was able to explain 59% of the variance of the DV with only four IVs. This

[10] Additionally, these IVs show effects of multi collinear behavior. This can be seen in the negative effect the IV FF_groups_in on the DV, whereas in the interaction network, the effect of AT_groups_in on the DV is positive. Entering both variables separately showed positive effects for both of them. The high VIF scores of 6.25 and 4.13 also indicated that a significant amount of information resulting from these metrics was already captured by the IVs from model 4.2.

means that all three dimensions of social capital are well able to predict the amount of retweets a person receives from members of other groups.

Discussion

The results supported hypothesis 4a. Indeed, the analysis revealed that the more a person creates structural bridging social capital, the higher the chances that person has of receiving retweets from members of other groups. The amount of attention (LN_FF_vol_in) and the amount of interaction (LN_AT_vol_in) are particularly good at capturing this dimension.

Hypothesis H4b, which stated that "the more individual cognitive bridging social capital one realizes, the more one's tweets are retweeted by members of other groups", had to be rejected. Indeed, the effects of tie strength on the DV were very small (Beta of 0.1 for LN_AT_avg-tie_strength), and the effect of reciprocity in the FF network on the DV was even negative (Beta of -.17). This means that, in the context of bridging social capital, stronger ties do not lead to more retweets when the cognitive and structural dimensions are controlled for.

Finally, hypothesis 4c was accepted. The more a person creates individual cognitive bridging social capital, as measured by the number of different interests they are perceived to have, the more their content is retweeted by members of other groups. Thus brokers are not only defined by their structural dimension, but also by their ability to cognitively express a number of different interests at the same time.

7.5 Interdependence between bonding and bridging social capital on information diffusion

Hypothesis 5: The more group bonding social capital the group obtains, the less retweets it obtains from other groups.

Hypothesis 6: The more individual bonding social capital the individual obtains, the less retweets he obtains from members of other groups.

Hypothesis 7: The more group bridging social capital the group obtains, the less the group's tweets are retweeted inside the group.

Hypothesis 8: The more individual bridging social capital the individual obtains, the less retweets he obtains from members of his own group.

Method: Regression with path model

To answer the hypotheses concerning the cross effects of bridging and bonding social capital on information diffusion, the corresponding models were combined at both the individual and

group level. This means that two different path models were created: one for the analysis of cross effects at the group level and one for the analysis of cross effects at the individual level. Each of these path models (as shown in figure 7.2 and figure 7.3) contains the indicators of the best models for each quadrant of social capital: The group path model contains the indicators from models 1.6 and 3.7. The individual path model contains the indicators from models 2.6 and 4.6.

7.5.1 Cross effects of group bonding and group bridging social capital

Description

At the group level, information diffusion can follow different paths. These can be seen in figure A.7 in the appendix. Indeed, the figure shows that different groups possess tendencies to disseminate information in different ways. While some groups such as "screenwriter", "director", "climatechange", "branding" or "politics_news" seem to be highly retweeted by other groups, the members of these groups do not themselves heavily retweet each other. On the other end of the spectrum, some special interest groups, such as "astronomy_physics", "army_military_veteran", "tennis", "surfing", are almost never retweeted by other groups, but the members of these groups heavily retweet each other. The path analysis is going to show if the creation of group bonding social capital harms the amount of retweets the group receives from outside, and if the creation of group bridging social capital harms the amount of retweets the group exchanges within the group.

The path analysis in figure 7.2 uses the IVs from models 1.6 and 3.7 to measure group bonding and group bridging social capital. The IVs that capture group bonding social capital are the IVs that were determined in model 1.6. For the structural dimension of social capital, these are: the clustering in the interaction network (AT_bin_clustering), the density in the interaction network (LN_AT_density) and the average path length in the attention network (FF_bin_avg_path-length). The relational dimension is captured by the reciprocity in the FF network as described by the IV FF_reciprocity. The IVs that capture group bridging social capital are the IVs that were determined in model 3.7: The cognitive dimension is represented by the group's aggregate number of interests (Agg_number_of_interests). The structural dimension of group bridging social capital is captured by the number of mentions from other groups (LN_AT_volume_in), and the centrality of the group in the interaction network (AT_bin_closeness). The relational dimension is represented by the average tie strength in the FF network (LN_FF_avg).

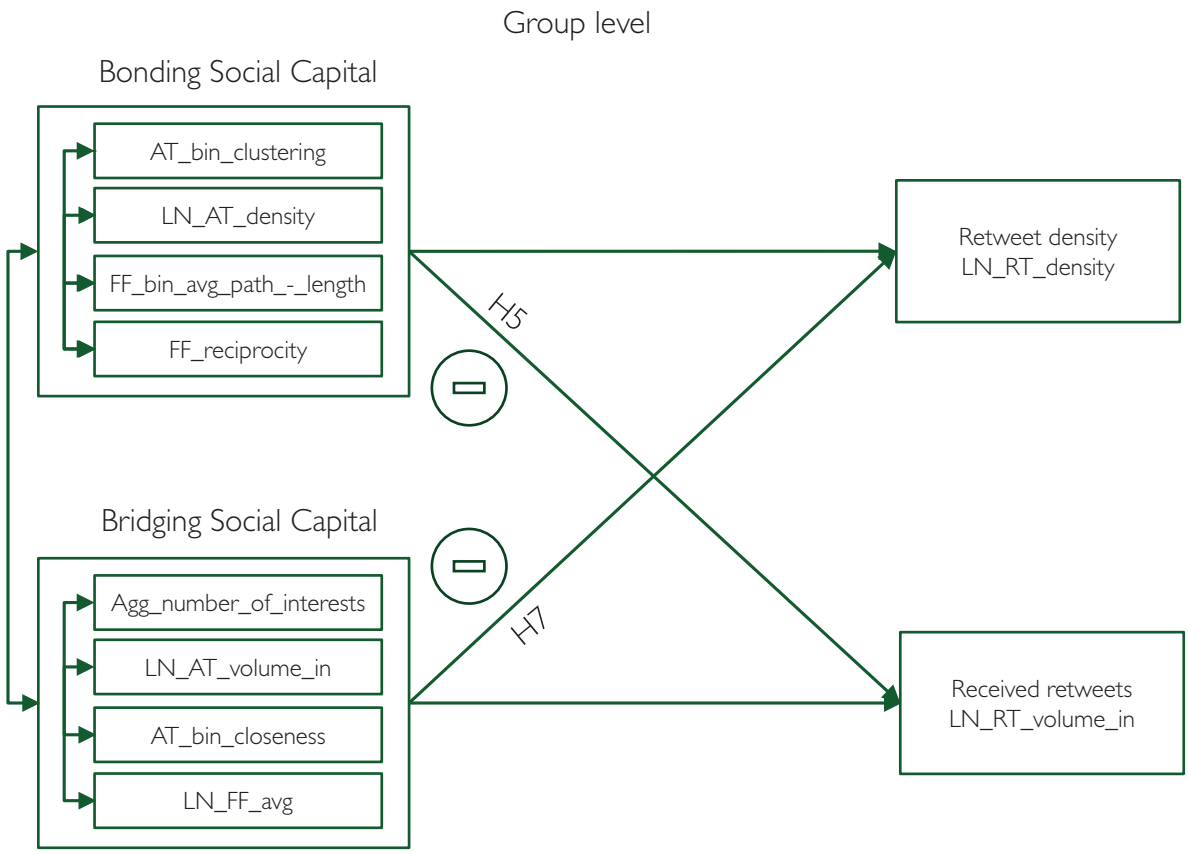


Figure 7.2: Path analysis of the cross effects on information diffusion using the IVs in models 1.6 and 3.7.

	LN_RT_density	LN_RT_volume_in
	Beta	Beta
IVs from model 1.6		
AT_bin_clustering	.34 **	.12
LN_AT_density	**.33** **	**- .19** *
FF_bin_avg_path_length	- .20 **	.00
FF_reciprocity	- .18 *	- .21 ***
IVs from model 3.7		
Agg_number_of_interests	.09	.28 ***
LN_AT_volume_in	- .09	.38 ***
AT_bin_closeness	- .02	.37 ***
LN_FF_avg	- .11	- .04
R^2	.52	.72
Adj. R^2	.49	.71
R^2 of model 1.6	.49	
R^2 of model 3.7		.66
R^2 change	.00	.05 ***

Table 7.9: Results of the path analysis on the cross effects of group bonding and group bridging social capital

Group path model of bonding and bridging social capital

Interpretation and discussion of the model's coefficients

The results of the path analysis in table 7.9 showed that there are significant moderate negative cross effects of the IVs on the opposite DV[11]. In particular, one structural indicator of group bonding, LN_AT_density, showed the strongest cross effect on information diffusion: Groups with a higher density in the interaction network (LN_AT_density), i.e., the group members strongly interact with each other, receive less retweets (Beta of -.19) from other groups. This confirms hypothesis H5, which states that the more group members focus on each other, the less "attractive" this group becomes to other groups. For the other group bonding IVs, the analysis found that their effects are either decreased or rendered insignificant when they are used to predict the amount of external information diffusion. In terms of the relational dimension of group bonding social capital, it seems that the less reciprocity a group shows (-.21), the more that group is retweeted by other groups.

The R^2 change value of .05 shows that model 3.7 can be significantly improved (.72 vs. 0.66) by taking into account the amount of interaction between group members. In terms of group bridging social capital, no cross effects on the opposite DV were found. In fact, none of the indicators of group bridging social capital particularly harmed the amount of information the group members exchanged with each other. Since no cross effects of group bridging social capital on internal information diffusion were found, hypothesis H7 was rejected.

7.5.2 Cross effects of individual bonding and individual bridging social capital

Descriptive view

The path analysis in figure 7.3 showed that the IVs that capture individual bonding social capital are the ones that were determined in model 2.6: For the cognitive bonding dimension, it is the IV bonding_Ranking_in_interest. For the structural bonding dimension, the indegree in the AT network is represented by the IV LN_bonding_AT_vol_in, the centrality in the AT network by the IV bonding_AT_bin_close and the centrality in the FF network by the IV LN_bonding_FF_bin_page. Finally, the relational bonding dimension is captured by the reciprocity in the FF network, which is described by the IV bonding_FF_rec. The IVs that capture the individual bridging social capital are the ones that were determined in model 4.6: For the cognitive dimension, this corresponds to the IV LN_bridging_Number_of_interests. For the structural dimension, the indegree in the AT network is represented by the IV LN_bridging_AT_vol_in and the indegree in the FF network (LN_bridging_FF_vol_in). The relational dimension is represented by the reciprocity in the FF network (bridging_FF_rec).

[11] This means that the sign of the influence of the IV on the DV is reversed

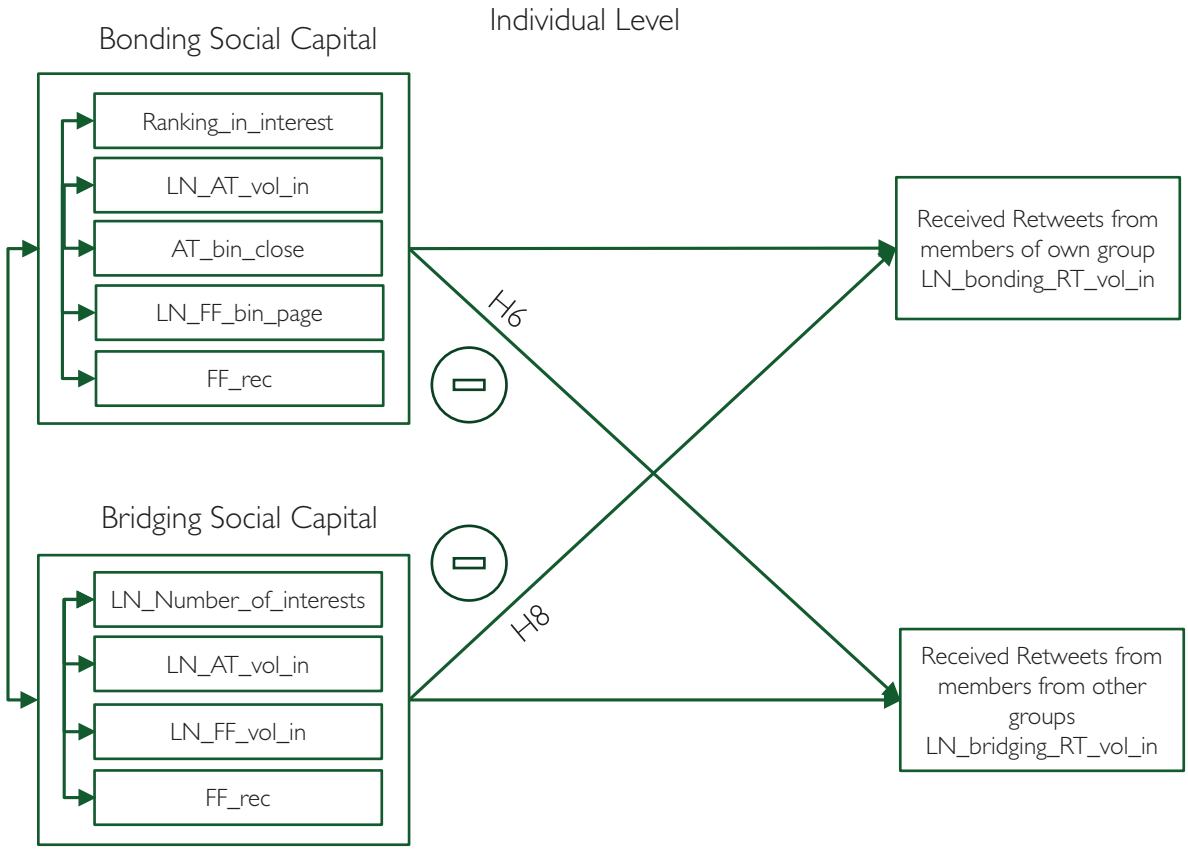


Figure 7.3: Path analysis of the cross effects of the IVS on information diffusion in models 2.6 and 4.6.

	LN_bonding_RT_vol_in	LN_bridging_RT_vol_in
	Beta	Beta
IVs from Model 2.6		
bonding_Ranking_in_interest	- .11 ***	- .02 ***
LN_bonding_AT_vol_in	**.45** ***	**- .10** ***
bonding_AT_bin_close	.22 ***	.05 ***
LN_bonding_FF_bin_page	.15 ***	.07 ***
bonding_FF_rec	- .05 ***	- .07 ***
IVs from Model 4.6		
LN_bridging_Number_of_interests	.04 ***	.12 ***
LN_bridging_AT_vol_in	**- .11** ***	**.52** ***
LN_bridging_FF_vol_in	**- .08** ***	**.25** ***
bridging_FF_rec	- .01	- .12 ***
R^2	.48	.60
Adj. R^2	.48	.60
R^2 of model 2.6	.45	
R^2 of model 4.6		.59
R^2 change	.03 ***	.01 ***

Table 7.10: Results of the path analysis on the cross effects of individual bonding and individual bridging social capital

Individual path model of bonding and bridging social capital

Interpretation and discussion of the model's coefficients

The results of the path analysis in table 7.10 showed that cross effects were indeed found for some of the IVs. Initially, the model showed that almost all IVs had a significant effect on information diffusion. Yet studying the Beta's of each IV revealed that the effect of bonding social capital metrics that accurately predict information diffusion within the group, almost disappears when it is used to predict external information diffusion (e.g., bonding_AT_bin_close .22 vs. 0.05). The same observation can be made for the indicators of bridging social capital: For example, LN_bridging_Number_of_interests has a Beta of .12 for the DV LN_bridging_RT_-vol_in, but only a Beta of .04 for the DV LN_bonding_RT_vol_in. This finding is an indication that the measures selected indeed measured what they should; namely, bonding and bridging social capital. Moreover, the path analysis slightly improved (1% and 3%) the amount of variance explained in the corresponding DVs (internal information diffusion and external information diffusion).

According to hypothesis H6 and H8, we expected to find social capital indicators that show

a positive influence on internal (respectively external) information diffusion but a negative influence on external (respectively internal) information diffusion. This was found to be true for both hypotheses. The model showed that the more mentions a person obtains from members of their own group (LN_boding_AT_vol_in), the less that person obtains retweets (LN_bridging_RT_vol_in) from members of other groups (compare the Beta of .45 vs. -.10). This means that individuals that primarily interact with members of their own group harm their chances of obtaining retweets from members of other groups.

For bridging social capital, the results of the analysis lead to similar observations: The more a person is able to obtain mentions from members of other groups (LN_bridging_AT_vol_in), the less these individuals are retweeted by members of their own group (compare Beta of .52 vs. -.11). Additionally, the more individuals are able to garner attention from members of other groups (LN_bridging_FF_vol_in), the less retweets they obtain from members of their own group (compare Beta of .25 vs. -.08). This confirms that primarily appealing to members outside of the group harms an individual's chances of obtaining retweets from members of their own group. Hypotheses H6 and H8, which postulated that the creation of the opposite social capital on an individual level would harm information diffusion, were thus accepted.

