METHODOLOGICAL CHALLENGES OF STUDYING SOCIAL MEDIA FROM THE PERSPECTIVE OF INFORMATION MANIPULATION

A Thesis
Presented to
The Academic Faculty

By

Bence Kollanyi

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Public Policy
in the School of Public Policy

Georgia Institute of Technology

August 2014

METHODOLOGICAL CHALLENGES OF STUDYING SOCIAL MEDIA FROM THE PERSPECTIVE OF INFORMATION MANIPULATION

Approved by:

Dr. Hans Klein, Advisor

School of Public Policy

Georgia Institute of Technology

Dr. Michael L. Best

School of Interactive Computing

Georgia Institute of Technology

Dr. Robert Rosenberger

School of Public Policy

Georgia Institute of Technology

Date Approved: July 1, 2014

ACKNOWLEDGEMENTS

First, I would like to thank Professor Hans Klein from the School of Public Policy, my advisor, for his generous support and continuous trust in my work. Dr. Klein's dedication, knowledge, and never-ending curiosity inspired me to explore a new research field as a Graduate Research Assistant and to write this thesis. I am grateful for the time we spent thinking together. By carefully reading the earlier versions of my thesis, asking the right questions, and providing insightful feedback, he has guided me throughout this exciting journey.

Dr. Klein has also introduced me to a smart and innovative research community focusing on computer security. From this group, I would like to specially thank Professor Wenke Lee for his support and the opportunity to be part of a team working together on an NSF proposal with brilliant researchers across the country. I learned a lot about how to motivate and bring together people for a common goal. I wish to extend this acknowledgment to Professor Nick Feamster who provided early feedback to my work. I truly enjoyed working together with Dr. Lee and Dr. Feamster developing a common vocabulary for computer scientists and researchers with a background in public policy.

As a member of this community and a philosopher interested in technology, Professor Robert Rosenberger, contributed to my research with thoughtful comments.

I also wish to express my sincere thanks to Professor Michael Best. I am grateful for the years I spent in his Technologies and International Development lab and the opportunity to work together with the most diverse group of people. This community provided the opportunity to work with real data. I would like to acknowledge Amanda

Meng for our collaboration on social media and elections in Africa, and Thomas Smyth for sharing his love and knowledge of this wonderful continent.

Finally, I would like to thank my wife, Klara Benda, for her love, support and encouragement. I would not have started on this road without her.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The first part of the thesis gives a systematic overview and conceptual analysis of the literature on studying misinformation and disinformation in social media, with a special focus on research projects using large scale data obtained from Twitter and Facebook. The literature review gives a detailed overview of the scope of data collected by the various research projects; the means of accessing the data, which are rooted in the concrete socio-technical arrangement of the various platforms, and the type of analytical tools they apply. Furthermore, it also maps the various theoretical questions behind the research projects. The author of the thesis also gives his own definition of information manipulation and describes a conceptual model of information manipulation in the context of social media.

The second part of the thesis applies some of the insights from the literature review to a large Twitter data set collected during the monitoring of African elections. The analysis follows a qualitative approach and focuses on case studies created from specific incidents during the elections. Each of these incidents illustrates a special aspect of the problem of information manipulation in online social media.

CHAPTER 1. INTRODUCTION

The study of misinformation and deception in social media has produced a large body of scientific work in the recent years. This research problem involves a combination of the technical and social dimensions, which often requires a multidisciplinary approach, so it is not surprising that these research projects bring together scholars from communication studies, psychology, and computer science. In my thesis, I will build on this interesting and complex literature in view of obtaining insights for the analysis of a large social media data set.

The thesis consists of two distinct parts. In the first section of my thesis, I will provide a systematic survey and a conceptual analysis of the literature on information manipulation in social media, with a special focus on research projects using large scale data obtained from two social media platforms, Twitter and Facebook. The literature review gives a detailed overview of the scope of data collected by the various research projects; the means of accessing the data, which are rooted in the concrete socio-technical arrangement of the various platforms, and the type of analytical tools they apply. Furthermore, it also maps the various theoretical questions behind the research projects.