7.5.3 Conclusion

This chapter evaluated 15 hypotheses (See table 7 in the appendix for a description of the hypotheses and indicators) concerning the effects of the structural, relational and cognitive dimensions of bonding and bridging social capital for groups and individuals. This chapter offered a descriptive analysis of the indicators for each social capital dimension of each hypothesis, and verified their significance using multivariate regression methods. Ten out of the fifteen hypotheses that were formulated regarding the influence of social capital on information diffusion hypotheses were confirmed by the statistical results. The results showed that, in Twitter, the structural and cognitive dimensions of social capital positively affect information diffusion, whereas the relational dimension does not. Additionally, the statistical results of the path analysis showed the negative cross effects of different types of social capital on information diffusion.

8 Conclusion

8.1 Summary of the results

This dissertation has investigated information diffusion in interest-based social networks in Twitter and determined the effect of social capital acquired in such networks on the amount of information diffused within them. The findings of this study show that the accumulation of bridging and bonding social capital at group and individual levels leads to better information diffusion. The results also confirmed that the cognitive, structural and relational dimensions of social capital are present at the group and individual levels. The next sections will summarize their influence and cross effects on information diffusion on a group and individual level.

The influence of group bonding social capital on information diffusion

Regarding the positive influence of group bonding social capital on the number of retweets that the group exchanges as a whole, it was shown that the more bonding group social capital is created by the group, the more retweets diffuse through the corresponding network. This confirms the hypotheses constructed around RQ 1. The final model 1.6 (see table 7.2 in section 7.1.2) of group bonding social capital, which contains four variables, was able to explain 50% of the information diffusion within the group. The results show that groups sampled according to the cognitive dimension of social capital share explicit ties that are captured in the structural dimension of social capital. The analysis confirmed hypothesis 1a, showing that in the structural dimension, a higher density, more clustering within the interactional network and a shorter path length between actors, lead to greater information diffusion within the group. In terms of the relational dimension, hypothesis 1b could not be confirmed - on the contrary, the results show that *less* reciprocity in the attention and interaction network leads to more information diffusion within the group. The broadness of the interest category had no significant effect on the internal information diffusion. In this investigation, the aim was to assess the influence of group bonding social capital on information diffusion within the group. The results of this study show that the explicit ties between individuals, established due to the cognitive dimension of social capital, are responsible for the majority of information diffusion within the group. Additionally, the negative effect of reciprocity on information diffusion shows that the process is dominated by hierarchical relations between members, which will be described in the paragraph below.

The influence of individual bonding social capital on information diffusion

The hypotheses of RQ2, which posited that individual bonding social capital has a positive influence on the amount of retweets an individual can obtain from group members, were also verified. The final model 2.6 of individual bonding social capital (see table 7.4 in section 7.2.2) was able to explain 46% of the information diffusion for an individual using only 5 variables. Hypothesis 2a was confirmed, showing that, in the structural dimension, more network centrality in the attention and interaction network allowed individuals to receive more retweets from members of their own group. Additionally, the more an individual is followed by central actors, as indicated by the PageRank metric, the more that person is able to obtain retweets. These findings confirmed the hierarchical relationship between group members, also known as opinion leader behavior in interest-based social networks. In terms of the relational dimension, hypothesis 2b was rejected. The results showed a significant negative effect of reciprocity in the attention network, indicating that "following back" does not automatically lead to more retweets once the centrality of the actors is taken into account. Additionally neither reciprocal interactions nor a strong average tie-strength to other actors had a strong enough effect on the number of received retweets. Hypothesis 2c was confirmed, showing that the more an individual's interests are similar to the group's interests (as measured by the person's nominations for this interest), the more retweets that person will obtain from group members. Thus, a significant finding of this study is that only two dimensions of individual bonding social capital play a significant role on information diffusion within the group. While the structural dimension shows strong tendencies for opinion leaders dominating information diffusion inside the group, the overlap of interests between an individual and a group benefits that person since they will receive more retweets from group members. Since the relational dimension of social capital had no effect on the amount of received retweets, this study shows that Twitter indeed is not a network to deepen and manage existing social relationships.

The influence of group bridging social capital on information diffusion

During the course of this study, the hypotheses built around RQ3 were confirmed: there is indeed a positive influence of group bridging social capital on the amount of retweets a group is able to obtain from other groups. The final model 3.7 (see table 7.6 in section 7.3.2) of group bridging social capital was able to explain 67% of the retweets the group obtained from other groups, using only four variables. Hypothesis 3a, which relates to the structural dimension of bridging social capital, was also confirmed. It was shown that the more central the group became, as measured by the attention and interaction the group received as a whole, the more retweets it was able to obtain from other groups. The group's centrality, measured using the group's closeness, was the best predictor to explain how the group's information was diffused by other groups. Groups that held an overall central position in the topical network had higher chances of receiving retweets from other groups. However, the group's intermediary position in the group network was not significant, showing that the diffusion process is not

dominated by the brokerage of certain interest groups, but rather by their centrality in the ecosystem. Hypothesis 3b, which reveals the relational dimension of group bridging social capital, could not be confirmed. The regression results for the group network showed that the smaller the average tie strength in the attention network, the more retweets the group obtained. This indicates that hierarchical tendencies exist even at a group level, where central groups do not depend on the relational dimension of social capital to obtain retweets from other groups. Hypothesis 3c was confirmed, thus revealing a strong effect of the cognitive social capital dimension on information diffusion at the group level: The more diverse the group's cognitive dimension was, the more retweets it was able to obtain. The more members with a high number of interests the groups had, the more retweets these groups got from other groups. Here interest groups such as "politics", "news" and "magazines", had a significantly high proportion of members that had multiple interests. However, the overall broadness of the interest category had no significant effect on the internal information diffusion. The following conclusions concerning the impact of group bridging social capital on information diffusion can be drawn from this study: Overall, the structure of the attention and interactional network facilitates information diffusion from the core to the periphery, and information diffuses from group to group because these groups perceive each other as being similar. On one hand, groups that share highly similar cognitive interests, such as "energy", "climate change", "sustainability" and "ecology" for example, tend to exchange a great amount of information. On the other hand, the structural core of this ecosystem is populated by groups whose cognitive social capital dimension is perceived as highly diverse. Such groups, which are classified under keywords such as "news", "politics" or "tech", play a crucial role in the Twitter ecosystem since they are the powerhouses of the information diffusion process. Indeed, the information shared by these groups tends to be very diverse on the cognitive level and reaches a wide audience since it is retweeted by all other groups.

The influence of individual bridging social capital on information diffusion

During the course of this study, the hypotheses of RQ4 were confirmed, thus showing that individual bridging social capital positively influences the amount of retweets an individual can obtain from people in other groups. The final model 4.6 (see table 7.8 in section 7.4.2) of individual bridging social capital, which used only four variables, explained 59% of the retweets the individual got from people in other groups. Hypothesis 4a, which is related to the structural dimension of social capital, was confirmed. The results showed that the more central the individual was for members of other groups, the more retweets he received. However, the structural bridging social capital was due in great part to the actor's central position, as measured by the actor's indegree and centrality in the attention and interactional network, rather than his brokerage position in the global network, as measured by the betweenness of the actor. Hypothesis 4b expressed the relational dimension of social capital by postulating that tie strength would positively influence information diffusion. The findings led to the rejection of this hypothesis: The analysis showed a significant negative effect between reciprocity in the

attention network, and information diffusion. Additionally neither reciprocal interactions nor a higher average tie-strength to other members had an effect on the amount of retweets received from other group members. Hypothesis 4c explored the influence of the individual's cognitive social capital dimension on information diffusion; this study confirmed this hypothesis. Indeed, the bridging cognitive social capital, which was measured by the number of interests that an actor had, emerged as a reliable predictor of information diffusion. It thus appears clear that individual bridging social capital has an effect on information diffusion: Two dimensions of individual bridging social capital have an influence on the amount of retweets an individual is able to receive from members outside his group. Obtaining an overall central position in the overall network (that is, the structural dimension) has the highest positive influence on an individual's probability of getting retweets. However, potentially highly relevant brokerage network metrics such as betweenness were rendered insignificant, due to the high number of total edges in the network and low constraints on edge creation in Twitter. The positive effect of the relational social capital on information diffusion could not be confirmed. For users that have reached a high popularity in Twitter, the relational dimension of social capital is impossible to maintain due to the high number of social ties that they possess. Finally being perceived as cognitively diverse in one's interests is also important for getting retweets from members of other groups.

The cross effects of social capital on information diffusion

In order to explore the cross effects between bridging and bonding social capital on information diffusion, the final models for social capital were used to describe social capital and the outcome variables were both considered in a path model. Hypothesis 5 posited that group bonding social capital would have a negative influence on external information diffusion. The findings confirmed this hypothesis. The statistical evidence suggests that the higher the group's density in the interactional network, the less retweets it receives from other groups. A plausible interpretation of this result is that groups that are highly focused on themselves – that is, whose members strongly interact with each other, and thus build more bonding social capital - do so at the cost of becoming less interesting for other groups, thus decreasing their chances of becoming retweeted by other groups. Hypothesis 7 postulated that group bridging social capital would have a negative influence on internal information diffusion. This could not be confirmed by the findings as neither the cognitive, relational nor structural group bridging metrics had a significant effect on internal information diffusion. This means that groups that are appealing to other groups (and thus build group bridging social capital) do not harm their internal information diffusion. These cross effects were also confirmed at the individual level. Hypothesis 6, which posited that individual bonding social capital would have a negative influence on information diffusion, was confirmed during the course of this study. The results showed that the more individuals interacted with group members, the less the individual's information was retweeted by users external to the group. This confirms the finding that building up individual bonding social capital comes at the cost of being perceived as a less

		Bonding		Bridging	
		Group	Individual	Group	Individual
Positive influence	Cognitive	-	✓	✓	✓
	Relational	✗	✗	✗	✗
	Structural	✓	✓	✓	✓
Negative influence	Internal Diff.	-		✗	✓
	External Diff.	✓	✓	-	

Table 8.1: Overview of the acceptance (✓) and rejection(✗) of hypotheses on different dimensions of social capital and their influence on information diffusion.

interesting person to retweet from the point of view of outsiders to the group. This means that to some extent, it is harder for opinion leaders of a certain interest group to appeal to members of other groups, or simply that their opinion leadership is limited to their specific interest group. Hypothesis 8, which posited that individual bridging social capital would have a negative influence on information diffusion, was also confirmed by the findings. Indeed, the amount of incoming interactions and attention from members outside of their group had a strong negative effect on the amount of retweets an individual obtained from members within their group. One can assume that individuals, who are able to acquire bridging social capital beyond their group, do so by significantly decreasing their chances of receiving retweets from members inside their group. This means that people that are structurally perceived as brokers have to pay a certain price, that is the price of getting less retweets from group members. Yet, being perceived as highly cognitively diverse slightly improves the chances of receiving retweets from group members.

Summary of results

This thesis has introduced a framework for social capital in Twitter that provides measurements to capture the structural, relational and cognitive dimensions of bonding and bridging social capital for groups and individuals. Ten out of the fifteen hypotheses that were formulated regarding the influence of social capital on information diffusion hypotheses were confirmed by this study (see table 8.1 and table 7 in the appendix for a description of hypotheses and indicators). It could be shown that in Twitter the structural and cognitive dimension of social capital positively affect information diffusion, while the relational dimension does not. Additionally the thesis showed the negative cross effects of different types of social capital on information diffusion. The section below will now highlight the theoretical, methodological and practical contributions and implications of the research findings.

8.2 Contribution

8.2.1 Theoretical contribution

The theoretical contribution of this work is the creation of a social capital framework for Twitter that can be used to predict the information diffusion process. This framework is based on three theoretical pillars: a) the insights gained from social capital theory and the research on online or virtual social capital, b) the insights gained from the information diffusion concepts and c) the specific insights relating to information diffusion in Twitter.

The systematic literature review on social capital has provided important insights into the operationalization of social capital in interest-based social networks such as Twitter. This work has contrasted the different cultural and structural views of social capital and introduced Nahapiet's multidimensional view, which takes into account the structural, relational and cognitive dimensions of social capital. In order to measure social capital, this study has distinguished three different components: the unit of analysis (group vs. individual), the type of social capital (bridging vs. bonding), and its source (offline vs. online). The literature review summarized the attempts to measure online social capital and highlighted these measurements lack of ability to use existing data traces in online social networks.

In order to form hypotheses on how social capital affects information diffusion, this work reviewed the literature on information diffusion. One of the significant findings to emerge from this review is that the main theories and concepts of innovation diffusion and information can be connected to the social capital theory. The review highlighted how information diffusion in groups can be explained by various structural factors (e.g., opinion leaders), relational factors (e.g., strong ties) and cognitive factors (e.g., shared norms). It also discussed the current insights on information diffusion in Twitter and categorized them according to the structural, relational and cognitive dimensions of social capital. It has shown that researchers have proposed various structural network metrics, and analyzed the relational and cognitive components of information diffusion, yet none of the literature integrates these insights into one common theoretical framework.

In order to fill this gap, this dissertation conceptualized a social capital framework for Twitter that makes use of the publicly available behavioral data traces. This framework offers conceptualizations of the structural, relational and cognitive dimensions of social capital. Structural and relational dimensions cannot be considered as attributes of a person, but rather can only be considered in relation to others, so the three dimensions of social capital were expressed as ties between individuals. The framework also took into consideration the behavioral insights gained from previous research into Twitter, and provided different network layers that were successfully combined with the social capital theory. Finally, it also explained, from a network theoretical perspective, how these types of network layers could relate to one another using a multiplex network approach. This constitutes the cornerstone of the social capital concept for Twitter because it helps determine which types of ties can be used to

derive the cognitive group concepts. Defining the notion groups helped conceptualize the difference between group and individual social capital and to define the bridging and bonding social capital concepts in Twitter. This allowed the creation of a set of hypotheses regarding the influence of social capital on information diffusion in Twitter. This set of hypotheses describes the influence of the structural, relational and cognitive dimensions of bridging and bonding social capital on information diffusion at a group level (H1a, H1b, H3a, H3b, H3c) and at an individual level (H2a, H2b, H2c, H4a, H4b, H4c). It also takes into account the interdependencies between bonding and bridging capital on both levels (H5, H6, H7, H8). Thanks to the resulting framework and its operationalization, the hypotheses concerning the influence of social capital on information diffusion in interest-based social networks were addressed at a fine-grained level in a way that previous studies were unable to do.

8.2.2 Methodological contribution

The methodological contribution of this work is threefold: first, it provides an operationalization of bonding and bridging social capital for the structural, relational and cognitive dimensions of groups and individuals, using the available data traces on Twitter and established network measures. The second contribution is the introduction of a method that helps to create a "folksontology" of users' interests, based on both the raw unstructured Twitter data (folksonomy) and on the ordered preexisting dictionaries of user's interests (ontology) online. The third contribution is the development of a network sampling method that allows for the sampling of interest-based networks on Twitter. These three methodological contributions will be detailed below.

The first methodological contribution is the operationalization of the bridging and bonding social capital framework for Twitter. This was done by providing exact measures for the structural, relational and cognitive components of bonding and bridging social capital, at the individual and group level. In order to operationalize the structural component, existing network measures for entire networks and for individual actors were introduced as measures of structural social capital. Regarding its relational dimension, existing network measures expressing tie strength and reciprocity were used. In order to measure the cognitive dimension of social capital, an own approach was developed. This approach is not based on the tweets that the user writes, yet, by using Twitter lists, it is not only able to determine how strongly an individual exhibits a certain interest but it allows the capturing of multiple interests per user.

The second contribution of this work is the development of a method to create a folksontology of users' interests that is based on both the raw unstructured Twitter data (folksonomy) and on the preexisting online interest dictionaries (ontology). Using this method has three advantages in comparison to using only a predefined ontology: it collects Twitter users' publicly exhibited interests, it merges lexicographically similar keywords, and it represents Twitter users' interests by selecting the most commonly used keyword. Since the interests retrieved (represented by keywords) are merged with an ontology, this method can organize these keywords according

to a predefined hierarchy and thus allows user interests to be expressed at the same level of granularity.

The third methodological contribution of this dissertation lies in the creation of a sampling method that allows for the sampling of Twitter users with common interests. The proposed seed-based method has the following advantages in comparison to other methods: it can take a large sample of candidates into consideration (who represent certain specific interests), it produces reliable, stable results independent of the initial seeds, it merges keywords that represent the same interest group and, allows the final candidates to change their original group assignment, which results in groups that strongly share the cognitive dimension of interest. The result is a highly detailed network of over 166 interest groups that provides a semantic map of user interests in Twitter and is used as the basis for all further analyses. The resulting map of user interests has - to the knowledge of the author - never been compiled at such a high level of detail[1].

8.2.3 Practical contribution

The present study makes several practical contributions to a number of domains, in ways that will be described below.

The practical relevance to community management

The group bonding measures that have been introduced can be used for community management or brand management to assess the value or vitality of a given community. Metrics, relying on the number of fans or followers of an account, for example, often do not reveal particular insights into the activity or engagement of a community since fans or followers can simply be bought online. Yet many practitioners today use their Facebook likes or Twitter followers as key performance indicators of their activities, even though they know that a high number of fans only leads to meager conversions if those fans do not share the same cognitive dimension as that of the company. Instead, measuring the structural and cognitive bonding social capital of such groups would allow community managers to really assess the value of their community. They can observe the potential deficits of a certain social capital dimension in their community and plan interventions to bolster that dimension. Before and after such an intervention, the bonding social capital metric might be used to measure the potential progress of such community-building activities. The group bridging social capital measures can additionally be used to assess the community's overall position in the Twitter network, and thus to predict the spread of information outside of the community. The results of this work show that groups with a high bridging social capital will also excel at receiving retweets from other interest groups. This

[1] Up until now, only five to ten interest groups have been studied by researchers (Wu et al., 2011; Choudhury et al., 2010).

can be beneficial if the interest community wants to promote issues such as sustainability or healthy living.

The practical relevance to politicians and governments

The descriptive analysis of group bridging social capital revealed that political interests are at the very core of users' interest networks. This means that despite the fear of polarized discussions and the emergence of "echo chambers", the constantly changing issues that are discussed in politics help this group maintain high cognitive bridging social capital. Moreover, this work has shown that cognitive bridging social capital is a relevant dimension of group bridging social capital and leads to information diffusion. This might explain why other interest groups are interested in political groups and why they are willing to retweet their information. This means that political communication via social media has a bright future: Twitter users seem to actually listen to what politicians have to say and frequently retweet this information. Additionally, group bridging social capital does not seem to harm communication within the group (see results of H7). However, this does not mean that the echo chamber hypothesis is rendered invalid: because of partisan affiliation and homophily, politicians might still prioritize communication with members of their own group rather than members outside the group (Plotkowiak et al., 2010). If this is the case, they should be aware that according to the results (H1, H8) of this thesis, this will only increase the chances of their information staying stuck inside the interest group, thus possibly harming their chances of receiving retweets from other groups.

Furthermore, the method that was used to sample user networks that represent certain interests can be used by politicians or governments as a radar. In this case, governments and politicians can "tune into" the most important representatives of different interest areas and directly capture users' reactions to their activities, according to the various interest groups. In this way, the representatives become social media signals for politicians or governments. Such a practice might be useful if politicians want to determine how a certain intervention on, for example, energy reforms will be perceived by car drivers vs. cyclists. Of course, this approach cannot replace representative polls, but allows for a near real-time capture of opinions and sentiments.

The practical relevance to journalists and news organizations

Since Twitter is considered a hybrid of a social network and a news media organization (Kwak et al., 2010), the theoretical implications of this study and the methods used have major practical implications on the role of journalists in the new sourcing, processing and dissemination processes that will be described below and have been discussed in greater detail in (Plotkowiak et al., 2012).

Sourcing processes: Interest-based networks are becoming a source of information that can be processed by the emerging "citizen journalism 2.0". In this context, journalists and news

agencies can use the method of sampling networks of users with certain interests in order to find entire communities of interest. These can then be used as social media sources in times of crisis or when new issues are emerging (e.g., the uprisings in Iran, Iraq or Syria or the financial crisis). The users appear as valuable real-time social media sources that represent different interest groups with unfiltered opinions – which are often hard to come by. Journalists can thus deeply embed themselves in their areas of interest and expertise in order to obtain the most valuable information as quickly as possible.

Brokerage processes: This work has shown that building up individual bridging social capital helps individuals obtain retweets from other groups (H4). Since the interest-based Twitter ecosystem allows news agencies and journalists to connect to various interest communities by bonding with the sources while potentially gaining followers from very different interest groups, this gives them the chance to establish themselves as topic brokers between interest groups. The journalistic routine can thus be potentially seen as a two-sided service: first, journalists can offer the service of lending their voice to special interest groups by simply retweeting or forwarding selected material to a much broader audience, and second, they can provide their audience with aggregated insights from special interest groups that are difficult for an unskilled user to understand. Journalists can thus directly become their own brand, by forwarding selected information from sources on Twitter to their audience. Additionally, the descriptive analysis of group bridging social capital has shown that the "news", "journalists" and "magazines" groups are highly effective at obtaining retweets from other groups, suggesting their central role as multipliers or news powerhouses. In fact, the results of this thesis suggest that the discussion on new gatekeeping processes in social media will need embrace new roles of cognitive brokers who can transform information obtained from a specific interest group in such a way that it appeals to members of other interest groups.

Dissemination processes: This work has shown that it is beneficial to build up social capital in the interest group to be retweeted (H2). Thus, in order to reach a wide audience and to guarantee the news is relevant to them, news corporations and journalists have to establish their Twitter accounts in terms of their interest area – that is, embed their account into a specific interest group. This work has shown that the more the account develops individual bonding social capital, the higher the chances that it will receive retweets. For a journalist, developing individual bonding social capital means sharing the same cognitive dimension as his or her audience (H2c), and becoming structurally central by gaining followers that are relevant to his or her interest groups (H2a). This means that, every day, a journalist needs to tend to his or her reputation as an aggregator or curator of news for this interest group.

Established newspapers and news agencies will need to think about how these new sourcing, brokering and dissemination processes in social media can lead to new business models that benefit the audience, journalists and corporations.

Practical relevance for marketers and advertisers

Since Twitter is also a major outlet for brands, corporations, marketers and advertisers, which are continuously developing strategies and processes to effectively partner with their consumers, the results of this thesis have important implications for these entities. In fact, the results of this dissertation highlight that, from a marketing perspective, it might not be too far-fetched to consider a user's interests as the most valuable resource an online social network can reveal — interests that users inherently express through the relations they establish in their network. This form of detailed knowledge concerning the consumers of a brand has never been possible before today, and in some cases, is the best-kept secret of these networks, such as Facebook[2]. As a potential application of this work, the seed-based sampling method described above, can easily be adapted by marketers or advertisers to target[3] highly specific networks of people whose interests might correspond to the product or service they are offering.

Acknowledging that a shared cognitive dimension leads to shared structural social capital dimensions implies that the target group cannot simply be seen as a disconnected audience, but instead as a networked interest group whose connections really matter to its members. Marketers can take advantage of the results of this study, which have shown that individuals with high individual bonding social capital (containing high cognitive and structural social capital) excel at spreading information in such interest groups (H2). Marketers can also take advantage of the results showing that people with high bridging social capital are adept at spreading information outside of an interest group (H4). Indeed, seeking cooperation with opinion leaders from different interest groups in order to have them endorse a certain product might be beneficial to both the user and the company. This sort of endorsement is different from how celebrity endorsement[4] is practiced today. The current strategies of such product endorsements on Twitter are often based on the number of followers alone, which leads companies to select users that seem influential but are actually only marginally (Gayo-Avello, 2010) relevant to the interest groups the product is targeted at.

Finally, at the group level, this study offers two potential ways for viral marketing campaigns to advertise a new product in an interest community. First, this dissertation has provided additional evidence[5] that internal group information diffusion is influenced by the structural and cognitive dimensions of the group (H1). Therefore, campaigns that take all of these dimensions into account may culminate in great success. Bauer et al. (2005) envisioned this approach under the name of 'virtual communities' noting that "[...] communities that provide social capital are a key instrument to establish satisfaction, trust and commitment within a

[2] http://www.adweek.com/news/technology/facebook-giving-some-brands-sneak-peek-fans-other-likes-144486

[3] http://advertising.twitter.com/2012/08/interest-targeting-broaden-your-reach.html

[4] http://www.usatoday.com/tech/news/story/2011-11-03/celebrity-twitter-endorsements/51058228/1

[5] Choudhury notes that marketers "could benefit from considering specific sets of users on a social space who are homophilous with respect to their interest"[p.3] (2010).

business relationship."[p.91]. The capacity to measure the social capital of their own brand communities will thus enable marketers to establish satisfaction, trust and commitment in the networked relationship with their audience. Community managers of brands in social media are struggling to understand how to use the emergent behavioral traces and the actual interests and preferences of users in such fan pages or groups. These managers can use the concept of social capital as a valuable tool to better measure and manage their brand community and, as a result, help more effectively and efficiently disseminate their message to their users, or generate more conversions. Furthermore, the results of the study on group bridging social capital (H3) have shown that the network of user interest groups is not actually fragmented into different areas, but rather offers a highly interwoven network of interests. Thus, advertisers can determine the right target group, by taking into consideration the neighboring groups of a particular interest group; indeed, the chances are high that their message or product will also be relevant to these people.

These potential interest-based improvements to social media marketing can lead to higher response rates and lower losses due to selective advertising, as well as higher incomes for advertisers and relevant advertising for customers. Combining demographic data with network-based and interest-based data will give marketers the ability to combine both worlds and, in doing so, achieve even higher relevance: for instance, they would be able to promote a high-priced running shoe to *well-connected* users who have the *right household income* and share the *interest in running*.

The practical relevance to knowledge exchange systems

This work also has practical relevance to the knowledge exchange systems that exist within organizations. The implications stem directly from the analysis of the interdependency of the group bridging and bonding social capital analysis of the interest groups. On one hand, the results showed that more group bonding social capital led to more intra-group knowledge exchange but harms the intra-group knowledge exchange (H5); on the other hand, the results showed that more group bridging social capital led to more inter-group knowledge exchange, without harming the intra-group knowledge exchange (H7).

These observations thus create an opportunity for the development of new knowledge exchange and management systems for organizations that take advantage of these mechanisms: on one hand, such systems should endorse information sharing inside departments, but on the other, they should also create new processes or incentives to exchange information between departments. The combination of sociological concepts and computer science will breed new forms of knowledge exchange and management systems able to support these processes: "while computer science has not yet embraced the social capital concept, there are many computer applications that have the potential to augment social capital of its users"[p.89] (Huysman and Wulf, 2005).

Therefore, it seems plausible that the next generation knowledge exchange systems will make use of the social capital concept by taking into account the cognitive, structural and relational social capital dimensions in order to enhance knowledge exchange. Such systems have already been proposed by Husymann (2005) who postulated that "communities with high social capital will be more inclined to [...] share knowledge than ones with low social capital. Future research into the various dimensions of social capital will enhance our understanding of how technology can support communities."[p.91]. The findings concerning the different dimensions of social capital in interest networks suggest how next generation systems might work, and reveal that they might not actually be that different from current social media systems, in that they might copy the same roles that individuals play in Twitter. It can also be boldly suggested that new roles in organizations might appear: roles where people need to facilitate inter-group knowledge exchange, in a similar way to what journalists have been doing for decades.

Twitter and the filter bubble

The findings of this thesis can contribute to the recent discussion on the filter bubble lead by Eli Pariser (2011). He describes the filter bubble as the phenomenon where search engines and social networks show different results to different people according to their political orientation. A politically-left user might search for "BP" in Google and see results related to the oil spillage in Mexico, whereas the top search results of a conservative user might be the investor pages of BP. A similar phenomenon can be seen in social networks such as Facebook, where an algorithm filters the News Feed in order to show only relevant posts to the users. The problem is that users don't get to decide what passes through the filter and what doesn't: Yahoo, Google and Facebook do. More importantly, users don't see what gets edited out, so they don't even know what they are missing. In contrast, the news feed in Twitter, which displays the tweets of people users' follow, is not filtered. This might lead users to naively believe that Twitter allows them to overcome or "pop" the filter bubble. This is not actually the case. Indeed, this work has shown that the biggest concern for users is not the filter bubble created by others, but rather the filter bubble users create themselves every day, simply because there is a human limitation to the amount of information a person can process. In Twitter, when users follow their homophile instinct and consume only information from like-minded people, they are exposed uniquely to information that fits their worldview. The dangers of this are epitomized by Eli Pariser's assertion[6] that "a world constructed from the familiar is a world in which there's nothing to learn [since this is] invisible auto-propaganda, indoctrinating us with our own ideas". Indeed, this perfectly applies to Twitter, too. It is therefore maybe time for people to abandon the naive idea that content is unfiltered in Twitter, be less worried about the filter bubbles created by algorithms, and focus more on the filter bubble they entangle themselves in on social media platforms. Indeed, we should actively seek to expose ourselves to other interests and other groups by acting more like brokers on Twitter, and help spread information from different

[6] see `http://www.economist.com/node/18894910?story_id=18894910`

interest groups to our own followers. In this way, we expose not only ourselves to divergent points of view; we also expose all of our followers to new diverse material that stimulates critical thinking.

8.3 Limitations and outlook

The findings in this work are subject to a number of limitations that are largely tied to the focus of this study. Indeed, the goal of this dissertation was to create a social capital framework for Twitter and show that social capital had an effect on information diffusion. This meant that only the phenomena explained by the social capital theory could be taken into account, while other phenomena that explain information diffusion had to be omitted. However, these limitations also provide opportunities for future research into how social capital affects information diffusion.

For this study, the size of the interest groups was limited: although the group size of interest groups was chosen according to theoretical considerations (such as the Dunbar Number and empirical results on Twitter), analyzing groups with more members would have made the already large data corpus (see section 6.6) even harder to process. It is nonetheless interesting to explore what effect group sizes of a magnitude of 10 or 100 might have on the results. It is conceivable that higher group sizes would result in less and less cognitively similar groups, which might be reflected in lower attention or interaction among members. Closely related to this is the question of how higher group sizes could affect the possibility of users being allocated to different groups than their initial one. The number of interest groups that have been analyzed is also an interesting factor to take into account. While the final 166 interest groups led to significant statistical results, it would be interesting to know how a much higher number of groups would affect the results. Indeed, an increase in the number of interest groups included in the study would have led to an increase in the specificity of these interest groups, since the sampling followed a top-down approach that started with the most prominent groups. However, the methodological chapter showed that for extremely specific interest groups, it would become close to impossible to collect 100 members or more, thus making the groups harder to compare. Still, from a practical perspective, marketers looking for commercial applications of these findings might require systems than can analyze a greater number of interests than just 166. Indeed, such systems would be able to capture subtle nuances in user interests at the product level.

Limitations also stem from exploring only the selection process in interest-based social networks and the assumption that interests lead to social ties, which then give birth to interactions, and thus support information diffusion. It might have been interesting to also study the influence processes in networks, which lead to the adoption of certain similarities, and thus might change the interests of a user. The first reason why this dissertation has refrained from researching these processes is that they do not fit in with the social capital theory, which

predicts that networks lead to the achievement of certain goals. Indeed, the adoption of an interest (the cognitive dimension) is a network outcome, not a goal. The second reason is that influence processes are much less likely to be found (Aral et al., 2009) in networks than selection processes. The third reason is that the research questions on influence processes can only be answered empirically in longitudinal analyzes of the network, which, at the time of writing, is not statistically feasible for networks of this size. However, disentangling the selection and influence processes in social networks is a very promising area of research that can potentially be answered by experimental research designs. Such research designs would additionally allow researchers to analyze the interactions between the structural, relational and cognitive dimensions of social capital offer insights how these dimensions harm or reinforce each other.

The next limitation deals with the focus on certain attributes: A user's interest is only one of many attributes that can be collected on social networks. The literature review has shown that there are a multitude of other attributes (e.g., political orientation, geographical location or language) that might lead to homophilious networks, whereas in this work, only the user's interests were investigated. Thus, a very important question that should be answered in the future is which combination of the various attributes of a user best explains tie creation in online social networks. This would allow researchers to determine which set of attributes would best predict interpersonal information diffusion.

Ultimately, the study did not include a content analysis of the messages exchanged or an analysis of their structure. There is an entire research field in communication science that focuses on the news value of items (Lippmann, 1922; Ruhrmann, 2003), and it could be used to explain why certain messages lead to retweets and others do not. The outcomes of these research questions might complement the results of this study. It is very plausible that the cognitive dimension of a user, that is the manner in which a certain user fits in with the interest group, is highly reflected in the messages he or she shares with the community. Therefore, future research might explore in greater detail the different types of messages that travel through these interest networks, and how these can be manipulated in order to achieve higher virality within the target groups. So far, the results have shown that the more bonding or bridging social capital an individual acquires in these networks, the more his messages will be retweeted. This indirectly confirms Valente's prediction that the "messenger is the message" (2010). However, it would be interesting to see what the messenger's messages actually look like. This might help determine how well the messages of group opinion leaders are aligned with the group's interest. It might also reveal how brokers between interest groups adapt their messages in such a way that they appeal to multiple audiences. Further research into these questions might reveal if messages differ in content and structure across different interest groups. The insights from such a detailed analysis of the users' messages might be used to create better campaigns for marketers, better political messages for governments or politicians, and complement the research on news value theory.

Mapping users' interests

Finally, a very interesting research direction, beyond the study of information diffusion, could be the creation of user interest maps. A map of interests, which is based on the aggregation of interaction between many interest groups, can be used as a tool to extract a user's interests solely from his social ties. The logic behind this is simple: If a user is interested in a given topic, over time there is a high probability that he will create explicit ties to at least one of the 100 members that represent this interest group. This means that we do not necessarily have to look at the lists the user is listed on or study his messages in order to find out what interests he might represent, but rather that his social ties on Twitter can be used to determine these interests. A histogram of each and every one of the user's interests can be created by tracing his social ties back to the members of the groups they represent: for example, a user might follow 5 people from the "sustainability" group, 3 people from the "automotive" group, and 2 people from the "politics_news" group. We can infer from these groups that his actual interests might be distributed in a similar way. So when Twitter asks you to "follow your interests", to a certain degree this actually means:

> Your interests are defined by whom you follow.