The purpose of this systematic overview of the literature is to look at the major themes and data collection in the literature, to compare the two platforms, Facebook and Twitter in this respect, and to identify the gaps in the literature. The following are the three major questions that I attempted to answer with respect to each of the papers reviewed:

1. How did the researchers collect the data from social media?

2. How did they interpret and analyze the data?
3. Finally, what were the major themes covered by the research?

In the second half of my thesis, I will apply some of the insights and research findings from the literature to the analysis of a large social media dataset collected as part of an effort to use social media to monitor national elections in four African elections. The analysis has followed a qualitative approach and I have created case studies around specific incidents during the election. Each of these incidents illustrates a special aspect of the problem of information manipulation in online social media. Furthermore, the incidents allowed me to have firsthand experience with social media data and test some of the findings I have extracted from the literature.

1.1. Defining misinformation and disinformation

Communication is a basic human desire. In communication, participants share information with each other. There are social norms which tell us that this information has to be accurate, true, and complete (Karlova and Fisher, 2013). Still, as Karlova and Fisher point out, there are situations where people share information which is false or incomplete. Misinformation is generally defined by the literature as inaccurate information, while disinformation is defined as intentionally false information. To understand the key difference between the two, we have to focus on the intention and not on the truthfulness of the information. In other words, in order to identify misinformation it is enough to look at the content of the message, but for disinformation we have to look at the intentions, or the “state of mind” of the person who is creating or sharing the information.

Nevertheless, both misinformation and disinformation can be (partially) useful for the receivers of the information. Because of this, informativeness is an important characteristic of all kinds of information. As Karlova and Fisher (2011) remind us “misinformation, albeit false, is still information and, therefore, can still be informative”. Contrary to the everyday use of the terms, disinformation should not be seen as a subset of misinformation. The authors illustrate this point with a simple example where a person tries to deceive another person by saying that a movie starts at 3:30 PM when it is supposed to start at 3PM. However, this misinformation can easily turn out to be true information if the theater projector gets broken, and the projectionist has to replace the light bulb, so the movie indeed starts with a 30 minute delay (ibid.). In another article, Karlova and Lee (2011) claim that misinformation can also be information open to multiple interpretations (i.e. ambiguous), it can be uncertain, or it can be too vague or unclear.

1.2. Forms of misinformation

In order to better understand the forms that misinformation can take, it is useful to look at the existing work which presents and compares the various types of misinformation. Besterman’s (2013) typology is a good starting point, because it both gives an overview of the related literature and lists various types of misinformation. Besterman discusses the following four distinct types of misinformation:

(1) Rumor

(2) Conspiracy theory

(3) Fiction

(4) Simple factual error

A rumor (or sometimes referred to as a false rumor) is a claim which is based on false information. Besterman cites an important observation by Sunstein about the credibility of rumors. As Sunstein (2009) points out, rumors do not gain their credibility from “direct evidence(s) known to support them”, instead, they are perceived as credible information “because other people seem to believe them”. It follows that in order to create a false credibility, it is enough to show that there are a large number of people who believe the information. In the literature of persuasion, this is exactly what researchers call creating a false social proof (see Cialdini, 2007).

While rumors can be simple claims, conspiracy theories are manifested in the form of explanations. A conspiracy theory explains how certain people, usually powerful people, do certain things while they “conceal their role” (Sunstein & Vermeule, 2009, cited by Besterman, 2013).

Fictions were included in the list, because the literature has shown that people believe fictional information even if it is pretty obviously fictional. For example, Lewandowsky showed how the audience believe fictional stories because the information is coming from the media (Lewandowsky et al, 2012).