This simple conclusion paves the way for future research and very interesting applications.

Solving the cold start problem[7]: When a user starts to use a new application (e.g., visits a shopping website for the first time, or turns on his new Smart TV), a problem can appear: the application might be incapable of recommending anything to the user because it simply does not know what the user likes. In this case, researchers speak of a cold start problem. However, if that user can be found on Twitter, such applications might analyze the publicly expressed ties on his Twitter account and, in combination with the map of interests developed, would be able to infer the user's basic interests simply from the ties that he formed on Twitter. This means that, in the best of cases, applications do not need to ask the user what his or her interests are, but rather automatically "know" them, with no effort required from the user. Such systems would create new types of TV watching experiences, where the TV shows the user the shows he might be interested in, and the user simply leans back and consumes his personalized stream.

Assembling automatic news streams: The second potential application would be the creation of new generations of news applications where the system automatically recommends news items that might interest the user. Such systems already partially exist[8], but they only take into account the network structure of the user's ego-network and so only partly capture the user's cognitive and structural dimensions. It would be useful to explore how a combination of the structural, relational and cognitive dimensions of social capital would create new generations of news filtering and aggregation systems. In these systems, news streams would automatically

[7] `http://en.wikipedia.org/wiki/Cold_start`

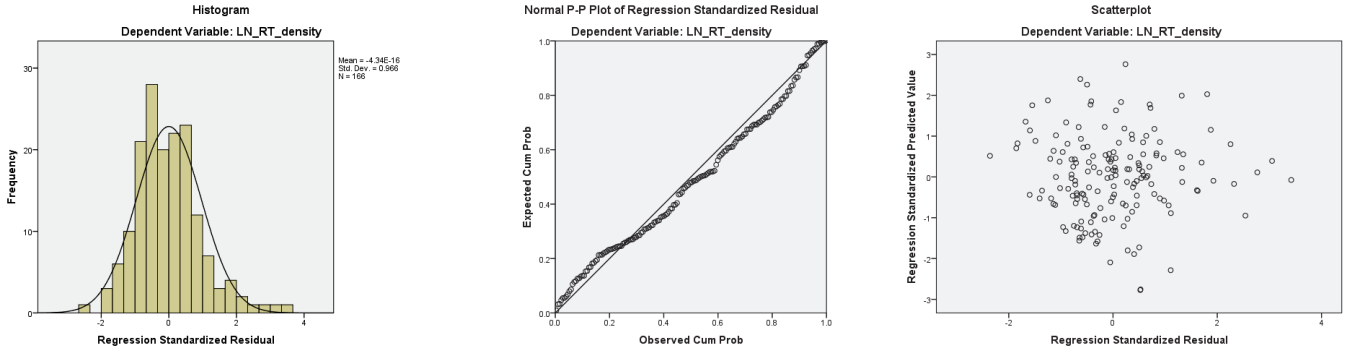
[8] See, e.g., `http://tweetedtimes.com` or `http://paper.li`

originate from interest-based social networks, while the social ties of the user reveal his interests.

Determining the interests of entire communities: The third practical application of the capacity to reveal a user's interests lies in providing aggregated overviews of entire groups of users. Fans of a certain brand, product or newspaper can thus be evaluated as a whole and reveal insights into their overall interests. As an example, the coffee brand Starbucks might find out what their followers like other than coffee and then try to serve those interests. Possessing such information can allow marketers to optimize their communication strategy and allow news organizations and journalists to provide more relevant news to their readers.

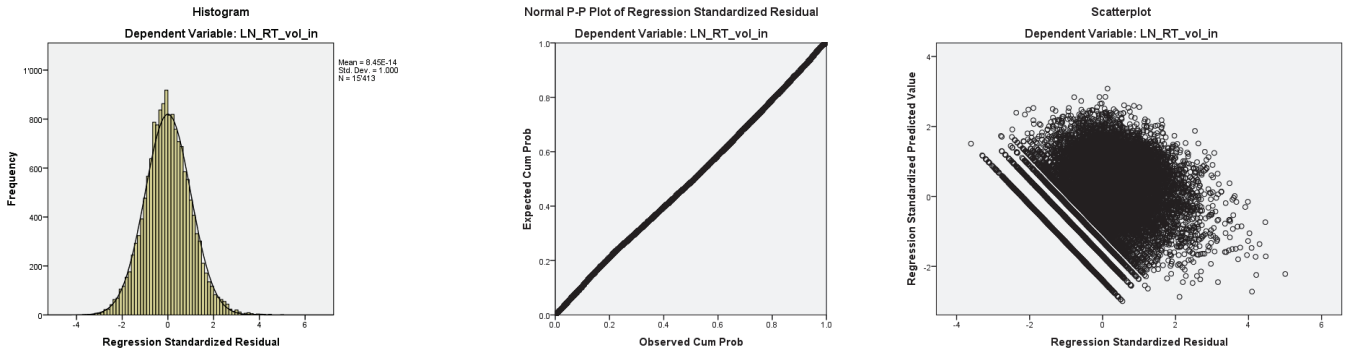
Appendix

Additional tables and results



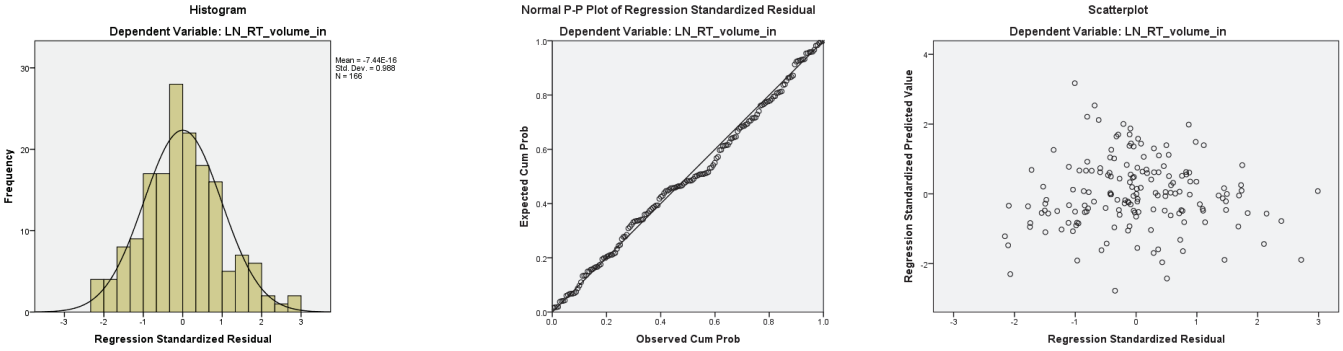
(a) Histogram of Residual **(b)** P-P Plot of Residual **(c)** Scatterplot of Residual

Figure A.1: Regression residuals of group bonding social capital



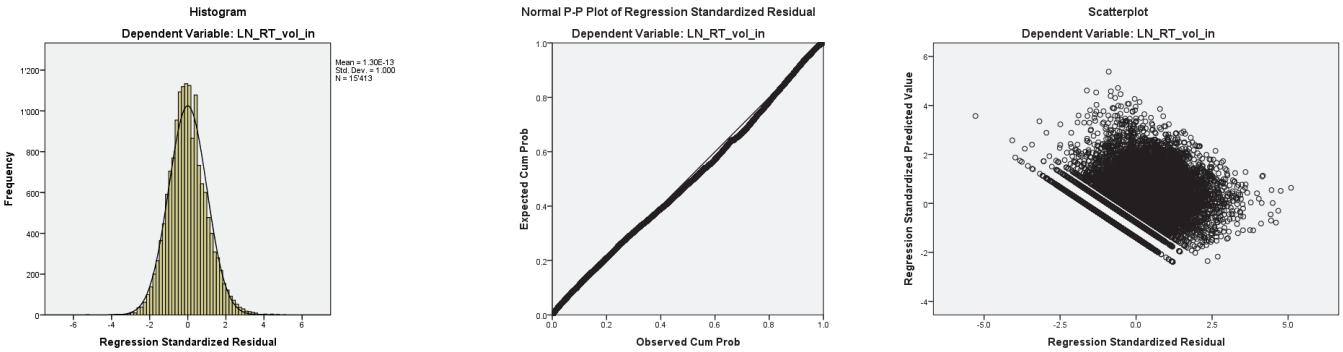
(a) Histogram of Residual **(b)** P-P Plot of Residual **(c)** Scatterplot of Residual

Figure A.2: Regression residuals of individual bonding social capital



(a) Histogram of Residual **(b)** P-P Plot of Residual **(c)** Scatterplot of Residual

Figure A.3: Regression residuals of group bridging social capital



(a) Histogram of Residual **(b)** P-P Plot of Residual **(c)** Scatterplot of Residual

Figure A.4: Regression residual of individual bridging social capital

Tests of Normality

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
RT_density	.07	166.00	.04	.95	166.00	.00
Member_count	.31	166.00	.00	.45	166.00	.00
FF_bin_density	.04	166.00	.20*	.99	166.00	.14
AT_density	.13	166.00	.00	.84	166.00	.00
FF_bin_clustering	.08	166.00	.01	.97	166.00	.00
AT_bin_clustering	.06	166.00	.20*	.99	166.00	.31
FF_bin_avg_path_length	.13	166.00	.00	.87	166.00	.00
AT_bin_avg_path_length	.13	166.00	.00	.90	166.00	.00
FF_reciprocity	.05	166.00	.20*	.99	166.00	.15
AT_reciprocity	.06	166.00	.20*	.99	166.00	.39
LN_AT_avg	.06	166.00	.20*	.99	166.00	.39

Table 2: Kolmogorov-Smirnov and Shapiro-Wilk tests of normality for group bonding IVs and DV

Tests of Normality

	Kolmogorov-Smirnov		
	Statistic	df	Sig.
RT_vol_in	0.29	15413.00	0.00
Ranking_in_interest	0.07	15413.00	0.00
FF_bin_in_deg	0.05	15413.00	0.00
AT_vol_in	0.35	15413.00	0.00
FF_bin_close	0.10	15413.00	0.00
AT_bin_close	0.09	15413.00	0.00
FF_bin_page	0.10	15413.00	0.00
AT_bin_page	0.24	15413.00	0.00
FF_rec	0.03	15413.00	0.00
AT_rec	0.11	15413.00	0.00
AT_avg	0.30	15413.00	0.00

Table 3: Kolmogorov-Smirnov tests of normality for individual bonding IVs and DV

Tests of Normality

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
RT_volume_in	.18	166.00	.00	.77	166.00	.00
Agg_number_of_interests	.17	166.00	.00	.83	166.00	.00
Member_count	.31	166.00	.00	.45	166.00	.00
FF_volume_in	.24	166.00	.00	.65	166.00	.00
AT_volume_in	.30	166.00	.00	.50	166.00	.00
FF_bin_betweenness	.30	166.00	.00	.52	166.00	.00
AT_bin_betweenness	.26	166.00	.00	.66	166.00	.00
FF_bin_closeness	.15	166.00	.00	.96	166.00	.00
AT_bin_closeness	.03	166.00	.20*	1.00	166.00	.95
FF_bin_pagerank	.27	166.00	.00	.57	166.00	.00
AT_bin_pagerank	.32	166.00	.00	.48	166.00	.00
FF_rec	.06	166.00	.20*	.98	166.00	.02
AT_rec	.08	166.00	.02	.96	166.00	.00
FF_avg	.08	166.00	.02	.96	166.00	.00
AT_avg	.11	166.00	.00	.91	166.00	.00

Table 4: Kolmogorov-Smirnov and Shapiro-Wilk tests of normality for group bridging IVs and DV

Tests of Normality

	Kolmogorov-Smirnov		
	Statistic	df	Sig.
RT_vol_in	.36	15413.00	0.00
Number_of_interests	.49	15413.00	0.00
FF_vol_in	.29	15413.00	0.00
AT_vol_in	.43	15413.00	0.00
FF_groups_in	.17	15413.00	0.00
AT_groups_in	.29	15413.00	0.00
FF_bin_betweenness	.45	15413.00	0.00
AT_bin_betweenness	.40	15413.00	0.00
FF_rec	.16	15413.00	0.00
AT_rec	.22	15413.00	0.00
AT_avg_tie_strength	.31	15413.00	0.00
AT_strength_centrality_in	.43	15413.00	0.00

Table 5: Kolmogorov-Smirnov test of normality for individual bridging IVs and DV

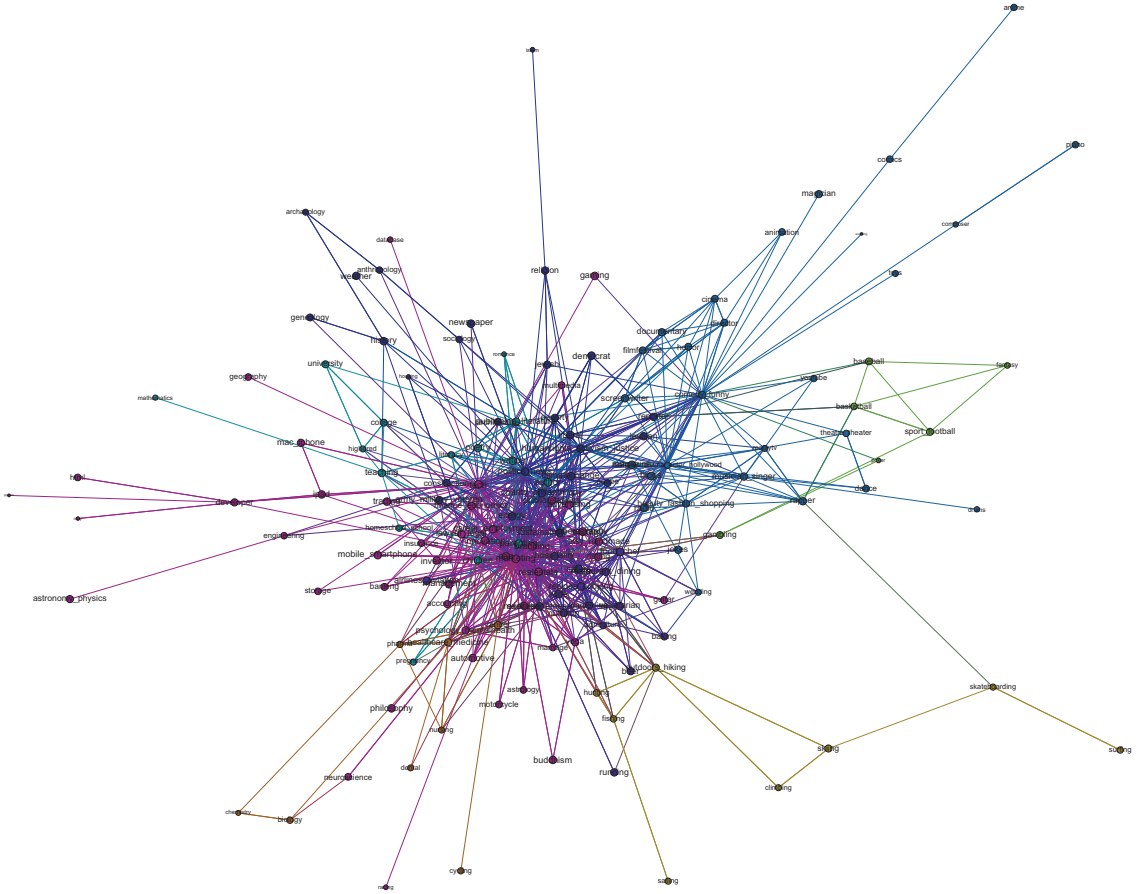


Figure A.5: Reduced version of the blockmodel of the FF network. Each node represents the network of 100 Twitter users for this group. For enhanced visibility edges with weight of < 300 have been removed; coloring acc. to modularity detection; size acc. to FF in-degree.

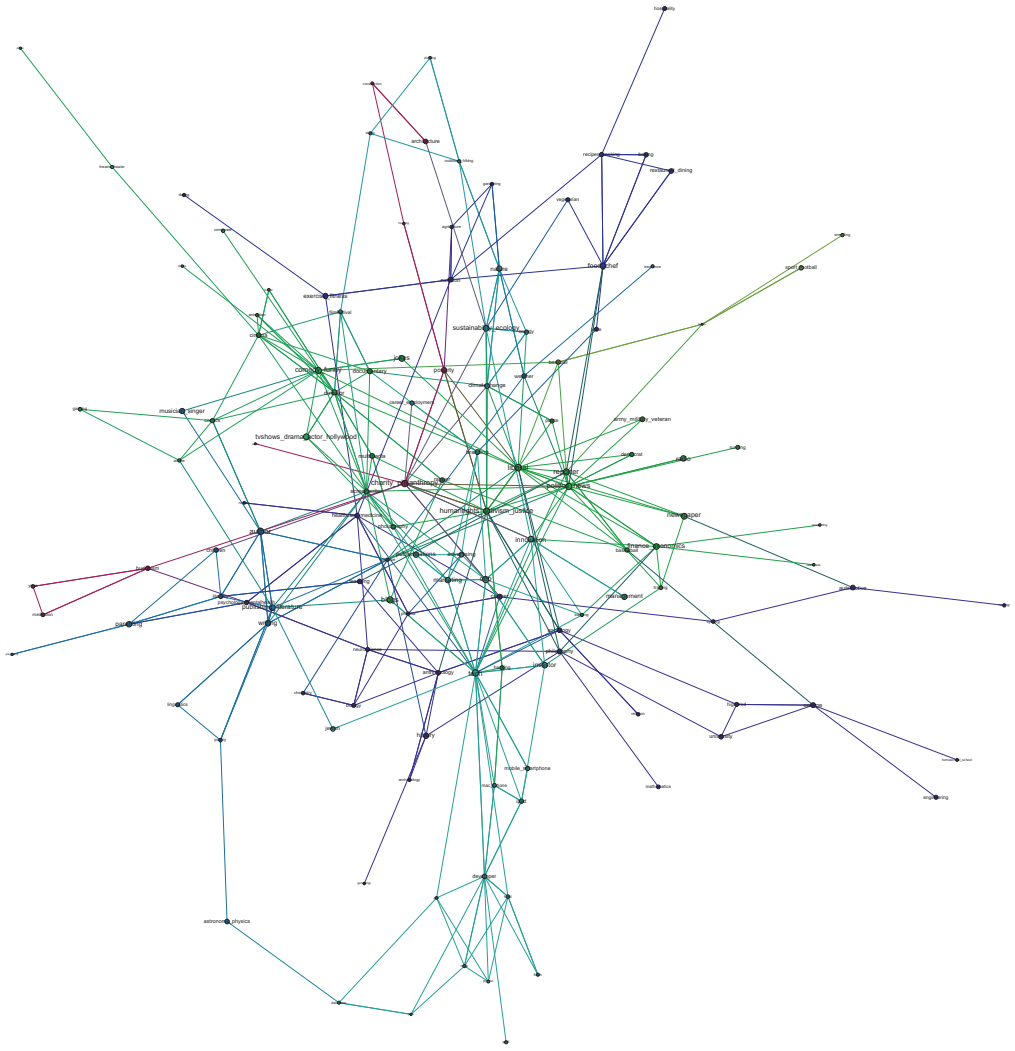


Figure A.6: Reduced version of the blockmodel of the RT network. Each node represents the network of 100 Twitter users for this group. For enhanced visibility edges with weight of < 100 have been removed; coloring acc. to modularity detection; size acc. to RT in-degree.

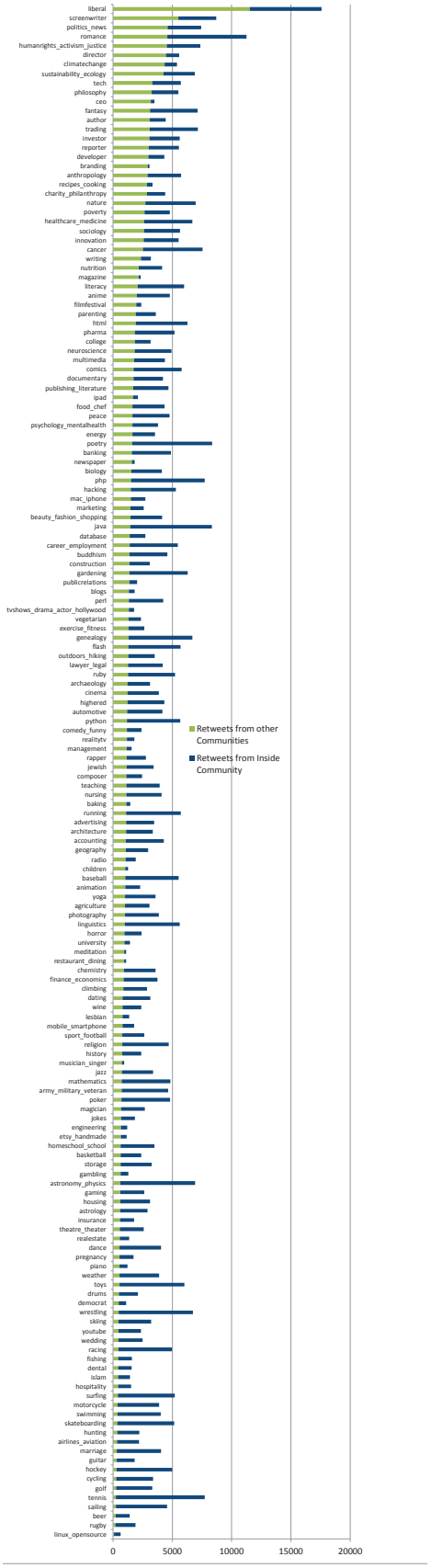


Figure A.7: Internal retweets disseminated inside group vs. retweets received from other groups

Figure A.8: Listings A-I

Figure A.9: Listings J-Z

Community name	Based on Keyword(s)	Number of lists (unique)	Member count (unique)	Average members per list	Min members per list	Max members per list
accounting	accounting	1100	3680	210.72	1	501
advertising	advertising	10720	22988	209.72	1	500
agriculture	agriculture	447	4215	164.57	1	481
airlines_aviation	airlines, aviation	8710	5964	80.16	1	408
animation	animation	596	4857	74.31	1	498
anime	anime	1017	13236	91.94	1	500
anthropology	anthropology	592	1300	69.06	1	178
archaeology	archaeology	610	1878	152.37	1	481
architecture	architecture	3211	14084	126.76	0	500
army_military_veteran	army, military, veteran	2947	9451	105.77	1	499
astrology	astrology	1239	1748	72.05	1	424
astronomy_physics	astronomy, physics	3570	5851	87.18	1	499
author	author	2532	40658	237.08	1	502
automotive	automotive	1689	5039	120.99	1	382
baking	baking	683	2302	135.19	1	450
banking	banking	133	1820	97.59	2	473
baseball	baseball	3363	11270	103.67	0	492
basketball	basketball	9497	6377	81.16	1	500
beauty_fashion_-shopping	beauty, fashion, shopping	25779	91348	127.50	1	565
beer	beer	107	4932	180.59	3	493
biology	biology	678	1563	54.23	1	276

blogs	blogs	1286	41108	132.34	1	501
branding	branding	1612	7301	391.04	0	498
buddhism	buddhism	907	2882	159.78	1	461
cancer	cancer	1568	6781	120.93	1	499
career_employment	career, employment	2913	12036	182.62	1	500
ceo	ceo	941	2738	52.01	1	500
charity_philanthropy	charity, philanthropy	2013	12034	125.18	0	485
chemistry	chemistry	248	1568	73.64	1	330
children	children	493	5024	105.43	1	492
cinema	cinema	742	11872	73.07	1	500
climatechange	climatechange	66	798	54.88	4	172
climbing	climbing	301	2410	81.21	1	383
college	college	681	10191	117.10	1	498
comedy_funny	comedy, funny	25319	25466	76.23	1	501
comics	comics	5843	15030	84.91	0	500
composer	composer	979	3607	184.53	2	500
construction	construction	759	6098	135.74	1	497
cycling	cycling	3530	15358	101.02	0	500
dance	dance	1535	14952	96.71	0	500
database	database	67	800	33.76	1	332
dating	dating	969	2627	104.34	1	466
democrat	democrat	1030	1731	80.91	1	391
dental	dental	984	4359	197.56	1	496
developer	developer	540	18075	133.58	1	502
director	director	296	3815	124.80	1	499
documentary	documentary	656	1121	96.27	1	367
drums	drums	508	598	44.60	1	122

energy	energy	1413	12582	105.65	1	501
engineering	engineering	231	3547	72.08	1	490
etsy_handmade	etsy, handmade	2870	21345	308.29	0	499
exercise_fitness	exercise, fitness	4571	28653	165.49	1	501
fantasy	fantasy	396	5763	70.52	1	371
filmfestival	filmfestival	336	977	120.83	5	212
finance_economics	finance, economics	7692	15430	85.17	1	499
fishing	fishing	552	5456	113.50	1	478
flash	flash	2022	3711	81.91	1	498
food_chef	food, chef	61025	44632	92.64	1	501
gambling	gambling	466	951	71.86	1	103
gaming	gaming	4647	16728	77.88	1	500
gardening	gardening	1286	6536	132.78	1	501
genealogy	genealogy	1165	1977	106.89	1	486
geography	geography	460	1395	122.57	2	401
golf	golf	17144	6817	66.12	1	497
guitar	guitar	822	5839	80.91	1	497
hacking	hacking	648	1782	63.73	1	158
healthcare_medicine	healthcare, medicine	4156	18701	152.27	1	500
highered	highered	612	6305	137.82	1	501
history	history	1307	9979	82.10	1	492
hockey	hockey	5064	14937	117.71	1	500
homeschool_school	homeschool, school	7079	11250	92.84	1	499
horror	horror	1025	4166	76.99	1	456
hospitality	hospitality	588	4324	85.02	1	378
housing	housing	793	1970	115.58	1	478

html	html	1369	2309	50.37	1	495
humanrights_-activism_justice	humanrights, activism, justice	2197	19746	109.97	1	501
hunting	hunting	465	2504	173.15	1	406
innovation	innovation	2672	14499	239.68	1	500
insurance	insurance	562	5325	108.56	1	501
investor	investor	1796	4079	90.91	1	498
ipad	ipad	123	2823	113.58	1	498
islam	islam	532	3228	73.93	1	462
java	java	2549	2312	38.05	1	469
jazz	jazz	1392	7422	177.05	1	494
jewish	jewish	548	2333	111.64	1	453
jokes	jokes	101	1347	72.12	1	497
lawyer_legal	lawyer, legal	1994	17664	174.70	1	499
lesbian	lesbian	674	2125	169.80	1	466
liberal	liberal	992	5334	191.01	1	495
linguistics	linguistics	315	1851	96.26	2	470
linux_opensource	linux, opensource	2097	8908	54.12	1	469
literacy	literacy	312	2955	78.25	1	500
mac_iphone	mac, iphone	4244	23853	68.75	1	498
magazine	magazine	3165	24106	75.71	1	500
magician	magician	663	1141	87.76	1	316
management	management	452	7405	124.56	1	500
marketing	marketing	31686	96980	216.19	1	501
marriage	marriage	324	1050	43.85	1	171
mathematics	mathematics	564	1666	111.91	1	483
meditation	meditation	1056	2322	104.58	1	330

mobile_smartphone	mobile, smartphone	2524	21393	92.80	1	503
motorcycle	motorcycle	479	4008	107.28	1	488
multimedia	multimedia	184	2802	70.21	1	500
musician_singer	musician, singer	20941	38673	90.59	1	501
nature	nature	995	12229	117.35	1	500
neuroscience	neuroscience	396	2258	190.99	1	479
newspaper	newspaper	179	3704	70.61	1	393
nursing	nursing	531	2407	81.18	1	239
nutrition	nutrition	939	9947	223.92	1	499
outdoors_hiking	outdoors, hiking	1399	11264	90.36	1	502
parenting	parenting	2821	13635	131.57	1	500
peace	peace	207	5385	81.68	1	491
perl	perl	2142	1228	57.65	1	496
pharma	pharma	1656	5722	87.29	1	496
philosophy	philosophy	235	4185	96.23	1	501
photography	photography	31298	17238	124.98	1	586
php	php	1616	4296	50.36	1	473
piano	piano	328	1536	58.50	1	234
poetry	poetry	846	7552	104.36	1	500
poker	poker	6839	10632	98.70	0	500
politics_news	politics, news	157453	47963	67.82	1	583
poverty	poverty	161	2447	65.37	1	498
pregnancy	pregnancy	224	1747	78.11	1	266
psychology_-mentalhealth	psychology, mentalhealth	1904	7569	125.78	1	500
publicrelations	publicrelations	2198	2642	469.46	1	500
publishing_literature	publishing, literature	3280	22559	131.59	0	501

python	python	3957	1572	54.55	1	474
racing	racing	1691	7508	106.15	1	466
radio	radio	4915	32164	78.74	1	501
rapper	rapper	232	4565	137.70	1	499
realestate	realestate	1847	11333	228.17	1	500
realitytv	realitytv	565	1618	143.31	1	500
recipes_cooking	recipes, cooking	3206	11231	238.17	1	493
religion	religion	1862	7117	108.97	1	500
reporter	reporter	184	7501	90.41	1	483
restaurant_dining	restaurant, dining	1459	24293	146.63	1	501
romance	romance	238	2348	99.89	1	306
ruby	ruby	9006	5298	75.80	1	449
rugby	rugby	1412	5948	76.05	1	498
running	running	2090	11091	72.20	1	499
sailing	sailing	1217	1762	65.58	1	248
screenwriter	screenwriter	815	2342	84.16	1	415
skateboarding	skateboarding	541	1322	113.11	1	261
skiing	skiing	690	2226	95.01	1	478
sociology	sociology	307	1691	132.70	1	429
sport_football	sport, football	59538	40821	73.60	1	501
storage	storage	923	3758	100.55	1	490
surfing	surfing	1819	2983	79.93	1	484
sustainability_ecology	sustainability, ecology	3400	17419	138.05	1	500
swimming	swimming	190	1811	50.65	1	335
teaching	teaching	915	4521	100.45	1	491
tech	tech	87409	36632	77.18	1	520
tennis	tennis	28316	2954	75.41	1	500

theatre_theater	theatre, theater	3344	18320	115.94	1	500
toys	toys	399	2083	64.17	1	384
trading	trading	2231	7304	133.27	1	492
tvshows_drama_-actor_hollywood	tvshows, drama, actor, hollywood	7003	23535	75.49	1	501
university	university	294	2347	55.61	1	180
vegetarian	vegetarian	458	2260	105.70	1	474
weather	weather	1936	7660	42.35	1	500
wedding	wedding	7072	21085	107.13	0	498
wine	wine	9371	37798	185.90	1	500
wrestling	wrestling	2237	7538	98.00	1	502
writing	writing	2947	42081	187.52	1	527
yoga	yoga	2782	10320	111.52	1	498
youtube	youtube	2635	7586	53.66	1	467
Total number of communities	Total number of keywords	Total	Total	Average	Avg.	Average
166	200	830448	1714748	112.33	1.01	455.64

Table 6: Overview of the interest communities, the used keywords, lists and members in the sampling process.

	Name	Dimension of social capital	Description
Group bonding social capital			
Research question:	How does group bonding social capital influence the overall diffusion of tweets inside the group?		
Type of analysis:	The unit of analysis is the group. The network is partitioned according to interest groups. Only actors from same group are contributing towards metrics.		
Hypotheses	H1a	Structural	The more structural bonding group social capital the group realizes, the more retweets are exchanged in the group network.
	H1b	Relational	The more relational bonding group social capital the group realizes, the more retweets are exchanged in the group network.
Dependent variable	RT_density	Outcome	Density in the RT network
Indicators	FF_density	Structural	Density in the FF network
	AT_density	Structural	Density in the AT network
	FF_clustering	Structural	Clustering in the FF network
	AT_clustering	Structural	Clustering in the AT network
	FF_avg_path_length	Structural	Average path length in the FF network
	AT_avg_path_length	Structural	Average path length in the AT network
	FF_reciprocity	Relational	Reciprocity in the FF network
	AT_reciprocity	Relational	Reciprocity in the AT network
	Member_count	Control variable	The broadness of the keyword representing the interest category, measured by the total number of unique members found on lists for this category.
Individual bonding social capital			
Research question:	How does a person's individual bonding social capital influence the diffusion of their tweets to other members of the group?		
Type of analysis:	The unit of analysis is the individual. The network is partitioned according to interest groups. Only actors from same group are contributing towards metrics.		

Hypotheses	H2a	Structural	The more structural individual bonding social capital one realizes in his own group, the more one's tweets are retweeted inside the group.
	H2b	Relational	The more relational individual bonding social capital one realizes in his own group, the more one's tweets are retweeted inside the group.
	H2c	Cognitive	The more cognitive individual bonding social capital one realizes in his own group, the more one's tweets are retweeted inside the group.
Dependent variable	RT_vol_in	Outcome	Number of retweets received from group members (Actor's weighted indegree for actors in the RT network)
Indicators	FF_in_deg	Structural	Actor's indegree in the FF network
	AT_vol_in	Structural	Actor's weighted indegree in the AT network
	FF_close	Structural	Actor's closeness centrality in the FF network
	AT_close	Structural	Actor's closeness centrality in the AT network
	FF_page	Structural	Actor's PageRank in the FF network
	AT_page	Structural	Actor's PageRank in the AT network
	FF_rec	Relational	Actor's reciprocity in the FF network
	AT_rec	Relational	Actor's reciprocity in the AT network
	AT_avg	Relational	Actor's average tie strength in the AT network
	Ranking_in_interest	Cognitive	Actor's number of listings received from lists listing persons for this interest group. Actors with most listings achieve the first rank.
Group bridging social capital			
Research question:	How does group bridging social capital influence the diffusion of the group's tweets to other groups?		
Type of analysis:	The unit of analysis is the group. Each group has been reduced to one node in the blockmodel network. Ties between group are an aggregation of individual ties.		