Finally, factual errors can also be considered misinformation, and they are often published by reputable sources. This also implies that people will likely believe them. Besterman does not address the question whether factual errors have been intentionally false (disinformation), or they are just not precise for unintentional reasons, e.g. mistyping a number in a news report. The errors made in Wikipedia, also cited by Besterman, have been documented to be of both kinds. Hermida (2012) claims that people recognize misinformation only if there are evidences and reliable counter information presented to them. However, this alternative information is often separated from or unconnected with the original information, especially in the context of Internet-based communication.

1.3. What is online social media?

The social web is a young phenomenon in the relatively short history of the web, which has recently enjoyed academic attention. Discussions have specifically focused on social media and social networks. In the introduction I will present the definitions given for each, and show how social networks and social media overlap. Overall, the definitions suggest that social networks may be seen as part of social media, which also includes social applications that do not rely on the personal network of users.

On the one hand, online social networks (OSNs) are Internet-based services where registered users experience, maintain, and visualize their social networks. OSNs are often based on real world relationships, but they can connect users who have never met in real life and still help the formation of an (online) community. On the other hand, basically all online social networks support some form of communication between the users, either direct communication between two persons or communication with more than one users at a given time.

There are multiple definitions for social network; the majority of them incorporate both social (or network-based) and communicational components. This duality is important not only for the definition of social media, but also in understanding how misinformation and disinformation can spread on social media.

One of the earliest definitions of the online social network can be found in a widely cited article by danah m. boyd[a] and Nicole B. Ellison (2010). Interestingly, the definition in this paper gives a limited view of social networks as it only focuses on the network aspect of the phenomenon, and puts the emphasis on users' ability to „articulate" and visualize their social networks:

> *"We define social network sites as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system."*

With more than 5000 citations[b] the paper is the de facto scholarly definition of social network sites. Subsequent definitions are often verbatim copies of the three points cited above. Meanwhile, the article itself has a more balanced picture of social networks. Ellison and boyd describe the particular ways of using online social networks, which also involves various forms of communication between the registered users of the social network. The authors mention a „wide variety of technical features" implemented in the social network sites, including the ability of sending messages to each other or leaving public comments on their friends' profiles. They also note that these private messages

[a] The author was born as Danah Michele Mattas, but she changed her name and decided to write it with only small letters. According to her website, this has to do with both personal and political reasons. Interestingly, on Wikipedia, her entry gets rewritten regularly; while some use the capitalized version, others keep rewriting it to the non-capitalized version, and vice versa.

[b] Based on Google Scholar which is a widely used academic search engine, but it is often criticized by distorting citation numbers by not only using peer review, academic publications. Still, the 5028 citations indicates that the paper is widely cited and probably had a strong influence on the literature on social media.

and comments are very popular on these sites. Later, the authors also specify other features, such as photo- or video-sharing and instant messaging, which are basic functions in many online social networks. At the same time, they claim that the „backbone" of such services are the profiles and the connections between users. The article also gives an overview of the history of social network sites, and it is often cited in connection with the claim that the SixDegrees site, started in 1997, was the first example of online social networks. boyd and Ellison describe how this early example of social network supported the creation of user profiles and friend lists. However, they also note that SixDegrees „promoted itself as a tool to help people connect with and send messages to others". Still, their original definition of social network sites does not include communication.

Aïmeur, Gambs, and Ho (2010) can be cited as the example of a definition that covers both aspects of social networks. They created their definition for 'Social Networking Site', the term they use for online social networks, with the following three elements: (1) users form a social network in the form of befriending or following each other, (2) there is communication between these users, and (3) they have the ability to restrict the access to the their content to „authorized users". This is a privacy centric definition, but other papers also mention controlled access, which lends a bounded character to social networks, as well as the need to register and create a profile before someone can interact with the users of the network.