Hypotheses	H3a	Structural	The more structural group bridging social capital the group realizes, the more its tweets are retweeted by other groups.
	H3b	Relational	The more relational group bridging social capital the group realizes, the more its tweets are retweeted by other groups.
	H3c	Cognitive	The more cognitive group bridging social capital the group realizes, the more its tweets are retweeted by other groups.
Dependent variable	RT_volume_in	Outcome	Number of retweets received from members of other groups (In the blockmodel network equals the group's weighted indegree in the RT network)
Indicators	FF_volume_in	Structural	The group's weighted indegree in the FF network
	AT_volume_in	Structural	The group's weighted indegree in the AT network
	FF_betweenness	Structural	The group's betweenness in the FF network
	AT_betweenness	Structural	The group's betweenness in the AT network
	FF_closeness	Structural	The group's closeness in the FF network
	AT_closeness	Structural	The group's closeness in the AT network
	FF_pagerank	Structural	The group's PageRank in the FF network
	FF_pagerank	Structural	The group's PageRank in the AT network
	FF_rec	Relational	The group's reciprocity in the FF network
	AT_rec	Relational	The group's reciprocity in the AT network
	FF_avg	Relational	The group's average tie strength in the FF network
	AT_avg	Relational	The group's average tie strength in the AT network
	Agg_number_of_interests	Cognitive	The aggregated (sum) number interest groups, for which the actors of this category have been listed.
	Member_count	Control Variable	see above

Individual bridging social capital

Research question:	How does a person's individual bridging social capital influence the diffusion of their tweets to members of other groups?		
Type of analysis:	The unit of analysis is the individual. For each actor the members of the own group are removed (filtered) and only members from other groups contribute to metrics.		
Hypotheses	H4a	Structural	The more individual structural bridging social capital one realizes, the more one's tweets are retweeted by members of other groups.
	H4b	Relational	The more individual relational bridging social capital one realizes, the more one's tweets are retweeted by members of other groups.
	H4c	Cognitive	The more individual cognitive bridging social capital one realizes, the more one's tweets are retweeted by members of other groups.
Dependent variable	RT_vol_in	Outcome	Number of retweets received from members of other groups. Equals the actor's weighted indegree in the filtered RT network.
Indicators	FF_vol_in	Structural	The actor's weighted indegree in the filtered FF network.
	AT_vol_in	Structural	The actor's weighted indegree in the filtered AT network.
	FF_groups_in	Structural	The number of other groups that have at least one attention tie to the actor.
	AT_groups_in	Structural	The number of other groups that have at least one interaction tie to the actor.
	FF_betweenness	Structural	The actor's betweenness in the unfiltered FF network.
	AT_betweenness	Structural	The actor's betweenness in the unfiltered AT network.
	FF_rec	Relational	The actor's reciprocity in the filtered FF network.
	AT_rec	Relational	The actor's reciprocity in the filtered AT network.
	AT_avg	Relational	The actor's average tie strength in the filtered AT network.

	Number_of_interests	Cognitive	The number interests, for which this actor has been listed.
Group bonding vs. bridging social capital			
Type of analysis:	The unit of analysis is the group. For each group the final indicators from final models 1.6 and 3.7 represent the bonding/bridging social capital. A path analysis computes their effect on the opposite dependent variable.		
Hypotheses	H5	Structural, Relational, Cognitive	The more group bonding social capital the group obtains, the less retweets it obtains from other groups.
	H7	Structural, Relational, Cognitive	The more group bridging social capital the group obtains, the less the group's tweets are retweeted inside the group.
Dependent variables	RT_density, RT_volume_in	Outcomes	see above
Indicators	AT_clustering	Structural	group bonding, see above
	AT_density	Structural	group bonding, see above
	FF_avg_path_length	Structural	group bonding, see above
	AT_volume_in	Structural	group bridging, see above
	AT_closeness	Structural	group bridging, see above
	FF_avg	Relational	group bridging, see above
	FF_reciprocity	Relational	group bonding, see above
	Agg_number_of_interests	Cognitive	group bridging, see above
Individual bonding vs. bridging social capital			
Type of analysis:	The unit of analysis is the individual. For each individual the final indicators from final models 2.6 and 4.6 represent the bonding/bridging social capital. A path analysis computes their effect on the opposite dependent variable.		
Hypotheses	H6	Structural, Relational, Cognitive	The more individual bonding social capital the individual obtains, the less retweets he obtains from members of other groups.

	H8	Structural, Relational, Cognitive	The more individual bridging social capital the individual obtains, the less retweets he obtains from members of his own group.
Dependent variables	bonding_RT_vol_in, bridging_RT_vol_in	Outcomes	see above
Indicators	bridging_AT_vol_in	Structural	individual bridging, see above
	bonding_AT_vol_in	Structural	individual bonding, see above
	FF_vol_in	Structural	individual bridging, see above
	AT_close	Structural	individual bonding, see above
	FF_page	Structural	individual bonding, see above
	bonding_FF_rec	Relational	individual bonding, see above
	bridging_FF_rec	Relational	individual bridging, see above
	Ranking_in_interest	Cognitive	individual bonding, see above
	Number_of_interests	Cognitive	individual bridging, see above

Table 7: Tabular overview of the hypotheses and indicators.

Bibliography

Abrahamson, E. and Rosenkopf, L. (1997). Social network effects on the extent of innovation diffusion. *Organization Science*, 8(3):289–309.

Ackland, R. (2005). Mapping the US political blogosphere: are conservative bloggers more prominent? In *BlogTalk Downunder 2005 Conference, Sydney*, pages 1–12.

Adamic, L. and Glance, N. (2005). The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43.

Adar, E., Zhang, L., Adamic, L. A., and Lukose, R. M. (2004). Implicit structure and the dynamics of blogspace. *Workshop on the Weblogging Ecosystem*, 13(1):16989–16995.

Adler, P. S. and Kwon, S.-W. (2002). Social Capital: Prospects for a New Concept. *Academy of Management Review*, 27(1):17–40.

Ahuja, G. (2000). Collaboration Networks, Structural Holes, and Innovation: A Longitudinal Study. *Administrative Science Quarterly*, 45(3):425–455.

Albert, R., Jeong, H., and Barabasi, A.-L. (1999). The diameter of the world wide web. *Nature*, 401(9):130–131.

An, J., Cha, M., Gummadi, K., and Crowcroft, J. (2011). Media landscape in Twitter : A world of new conventions and political diversity. *Artificial Intelligence*, 6(1):18–25.

Anagnostopoulos, A., Kumar, R., and Mahdian, M. (2008). Influence and correlation in social networks. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining KDD 08*, 10(1):1–7.

Analytics, P. (2009). Twitter Study – August 2009. pages 1–13. Retrieved 10.10.2009 from http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf.

Aral, S. (2010). Commentary–Identifying Social Influence: A Comment on Opinion Leadership and Social Contagion in New Product Diffusion. *Marketing Science*, 30(2):217–223.

Aral, S., Muchnik, L., and Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21544–21449.

Aral, S. and Van Alstyne, M. W. (2009). Networks, Information & Brokerage: The Diversity-Bandwidth Tradeoff. In *Social Science Research Network Working Paper Series*, pages 90–171.

Arndt, J. (1967). Role of Product-Related Conversations in the Diffusion of a New Product. *Journal of Marketing Research*, 4(3):291–295.

Atkin, R. (1974). *Mathematical structure in human affairs*. Heinemann Educational, London.

Atkin, R. (1977). *Combinatorial connectives in social systems*. Birkhauser, Basel.

Baker, W. E. and Iyer, A. (1992). Information Networks and Market Behavior. *Journal of Mathematical Sociology*, 16(4):305–332.

Bakshy, E., Hofman, J. M., Watts, D. J., and Mason, W. A. (2011). Identifying ' Influencers ' on Twitter. In *Communications of the ACM*, pages 1–10.

Banerjee, N., Chakraborty, D., Dasgupta, K., Mittal, S., Joshi, A., Nagar, S., Rai, A., and Madan, S. (2009). User interests in social media sites. *Proceeding of the 18th ACM conference on Information and knowledge management CIKM 09*, 3(1):1823–1826.

Baruch, Y. and Lin, C.-P. (2012). All for one, one for all: Coopetition and virtual team performance. *Technological Forecasting and Social Change*, 79(6):1155–1168.

Bass, F. (1969). A new product growth model for consumer durables. *Management Science*, 15(5):215–227.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi : An Open Source Software for Exploring and Manipulating Networks. *Artificial Intelligence*, 2(2):361–362.

Batagelj, V. and Mrvar, A. (2008). Analysis of kinship relations with Pajek. *Social Science Computer Review*, 26(2):224–246.

Bauer, H. H. and Grether, M. (2005). Virtual Community : Its Contribution to Customer Relationships by Providing Social Capital. *Journal of Relationship Marketing*, 4(1):91–110.

Becker, M. H. (1970). Sociometric Location and Innovativeness: Reformulation and Extension of the Diffusion Model. *American Sociological Review*, 35(2):267 – 282.

Berelson, B., Lazarsfeld, P. F., and McPhee, W. N. (1954). *Voting: A study of opinion formation in a presidential campaign*. University of Chicago Press, Chicago.

Berkman, L. F. and Syme, S. L. (1979). Social networks, host resistance, and mortality: a nine-year follow-up study of Alameda County residents. *American Journal of Epidemiology*, 109(2):186–204.

Berry, L. (1995). Relationship marketing of services–growing interest, emerging perspectives. *Journal of the Academy of Marketing Science*, 23(4):236 – 245.

Berth, R. (1993). Szenen und soziale Netzwerke: Was steckt dahinter? Empirische Daten führen zu einer neuen Sicht. In *Social Networks: Neue Dimensionen der Markenführung*, pages 13–45. Econ-Verl., Düsseldorf.

Bhandari, H. and Yasunobu, K. (2009). What is Social Capital? A Comprehensive Review of the Concept. *Asian Journal of Social Science*, 37(3):480–510.

Blei, D. M. (2012). Introduction to Probabilistic Topic Models. *Communications of the ACM*, 55(4):1–16.

Böhringer, M. (2009). Really Social Syndication: A Conceptual View on Microblogging. *Sprouts: Working Papers on Information Systems*, 9(31):1–14.

Bonacich, P. (1987). Power and Centrality: A Family of Measures. *American Journal of Sociology*, 92(5):1170.

Borgatti, S., Jones, C., and Everett, M. (1998). Network measures of social capital. *Connections*, 21(2):27–36.

Borgatti, S. P. (1998). A SOCNET Discussion on the Origins of the Term Social Capital. *Connections*, 21(2):37–46.

Borgatti, S. P., Everett, M. G., and Freeman, L. C. (2002). Ucinet for Windows: Software for Social Network Analysis.

Borgatti, S. P. and Foster, P. C. (2003). The Network Paradigm in Organizational Research: A Review and Typology. *Journal of Management*, 29(6):991–1013.

Borgatti, S. P. and Halgin, D. S. (2011). On Network Theory. *Organization Science*, 22(5):1168–1181.

Bourdieu, P. (1986). The forms of capital. In Richardson, J. G., editor, *Handbook of Theory and Research for the Sociology of Education*, chapter 9, pages 241–258. Greenwood Press.

Bowerman, B.L. , O'Connell R, R. (1990). *Linear statistical models: An applied approach (2nd ed.)*. PWS-Kent Publishing Company, Belmont, CA: Duxbury.

Bowman, S. And Willis, C. (2003). We Media: How Audiences are Shaping the Future of News and Information. Technical report, American Press Institute, Reston.

Boyd, D. M. and Ellison, N. B. (2008). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.

Boyd, D. M., Golder, S., and Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10.

Brass, D. J. (1995). A social network perspective on human resources management. *Research in Personnel and Human Resources Management*, 13(1):39–79.

Brooks, B., Welser, H. T., Hogan, B., and Titsworth, S. (2011). Socioeconomic status updates. *Information Communication Society*, 14(4):529–549.

Brown, J. J. and Reingen, P. H. (1987). Social Ties and Word-of-Mouth Referral Behavior. *Journal of Consumer Research*, 14(3):350–362.

Brzozowski, M. J. and Romero, D. M. (2011). Who Should I Follow ? Recommending People in Directed Social Networks. In *Proceedings of CSCW2011 Hangzhou, China*, pages 458–452.

Burt, R. (1987). Social Contagion and Innovation. *American Journal of Sociology*, 92(6):1287–1335.

Burt, R. S. (1995). *Structural holes: The social structure of competition*. Harvard University Press, Cambridge.

Burt, R. S. (1999). The Social Capital of Opinion Leaders. *Annals of the American Academy of Political and Social Science*, 566(1):37–54.

Burt, R. S. (2000). The network structure of social capital. *Research In Organizational Behavior*, 22(4):345–423.

Burt, R. S. (2010). Structural holes in virtual worlds. In *Working Paper*, pages 1–83. University of Chicago Booth School of Business.

Campbell, K. E. (1991). Name generators in surveys of personal networks. *Social Networks*, 13(3):203–221.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. In *Fourth International AAAI Conference on Weblogs and Social Media*, pages 10–17.

Chan, K. K. and Misra, S. (1990). Characteristics of the Opinion Leader: A New Dimension. *Journal of Advertising*, 19(3):53–60.

Charlie Nesson, T. (2009). The Iranian Election on Twitter: The first Eighteen Days. *Web Ecology Project*, 1(26):1–7.

Chen, Y.-H. C. Y.-H., Li, S.-S. L. S.-S., and Chen, Y.-T. C. Y.-T. (2010). Extracting Topics Information from Conference Web Pages Using Page Segmentation and SVM. In *2010 International Conference on Technologies and Applications of Artificial Intelligence*, pages 270–277.

Cheng, Y., Qiu, G., Bu, J., Liu, K., Han, Y., Wang, C., and Chen, C. (2008). Model bloggers' interests based on forgetting mechanism. *Proceeding of the 17th international conference on World Wide Web WWW 08*, 15(3):1129–1131.

Chiu, C., Hsu, M., and Wang, E. (2006). Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. *Decision Support Systems*, 42(3):1872–1888.

Choudhury, M., Sundaram, H., John, A., Seligmann, D. D., and Kelliher, A. (2010). "Birds of a Feather": Does User Homophily Impact Information Diffusion in Social Media? In *Arxiv preprint arXiv10061702*, pages 1–31.

Choudhury, M. D. and Hari, Y.-R. L. (2010). How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media ? In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 34–41.

Christopher, A. (2004). The Dunbar Number as a Limit to Group Sizes. Retrieved 12.03.2011 from http://www.lifewithalacrity.com/2004/03/the_dunbar_numb.html.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, New York.

Cohen, S., Doyle, W. J., Skoner, D. P., Rabin, B. S., and Gwaltney, J. M. (1997). Social ties and susceptibility to the common cold. *Jama The Journal Of The American Medical Association*, 278(15):1940–1944.

Coleman, J. and Katz, E. (1966). *Medical innovation: A diffusion study*. Bobbs-Merrill Company, New York.

Coleman, J., Katz, E., and Menzel, H. (1957). The Diffusion of an Innovation Among Physicians. *Sociometry*, 20(4):253 – 270.

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *American Journal of Sociology*, 94(1):95–120.

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonc, B., Flammini, A., and Menczer, F. (2011). Political Polarization on Twitter. *Networks*, 133(26):89–96.

Cornfield, M., Carson, J., Kalis, A., and Simon, E. (2005). Buzz, Blogs and Beyond: the Internet and the National Discourse in the Fall of 2004. *Pew Internet and American Life Project*, 23(1):1–33.

Coulter, R. A., Feick, L. F., and Price, L. L. (2002). Changing faces: cosmetics opinion leadership among women in the new Hungary. *European Journal of Marketing*, 36(12):1287–1308.

Counts, Scott, Pal, A. (2011). Identifying Topical Authorities in Microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 45–54.

Crawford, K. (2009). Following you: Disciplines of listening in social media. *Continuum: Journal of Media & Cultural Studies*, 23(4):525–535.

Cross, R. and Parker, A. (2004). *The Hidden Power of Social Networks*. Harvard Business School Press, Boston.

Cucerzan, S. (2007). Large-Scale Named Entity Disambiguation Based on Wikipedia Data. *Computational Linguistics*, 6(1):708–716.

Damme, C. V., Hepp, M., and Siorpaes, K. (2007). FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies. *Bridging the Gap between Semantic Web and Web*, 2(1):57–70.

Davenport, S. and Daellenbach, U. (2011). Belonging' to a Virtual Research Centre: Exploring the Influence of Social Capital Formation Processes on Member Identification in a Virtual Organization. *British Journal of Management*, 22(1):54–76.

Davenport, T. H. (2002). *The Attention Economy*. Harvard Business Press, Boston.

Doerfel, M. (1998). What constitutes semantic network analysis? A comparison of research and methodologies. *Connections*, 21(2):16–26.

Doerfel, M. L. and Barnett, G. A. (1999). A Semantic Network Analysis of the International Communication Association. *Human Communication Research*, 25(4):589–603.

Dunbar, R. I. M. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493.

Durbin, J. and Watson, G. S. (1957). Testing for Serial Correlation in Least Squares Regression. *Biometrika*, 44(1):57–66.

Ediger, D., Jiang, K., Riedy, J., Bader, D. A., and Corley, C. (2010). Massive Social Network Analysis: Mining Twitter for Social Good. In *39th International Conference on Parallel Processing*, pages 583–593.

Ellison, N., Heino, R., and Gibbs, J. (2006). Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment. *Journal of Computer-Mediated Communication*, 11(2):415–441.

Ellison, N. B., Steinfield, C., and Lampe, C. (2007). The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168.

Ellison, N. B., Steinfield, C., and Lampe, C. (2011). Connection Strategies: Social Capital Implications of Facebook-enabled Communication Practices. *New Media & Society*, 13(6):873–892.

Etzioni, A. and Etzioni, O. (1999). Face-to-Face and Computer-Mediated Communities, A Comparative Analysis. *The Information Society*, 15(4):241–248.

Everett, M. G. and Borgatti, S. P. (1994). Regular Equivalence - General-Theory. *Journal of Mathematical Sociology*, 19(1):29–52.

Everett, M. G. and Borgatti, S. P. (1999). The Centrality of Groups and Classes. *Journal of Mathematical Sociology*, 23(3):181 – 201.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305.

Flake, G. W., Tsioutsiouliklis, K., and Zhukov, L. (2004). Methods for Mining Web Communities : Bibliometric , Spectral , and Flow. In *Web Dynamics*, pages 45–68. Springer Verlag, Berlin.

Flap, H. (2004). *Creation and returns of social capital: a new research program*. Routledge, London.

Frank, O. (1979). Sampling and estimation in large social networks. *Social Networks*, 1(1):91–101.

Franzen, A. and Pointner, S. (2007). Sozialkapital: Konzeptualisierungen und Messungen. In *Sozialkapital Grundlagen und Anwendungen*, pages 66–90. VS Verlag für Sozialwissenschaften, Wiesbaden.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41.

Freeman, L. C. (1979). The networkers network: a study of the impact of a new communications medium on sociometric structure. In *Social Science Research Reports*, pages 1–36. School of Social Sciences University of California, Los Angeles.

Freeman, L. C. (2004). *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press, Vancouver.

Friedkin, N. E. (1982). Information flow through strong and weak ties in intraorganizational social networks. *Social Networks*, 3(4):273–285.

Fukuyama, F. (2001). Social capital, civil society and development. *Third world quarterly*, 22(1):7–20.

Gaag, M. V. D. (2005). Measurement of individual social capital. *Social Capital and Health*, 5(10):29–49.

Gaines, B. J. and Mondak, J. J. (2009). Typing Together? Clustering of Ideological Types in Online Social Networks. *Journal of Information Technology Politics*, 6(3):216–231.

Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., and Kellerer, W. (2010). Outtweeting the Twitterers- Predicting Information Cascades in Microblogs. In *WOSN'10 Proceedings of the 3rd conference on Online social networks*, pages 1–9.

Garcia-Silva, A., Kang, J., Lerman, K., and Corcho, O. (2012). Characterizing emergent semantics in Twitter lists. In *The Semantic Web: Research and Applications*, pages 530–544.

Gargiulo, M. and Benassi, M. (1997). *The Dark Side of Social Capital*. Lazarsfeld Center at Columbia University, New York.

Garton, L., Haythornthwaite, C., and Wellman, B. (1997). Studying Online Social Networks. *Journal of Computer-Mediated Communication*, 3(1):1–15.

Gatignon, H. and Robertson, T. S. (1985). A Propositional Inventory for New Diffusion Research. *Journal of Consumer Research*, 11(4):849–867.

Gayo-Avello, D. (2010). Nepotistic Relationships in Twitter and their Impact on Rank Prestige Algorithms. *Arxiv preprint arXiv*, 10040816(1):1–40.

Gayo-Avello, D., Brenes, D. J., Fernández-Fernández, D., Fernández-Menéndez, M. E., and García-Suárez, R. (2010). De retibus socialibus et legibus momenti. *Europhysics Letters*, 94(3):1–21.

Girvan, M., Newman, M. E. J., and Girvan, M, Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826.

Gitlin, T. (1978). Media Sociology: The Dominant Paradigm. *Theory and Society*, 6(2):205–253.

Gladwell, M. (2000). *The Tipping Point*. Little, Brown and Company, New York.

Goldenberg, J., Han, S., Lehmann, D. R., and Hong, J. W. (2009). The Role of Hubs in the Adoption Process. *Journal of Marketing*, 73(2):1–13.

Golder, S. A. and Yardi, S. (2010). Structural Predictors of Tie Formation in Twitter : Transitivity and Mutuality. In *Social Computing SocialCom 2010 IEEE Second International Conference on*, pages 88–95.

Gonçalves, V. and Ballon, P. (2011). Adding value to the network: Mobile operators' experiments with Software-as-a-Service and Platform-as-a-Service models. *Telematics and Informatics*, 28(1):12–21.

Gould, R. V. and Fernandez, R. M. (1989). Structures of Mediation: A Formal Approach to Brokerage in Transaction Networks. *Sociological Methodology*, 19(1):89 – 126.

Grabowicz, P. A., Ramasco, J. J., Moro, E., Pujol, J. M., and Eguiluz, V. M. (2012). Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media. *PLoS ONE*, 7(1):e29358.

Granovetter, M. (1978). Threshold Models of Collective Behavior. *The American Journal of Sociology*, 83(6):1420 – 1443.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380.

Greene, D., O'Callaghan, D., and Cunningham, P. (2012). Identifying Topical Twitter Communities via User List Aggregation. *arXiv preprint 1207.0017*, pages 1–9.

Greene, D., Reid, F., and Sheridan, G. (2011). Supporting the Curation of Twitter User Lists. *arXiv preprint arXiv:1110.1349*, pages 1–8.

Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the 13th conference on World Wide Web WWW 04*, pages 491–501.

Gupta, S., Slawski, B., Xin, D., and Yao, W. (2011). The Twitter Rumor Network : Subject and Sentiment. In *California Institute of Technology Notes*, pages 1–10.

Halpin, H., Robu, V., and Shepherd, H. (2007). The complex dynamics of collaborative tagging. *Proceedings of the 16th international conference on World Wide Web WWW 07*, 07(1):211–220.

Hanifan, L. J. (1920). The Community Center. *The Annals of the American Academy of Political and Social Science*, 67(1):130–138.

Hannerz, U. (1971). *Soulside: Inquiries into ghetto culture and community.* Frank Cass and Co. Ltd., New York.

Hansen, L. K., Arvidsson, A., Nielsen, F. A. r., Colleoni, E., and Etter, M. (2011). Good Friends, Bad News-Affect and Virality in Twitter. *Arxiv preprint arXiv11010510*, pages 1–14.

Hansen, M. T. (2002). Knowledge Networks: Explaining Effective Knowledge Sharing in Multiunit Companies. *Organization Science*, 13(3):232–248.

Harrison, C. W., Boorman, S. A., and Breiger, R. L. (1976). Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions. *American journal of Sociology*, 81(4):730–780.

Häuberer, J. (2010). *Social Capital Theory Towards a Methodological Foundation*. Springer, Berlin.

Haythornthwaite, C. (2002). Strong, Weak, and Latent Ties and the Impact of New Media. *The Information Society*, 18(5):385–401.

Heidler, R. (2006). *Die Blockmodellanalyse. Theorie und Anwendung einer netzwerkanalytischen Methode*. Deutscher Universitätsverlag, Wiesbaden.

Henri, F. and Pudelko, B. (2003). Understanding and analysing activity and learning in virtual communities. *Journal of Computer Assisted Learning*, 19(4):474–487.

Hernando, A., Villuendas, D., Vesperinas, C., Abad, M., and Plastino, A. (2009). Unravelling the size distribution of social groups with information theory on complex networks. *European Physical Journal*, 76(1):87–97.

Holland, P. W. and Leinhardt, S. (1981). An Exponential Family of Probability Distributions for Directed Graphs. *Journal of the American Statistical Association*, 76(373):33–50.

Honey, C. and Herring, S. C. (2009). Beyond Microblogging: Conversation and Collaboration via Twitter. In *HICSS 09 42nd Hawaii International Conference on*, pages 1–10.

Hong, L., Dan, O., and Davison, B. D. (2011). Predicting popular messages in Twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 57–58.

Hong, L. and Davison, B. D. (2010). Empirical Study of Topic Modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88.

Horrigan, J., Rainie, L., and Fox, S. (2001). Networks that nurture long-distance relationships and local ties. In *Pew Internet and American Life Project*, pages 1–28. Retrieved 10.10.2009 from http://www.pewinternet.org/˜/media/Files/Reports/2001/PIP_Communities_Report.pdf.

Hsiao, C.-C. and Chiou, J.-S. (2011). The effects of a player's network centrality on resource accessibility, game enjoyment, and continuance intention: A study on online gaming communities. *Electronic Commerce Research and Applications*, 11(1):75–84.

Huang, J. and Lin, C. (2011). To stick or not to stick: The social response theory in the development of continuance intention from organizational cross-level perspective. *Computers in Human Behavior*, 27(5):1963–1973.

Huang, J., Thornton, K. M., and Efthimiadis, E. N. (2010). Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 173–178.

Huber, F. (2009). Social capital of economic clusters: towards a network-based conception of social resources. *Tijdschrift voor Economische en Sociale Geografie*, 100(2):160–170.

Huberman, B. A., Romero, D. A., and Wu, F. (2009). Social networks that matter: Twitter under the microscope. *First Monday*, 14(1):1–4.

Huvila, I., Holmberg, K., Ek, S., and Widén-Wulff, G. (2010). Social capital in Second Life. *Online Information Review*, 34(2):295–316.

Huysman, M. and Wulf, V. (2005). The Role of Information Technology in Building and Sustaining the Relational Base of Communities. *The Information Society*, 21(2):81–89.

Iyengar, R., Van Den Bulte, C., and Valente, T. W. (2010). Opinion leadership and social contagion in new product diffusion. *Marketing Science*, 30(2):195–212.

Jacobs, J. (1961). *The Death and Life of Great American Cities*. Vintage Series. Random House, New York.

Jansen, B., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science*, 60(11):2169–2188.

Jansen, D. (2006). *Einführung in die Netzwerkanalyse: Grundlagen, Methoden, Forschungsbeispiele*. Leske + Budrich, Wiesbaden.

Jarvis, J. (2008). In Mumbai, witnesses are writing the news. In: Guardian, 1st of December, 2008. Retrieved 21.04.2010 from http://www.guardian.co.uk/media/2008/dec/01/mumbai-terror-digital-media.

Java, A., Song, X., Finin, T., and Tseng, B. (2009). Why We Twitter: An Analysis of a Microblogging Community. *Advances in Web Mining and Web Usage Analysis*, 5439(1):118–138.

Kadushin, C. (2004). Too much investment in social capital? *Social Networks*, 26(1):75–90.

Katz, J., Rice, R. E., and Aspden, P. (2001). The Internet, 1995-2000: Access, Civic Involvement, and Social Interaction. *American Behavioral Scientist*, 45(3):405–419.

Katz, L. and Powell, J. H. (1955). Measurement of the tendency toward reciprocation of choice. *Sociometry*, 18(4):403–409.

Kempe, D. (2003). Maximizing the spread of influence through a social network. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining KDD 03*, 41(3):137–146.

Kernighan, B. W. and Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307.

Kim, W., Jeong, O.-R., and Lee, S.-W. (2010). On social Web sites. *Information Systems*, 35(2):215–236.

Kobayashi, T., Ikeda, K., and Miyata, K. (2007). Information , Communication & Society Social capital online : Collective use of the Internet and reciprocity as lubricants of democracy. *Information Communication Society*, 9(12):582–611.

Kossinets, G., Kleinberg, J. M., and Watts, D. (2008). The Structure of Information Pathways in a Social Communication Network. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 435–443.

Kossinets, G. and Watts, D. (2009). Origins of Homophily in an Evolving Social Network1. *American Journal of Sociology*, 115(2):405–450.

Kozinets, R. V., Valck, K. D., Wojnicki, A. C., and Wilner, S. J. S. (2010). Networked Narratives : Understanding Word-of-Mouth. *Journal of Marketing*, 74(3):71–89.

Krackhardt, D. (1992). The strength of strong ties: The importance of philos in organizations. In *Networks and organizations Structure form and action*, pages 216–239. Harvard Business School Press, Boston.

Krackhardt, D. and Kilduff, M. (2002). Structure, culture and Simmelian ties in entrepreneurial firms. *Social Networks*, 24(3):279–290.

Krishnamurthy, B., Gill, P., and Arlitt, M. (2008). A few chirps about twitter. In *Proceedings of the first workshop on Online social networks WOSP 08*, pages 19–24.

Kulkarni, S., Singh, A., Ramakrishnan, G., and Chakrabarti, S. (2009). Collective annotation of Wikipedia entities in web text. In *Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining (2009)*, pages 457–466.

Kumpula, J. M., Onnela, J.-P., Saramäki, J., Kaski, K., and Kertész, J. (2007). Emergence of communities in weighted networks. *Physical review letters*, 99(22):228701.

Kwak, H., Chun, H., and Moon, S. (2011). Fragile Online Relationship : A First Look at Unfollow Dynamics in Twitter. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 1091–1100.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600.

Laniado, D., Eynard, D., Colombetti, M., and Milano, P. (2007). Using WordNet to turn a folksonomy into a hierarchy of concepts. In *Semantic Web Applications and Perspectives SWAP 2007*, pages 192–201.

Lapinski, M. K. and Rimal, R. N. (2005). An Explication of Social Norms. *Communication Theory*, 15(2):127–147.

Lazarsfeld, P. F., Berelson, B., and Gaudet, H. (1965). *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. Columbia University Press, New York.

Lee, C., Kwak, H., Park, H., and Moon, S. (2010). Finding influentials based on the temporal order of information adoption in twitter. In *Proceedings of the 19th international conference on World wide web WWW 10*, pages 1137–1138.

Lerman, K. (2007). Social Information Processing in News Aggregation. *IEEE Internet Computing*, 11(6):16–28.

Lerman, K. and Ghosh, R. (2010). Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks. In *Fourth International AAAI Conference on Weblogs and Social Media*, pages 90–97.

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., and Hurst, M. (2007). Cascading Behavior in Large Blog Graphs. *SIAM International Conference on Data Mining SDM 2007*, pages 1–21.

Levy, Y. and Ellis, T. J. (2006). A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. *Informing Science: International Journal of an Emerging Transdiscipline*, 9(1):181–212.

Lewis, S., Kaufhold, K., and Lasorsa, D. (2010). Thinking About Citizen Journalism. *Journalism Practice*, 4(2):163–179.

Lin, C. (2011). Modeling job effectiveness and its antecedents from a social capital perspective: A survey of virtual teams within business organizations. *Computers in Human Behavior*, 27(2):915–923.

Lin, N. (1999). Building a Network Theory of Social Capital. *Connections*, 22(1):28–51.

Lin, N. (2002). *Social capital: A theory of social structure and action*. Cambridge University Press, Cambridge.

Lin, N. and Dumin, M. (1986). Access to occupations through social ties. *Social Networks*, 8(4):365–385.

Lippmann, W. (1922). *Public opinion*. Harcourt, Brace and Co., New York.

Lu, Y. and Yang, D. (2011). Information exchange in virtual communities under extreme disaster conditions. *Decision Support Systems*, 50(2):529–538.

Mahajan, V. and Peterson, R. A. (1985). *Models for innovation diffusion*. Sage Publications, New York.

Maletzke, G. (1963). *Psychologie der Massenkommunikation*. Hans Bredow-Institut, Hamburg.

Mathwick, C., Wiertz, C., and De Ruyter, K. (2008). Social Capital Production in a Virtual P3 Community. *Journal of Consumer Research*, 34(6):832–849.

McAuley, J. and Leskovec, J. (2012). Learning to discover social circles in ego networks. In *25th Conference on Advances in Neural Information Processing Systems*, pages 548–556.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a Feather: Homophily. *Annual Review of Sociology*, 27(1):415–444.

Mehra, A., Kilduff, M., and Brass, D. J. (2001). The Social Networks of High and Low Self-Monitors: Implications for Workplace Performance. *Administrative Science Quarterly*, 46(1):121–146.

Menzel, H. and Katz, E. (1955). Social Relations and Innovation in the Medical Profession: The Epidemiology of a New Drug. *The Public Opinion Quarterly*, 19(4):337–352.

Merton, R. K. (1957). *Patterns of influence: Local and cosmopolitan influentials*. The Free Press, New York.

Michelson, M. and Macskassy, S. A. (2010). Discovering Users ' Topics of Interest on Twitter : A First Look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pages 73–79.

Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *CIKM 07 Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242.