In this document, I will use the term Online Social Network (OSN) to talk about any platform offering online services which support a public or semi-public communication based on a network of registered users. Semi-public refers to the ability of restricting access to the content for a group of registered users. By a network of registered users I mean users who are registered and have the option to create and store formal relationships with other users registered within the platform. In theory, it would be possible to build a platform which does not offer any services of communication outside of sending “friend requests” to other users, but I am not aware of such a site.

While some definitions discuss social network sites as a subset of social media, others use social networks and social media interchangeably. To avoid confusion, I find it more logical to say that OSN is a subset of social media. Furthermore, social media also includes technological platforms and services without a network of registered users or might even allow unregistered users to be engaged in communication, for example to post content or comment. BlogMask, for example, allows anonymous users to post blog entries without registering. Obviously, in this type of blog service, users can not establish connections and the site is not designed to maintain social networks. Other services, such as Anonyme.com, allow registration without a valid email address or any other personal information, and also provide some control over privacy, but do not support “social networking”.

The literature on social media, as opposed to the social network literature, tends to focus more on communication, and they usually pay less attention to the underlying network of users. For example, Kaplan and Haenlein (2010) defined social media as “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content”. The authors provide some details about the technological platform for web 2.0

with concrete examples like Really Simple Syndications (RSS) and Asynchronous Java Script (AJAX), which were trending at the time of the paper's writing.

Another approach, exemplified by a publication of the Organization for Economic Cooperation and Development about user generated content limits social media to non-professional users (OECD, 2007). The OECD definition of social media has three elements – one talks about the platform / service used for publication, and the other two describe the content and the way the content was created. More specifically, a social media platform is described as a publicly available website or a social network, so based on this definition, social media is a wider concept than social networking, or at least it is not limited to social networking. The next element of the description claims that user generated content is the outcome of a creative effort. Finally, the definition suggests that this user generated content is not created by professionals, or at least the content creation is not part of their daily work. This might be true for most of the content published in social media, but social media also receives contributions from professionals, as the corporate world, politicians, and even researcher use these outlets for their purposes. Because of this trend, I do not find it useful to limit the definition to content created by "non-professionals".

Kim et al use the term social web and the authors include both social media and social network under this term. Their definition contains both the network (the formation of online communities) and the communication element. According to the authors' understanding, social web is also based on user generated content (Kim et al., 2010).

Finally, Russo et al (2008) used a much broader definition of social media which included services that "facilitate online communication, networking, and/or collaboration" (p. 22). Their paper puts more emphasis on the communication part, and the authors note that „social media technologies are designed primarily as network communication tools (unlike telephone or email, which are first and foremost tools of one-to-one messaging)." (p. 22.) This is an important observation which partly explains why social media (and its subgenres) should be considered important communication tools.

1.4. Types of online social media

For a more detailed break-down of social media platforms we can revisit Kaplan and Haenlein's paper (2010), where the authors distinguish the following six types of social media:

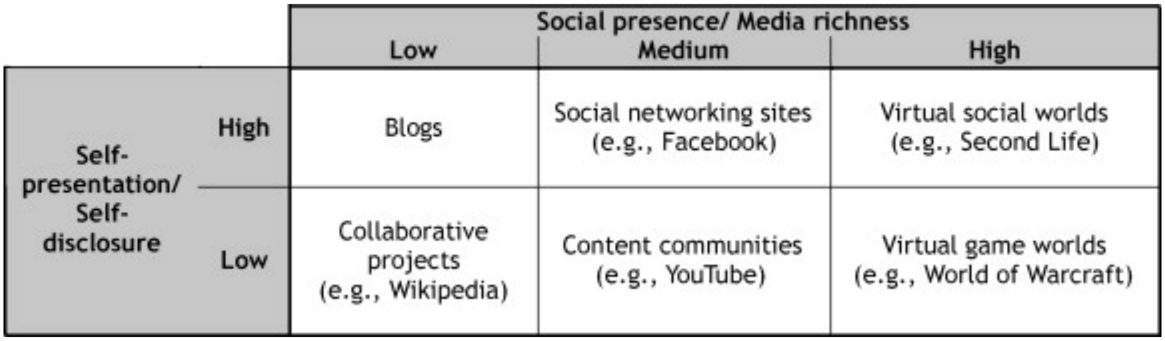
(1) collaborative projects (e.g. Wikipedia, OpenStreetMap)
(2) blogs (e.g. Blogger, WordPress)
(3) content communities (e.g. Flickr, YouTube)
(4) social network(ing) sites (e.g. Facebook, LinkedIn)
(5) virtual game worlds (e.g., EverQuest, World of Warcraft)
(6) virtual social worlds (e.g. Second Life, SmallWorlds)