Millen, D. R., Fontaine, M. A., and Muller, M. J. (2002). Understanding the benefit and costs of communities of practice. *Communications of the ACM*, 45(4):69.

Monge, P. R. and Contractor, N. S. (2003). *Theories of Communication Networks*. Oxford University Press, New York.

Moore, S., Shiell, A., Hawe, P., and Haines, V. A. (2005). The privileging of communitarian ideas: citation practices and the translation of social capital into public health research. *American Journal of Public Health*, 95(8):1330–1337.

Nahapiet, J. and Ghoshal, S. (1998). Social Capital, Intellectual Capital, and the Organizational Advantage. *Academy of Management Review*, 23(2):242–266.

Narayan, D. and Cassidy, M. F. (2001). A Dimensional Approach to Measuring Social Capital: Development and Validation of a Social Capital Inventory. *Current Sociology*, 49(2):59–102.

Newman, M., Barabasi, A.-L., and Watts, D. J. (2006). *The structure and dynamics of networks*. Princeton studies in complexity. Princeton University Press, New York.

Newman, M. E. J. (2002a). Assortative mixing in networks. *Physical Review Letters*, 89(20):1–5.

Newman, M. E. J. (2002b). The spread of epidemic disease on networks. *Physical Review E*, 66(1):1–12.

Newman, N. (2009). The rise of social media and its impact on mainstream journalism. *Reuters Institute for the Study of Journalism*, 8(2):1–5.

Norris, P. (2002). The Bridging and Bonding Role of Online Communities. *The Harvard International Journal of Press Politics*, 7(3):3–13.

Norris, P. (2003). Preaching to the converted? Pluralism, participation and party websites. *Party Politics*, 9(1):21–45.

Oh, H., Labianca, G., and Chung, M.-H. (2006). A multilevel model of group social capital. *Academy of Management Review*, 31(3):569–582.

Onnela, J. P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A. L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18):7332–7336.

Onyx, J. and Bullen, P. (2000). Measuring Social Capital in Five Communities. *The Journal of Applied Behavioral Science*, 36(1):23–42.

Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251.

Opsahl, T. and Panzarasa, P. (2009). Clustering in weighted networks. *Social Networks*, 31(2):155–163.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. *World Wide Web Internet And Web Information Systems*, 54(2):1–17.

Pariser, E. (2011). *The Filter Bubble*. Penguin Press.

Paxton, P. (1999). Is Social Capital Declining in the United States? A Multiple Indicator Assessment. *American Journal of Sociology*, 105(1):88–127.

Petroczi, A., Bazsó, F., and Nepusz, T. (2010). Measuring tie-strength in virtual social networks. *Connections*, 27(2):39–52.

Pitt, L., Merwe, R. V. D., Berthon, P., Salehi-Sangeri, E., and Barnes, B. R. (2010). Swedish BioTech SMEs: The Veiled Values in Online Networks. *Technovation*, 26(5):553–560.

Plotkowiak, T., Ebermann, J., and Stanoevska-Slabeva, K. (2010). A Longitudinal Social Network Analysis of German Politicians' Twitter Accounts. In *Sunbelt XXX conference*, page 789.

Plotkowiak, T., Stanoevska-Slabeva, K., Ebermann, J., Fleck, M., and Meckel, M. (2012). Die Rolle von Journalisten in Sozialen Medien am Beispiel Twitter. *Medien und Kommunikation*, 1(60):102–125.

Portes, A. (1998). Social Capital: Its Origins and Applications in Modern Sociology. *Annual Review of Sociology*, 24(1):1–24.

Pothen, A., Simon, H. D., and Liu, K.-P. P. (1989). Partitioning Sparse Matrices with Eigenvectors of Graphs. *SIAM Journal on Matrix Analysis and Applications*, 11(3):430–452.

Putnam, R. (1995). Bowling alone: America's declining social capital. *Journal of democracy*, 6(1):65–78.

Putnam, R. D. (2000). *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster, New York.

Rafaeli, S., Ravid, G., and Soroka, V. (2004). De-lurking in virtual communities: a social communication network approach to measuring the effects of social and cultural capital. In *37th Annual Hawaii International Conference on System Sciences 2004 Proceedings of the*, pages 1–10.

Ramage, D., Dumais, S., and Liebling, D. (2010). Characterizing Microblogs with Topic Models. *International AAAI Conference on Weblogs and Social Media*, pages 130–137.

Rao, A. R. and Bandyopadhyay, S. (1987). Measures of Reciprocity in a Social Network. *The Indian Journal of Statistics*, 49(2):141–188.

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., and Menczer, F. (2010). Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams. *arXiv preprint arXiv:1011.3768*, pages 1–10.

Rausch, A. and Stegbauer, C. (2006). Netzwerkanalysen internetbasierter Kommunikationsräume. In *Strukturalistische Internetforschung*, pages 149–168. VS Verlag für Sozialwissenschaften, Wiesbaden.

Reagans, R. and McEvily, B. (2003). Network Structure and Knowledge Transfer: The Effects of Cohesion and Range. *Administrative Science Quarterly*, 48(2):240–267.

Resnick, P., Bikson, T., Mynatt, E., Puttnam, R., Sproull, L., and Wellman, B. (2000). Beyond bowling together: Sociotechnical capital. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work CSCW 00*, pages 247–272.

Rheingold, H. (1993). *The Virtual Community: Homesteading on the Electronic Frontier*. Addison-Wesley, London.

Robey, D., Boudreau, M.-C., and Rose, G. M. (2000). Information technology and organizational learning: a review and assessment of research. *Accounting Management and Information Technologies*, 10(2):125–155.

Robu, V., Halpin, H., and Shepherd, H. (2009). Emergence of consensus and shared vocabularies in collaborative tagging systems. *ACM Transactions on the Web*, 3(4):1–34.

Roch, C. H. (2008). The Dual Roots of Opinion Leadership. *The Journal of Politics*, 67(1):110–131.

Rogers, E. M. (1995). *Diffusion of Innovations*. Free Press, New York, 4th. edition.

Rogers, E. M. and Cartano, D. G. (1962). Methods of Measuring Opinion Leadership. *Public Opinion Quarterly*, 26(3):435–441.

Rogers, E. M. and Kincaid, D. L. (1981). *Communication networks: Toward a new paradigm for research*. Free Press, New York.

Romero, D. M. and Kleinberg, J. M. (2010). The Directed Closure Process in Information Networks with an Analysis of Link Formation on Twitter. In *International Conference on Weblogs and Social Media ICWSM*, pages 138–145.

Romero, D. M., Meeder, B., and Kleinberg, J. M. (2011). Differences in the Mechanics of Information Diffusion Across Topics : Idioms , Political Hashtags , and Complex Contagion on Twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704.

Ruhrmann, G. (2003). *Der Wert von Nachrichten im deutschen Fernsehen : ein Modell zur Validierung von Nachrichtenfaktoren*. Leske + Budrich, Wiesbaden.

Ryan, B. and Gross, N. C. (1943). The diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology*, 8(1):15–24.

Schenk, M. (1993). Die Ego-zentrierten Netzwerke von Meinungsbildnern (”Opinion Leaders”). *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 45(2):254–269.

Schenkel, M. and Garrison, G. (2009). Exploring the roles of social capital and team-efficacy in virtual entrepreneurial team performance. *Management Research News*, 32(6):525–538.

Scott, P. J. P. (2000). *Social Network Analysis: A Handbook*. Sage Publications Ltd, New York.

Shirky, C. (2011). The Political Power of Social Media. *Foreign Affairs*, 90(1):28–41.

Singer, J. B. (2004a). More than ink-stained wretches: The resocialization of print journalists in converged newsrooms. *Journalism Mass Communication Quarterly*, 81(4):838–856.

Singer, J. B. (2004b). Strange bedfellows? The diffusion of convergence in four news organizations. *Journalism Studies*, 5(1):3–18.

Sinus Sociovision GmbH (2010). Die Sinus-Milieus. Retrieved 23.06.2010 from http://www.sinus-institut.de/.

Skoric, M. M. and Kwan, G. C. E. (2011). Platforms for mediated sociability and online social capital: the role of Facebook and massively multiplayer online games. *Asian Journal of Communication*, 21(5):467–484.

Smith, A. and Brenner, J. (2012). Twitter Use 2012. Technical report, Pew Research Center.

Smith, M. S. and Giraud-Carrier, C. (2010). Bonding vs. Bridging Social Capital: A Case Study in Twitter. In *2010 IEEE Second International Conference on Social Computing*, pages 385–392.

Smith, M. S., Giraud-Carrier, C., Ventura, D. A., Dewey, D. P., Knutson, C. D., and Embley, D. W. (2011). *A Computational Framework for Social Capital in Online Communities*. PhD thesis, Brigham Young University.

Steinfield, C., Dimicco, J. M., Ellison, N. B., and Lampe, C. (2009). Bowling Online : Social Networking and Social Capital within the Organization. In *Proceedings of the fourth international conference on Communities and technologies*, pages 245–254.

Steinfield, C., Ellison, N., and Lampe, C. (2008). Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology*, 29(6):434–445.

Stone, W. (2006). Measuring Social Capital. *The International journal of social psychiatry*, 52(4):360–75.

Suh, B., Hong, L., Pirolli, P., and Chi, E. (2010). Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 177–184.

Sunstein, C. R. (2002). The law of group polarization. *Journal of political philosophy*, 10(2):175–195.

Sunstein, C. R. (2009). *Infotopia. Wie viele Köpfe Wissen produzieren*. Suhrkamp, Berlin.

Takhteyev, Y, Gruzd, A, Wellman, B. (2010). Geography of Twitter networks. *Social Networks*, 34(1):73–81.

Täube, V. G. (2004). Measuring the Social Capital of Brokerage Roles. *Connections*, 26(1):29–52.

Teng, C.-Y. T. C.-Y. and Chen, H.-H. C. H.-H. (2006). Detection of Bloggers' Interests: Using Textual, Temporal, and Interactive Features. In *IEEEWICACM International Conference on Web Intelligence WI 2006*, pages 366–369.

Thelwall, M. (2010). Webometrics: emergent or doomed? *Information Research*, 15(4):1–28.

Trusov, M., Bodapati, A. V., and Bucklin, R. E. (2010). Determining Influential Users in Internet Social Networks. *Journal of Marketing Research*, 47(8):643–658.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting Elections with Twitter : What 140 Characters Reveal about Political Sentiment. *Word Journal Of The International Linguistic Association*, 280(39):178–185.

Tunkelang, D. (2009). A Twitter Analog to PageRank. pages 1–8. Retrieved 15.04.2010 from http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/.

Valente, T. W. (1993). Diffusion of Innovations and Policy Decision-Making. *The Journal of Communication*, 43(1):30–45.

Valente, T. W. (1995). *Network Models of the Diffusion of Innovations*. Hampton Press, New York.

Valente, T. W. (2010). *Social Networks and Health: Models, Methods, and Applications*. Oxford University Press, Oxford.

Valente Thomas W, Foreman, R. (1998). Integration and radiality: measuring the extent of an individual's connectedness and reachability in a network. *Social Networks*, 20(1):89–105.

Valenzuela, S., Park, N., and Kee, K. F. (2008). Lessons from Facebook: The Effect of Social Network Sites on College Students Social Capital. In *9th International Symposium on Online Journalism in Austin*, pages 1–39.

Valenzuela, S., Park, N., and Kee, K. F. (2009). Is There Social Capital in a Social Network Site?: Facebook Use and College Students' Life Satisfaction, Trust, and Participation1. *Journal of Computer-Mediated Communication*, 14(4):875–901.

Van Den Bulte, C. and Joshi, Y. V. (2007). New Product Diffusion with Influentials and Imitators. *Marketing Science*, 26(3):400–421.

Van Liere, D. (2010). How Far Does a Tweet Travel ? Information Brokers in the Twitterverse. In *Proceedings of the International Workshop on Modeling Social Media*, pages 1–6.

Vergeer, M., Lim, Y. S., and Park, H. W. (2011). Mediated relations: new methods to study online social capital. *Asian Journal of Communication*, 21(5):430–449.

Vernette, E. (2004). Targeting Women's Clothing Fashion Opinion Leaders in Media Planning: An Application for Magazines. *Journal of Advertising Research*, 44(1):90–107.

Vitak, J., Ellison, N. B., and Steinfield, C. (2011). The Ties That Bond: Re-Examining the Relationship between Facebook Use and Bonding Social Capital. In *44th Hawaii International Conference on System Sciences*, pages 1–10.

Walther, J. B. (1996). Computer-Mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction. *Communication Research*, 23(1):3–43.

Wasko, M. M. and Faraj, S. (2005). Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly*, 29(1):35–57.

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press, Cambridge.

Watts, D. (1999). Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology*, 105(2):293–527.

Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):5766–5771.

Watts, D. J. and Peretti, J. (2007). Viral Marketing for the Real World. *Harvard Business Review*, 85(5):22–24.

Webster, J. and Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2):13–23.

Weimann, G. (1994). *The influentials: People who influence people*. SUNY Press, New York.

Wejnert, B. (2002). Integrating models of diffusion of innovations: A conceptual framework. *Annual Review of Sociology*, 28(1):297–326.

Wellman, B., Haase, A. Q., Witte, J., and Hampton, K. (2001). Does the Internet increase, decrease, or supplement social capital. *American Behavioral Scientist*, 45(3):436–455.

Weng, J., Lim, E., Jiang, J., and He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270.

Widen-Wulff, G. (2004). Explaining knowledge sharing in organizations through the dimensions of social capital. *Journal of Information Science*, 30(5):448–458.

Williams, D. (2006). On and Off the 'Net: Scales for Social Capital in an Online Era. *Journal of Computer-Mediated Communication*, 11(2):593–628.

Woolcock, M. (2001). The place of social capital in understanding social and economic outcomes. *Canadian Journal of Policy Research*, 2(1):11–17.

Wu, S., Hofman, J. M., Watts, D. J., and Mason, W. A. (2011). Who Says What to Whom on Twitter. In *Proceedings of the 20th international conference on World wide web*, pages 705–714.

Xiao, H., Li, W., Cao, X., and Tang, Z. (2012). The Online Social Networks on Knowledge Exchange: Online Social Identity, Social Tie and Culture Orientation. *Journal of Global Information Technology Management*, 15(2):4–26.

Yamaguchi, Y., Amagasa, T., and Kitagawa, H. (2011). Tag-based User Topic Discovery using Twitter Lists. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 13–20.

Yang, J. and Counts, S. (2009). Predicting the Speed , Scale , and Range of Information Diffusion in Twitter. In *Proc. of the International AAAI Conference on Weblogs and Social Media*, pages 355–358.

Yardi, S. and Boyd, D. (2010). Tweeting from the Town Square: Measuring Geographic Local Networks. *Proceedings of the International Conference on Weblogs and Social Media*, pages 194–201.

Ye, Q., Fang, B., He, W., and Hsieh, J. P.-A. (2012). Can social capital be transferred cross the boundary of the real and virtual worlds? An empirical investigation of Twitter. *Journal of Electronic Commerce Research*, 13(2):145 – 156.

Yin, D. and Hong, L. (2011). Link Formation Analysis in Microblogs. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1235–1236.

Yoon, C. and Wang, Z.-W. (2011). The role of citizenship behaviors and social capital in virtual communities. *Journal of Computer Information Systems*, 52(1):106–115.

Young, H. P. (2005). The Spread of Innovations Through Social Learning. *Archives of Dermatology*, 141(12):1522–1534.

Young, K. (2011). Social Ties, Social Networks and the Facebook Experience. *International Journal of Emerging Technologies and Society*, 9(1):20 – 34.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473.

Zaman, T. R., Herbrich, R., and Stern, D. (2010). Predicting Information Spreading in Twitter. *Workshop on Computational Social Science and the Wisdom of Crowds, NIPS*, 104(45):17599–17601.

Zhang Y, W. Y. Y. Q. (2012). Community discovery in twitter based on user interests. *Journal Of Computational Information Systems*, 8(3):991–1000.

Zhao, D. and Rosson, M. (2009). How and why people Twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 243–252.

Zhao, J. and Ram, S. (2011). Examining the evolution of networks based on lists in Twitter. In *IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application*, pages 1–8.

Zhao, J., Wu, J., and Xu, K. (2010). Weak ties: subtle role of information diffusion in online social networks. *Physical Review E*, 82(2):1–18.

Zhao, W., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing Twitter and Traditional Media using Topic Models. *Advances in Information Retrieval*, 6611(1):338–349.

Zhong, Z.-J. (2011). The effects of collective MMORPG (Massively Multiplayer Online Role-Playing Games) play on gamers' online and offline social capital. *Computers in Human Behavior*, 27(6):2352–2363.

Author Vita

Personal information

Date of Birth	09.07.1981
Place of Birth	Zabrze
Citizenship	German

Employment history (Last 5 years)

Since June 2009	Research associate University of St.Gallen Institute for media and communications management Switzerland

Education (Last 5 years)

Since June 2009	University of St.Gallen Institute for media and communications management Switzerland Candidate for doctoral degree
February 2002 - June 2008	University of Mannheim Germany Business administration and computer science Diploma degree