Gundecha and Liu (2012) suggest another five categories:

(7) microblogs (e.g. Twitter, Sina Weibo)
(8) social news (e.g. Digg, Slashdot)
(9) opinion and review sites (e.g. Yelp, Amazon)
(10) social bookmarking (e.g. Delicious, CiteULike)
(11) answers (e.g. Yahoo! answers, WikiAnswers)

Kaplan and Haenlein (2010) also present a two dimensional classification, based on the intensity of self-disclosure and the richness of the medium characteristic of the types of social media platforms (Figure 1.1).

Figure 1.1: Types of social media based on the level of self-presentation and social presence, following Kaplan and Haenlein (2010)



		Social presence/ Media richness		
		Low	Medium	High
Self-presentation/ Self-disclosure	High	Blogs	Social networking sites (e.g., Facebook)	Virtual social worlds (e.g., Second Life)
	Low	Collaborative projects (e.g., Wikipedia)	Content communities (e.g., YouTube)	Virtual game worlds (e.g., World of Warcraft)

While in some cases it is not easy to decide which platform is more designed for self-disclosure, for example content communities, such as YouTube and its channels, are often more personal than blogs or even most user profiles on Facebook. However, it is still useful to think about both self-disclosure and social presence as two important characteristics of the social media.

In the following section, I will discuss misinformation and disinformation in social media. Self-presentation and social presence will be relevant for both. One way to spread disinformation is to trick users to believe that the information is coming from a source they can trust. In attempting to counter misinformation, platforms where self-presentation is less important can have an important role.

1.5. Misinformation and disinformation in social media

One form of communication in social media is (publicly) sharing information in the form of written messages and comments, pictures, or video messages. The information might come from a friend, a trusted source, or a complete stranger, who is an unknown user of an online social network. How can one decide whether the information is true and what are the clues one can rely on to make that decision? Furthermore, when it comes to assessing the validity of a piece of information that reaches us, is social media different from traditional media or face-to-face meeting with other people?

Before I get into the description of the two concrete platforms, Facebook and Twitter, and the discussion of various forms of information manipulation using those platforms, it is useful to review and compare traditional media and social media from the perspective of the clues that are available to users for detecting manipulated information.

Comparing journalism with social media, Hermida (2012) describes how social media has changed the process of determining which piece of information is reliable and which is not. Citing Brennen he suggests that facts are messy and cleaning them relies on interpretation. This process, according to him, "traditionally took place in newsrooms, away from the public eye, as journalists considered conflicting reports, weighed up incoming information and made decisions on what to publish." Similarly, Lewandowsky and his colleagues point out that the old "gate-keeping" mechanisms are not in place anymore; in the social media, there are no professional editors who follow the professional norms and traditions of journalism to double check the content before it gets published and (Lewandowsky et al., 2012). They argue that in social media, users have to devise their own strategies for identifying misinformation. They notably points out:

"Misleading information rarely comes with a warning label" (p. 111). It follows from this that users have to come up with good heuristics to identify misinformation. Misinformation can be detected either by judging the credibility of the source or by looking at the content itself. Still, most users have problems with identifying misinformation unless there is a clear retraction on behalf of the source, or they have access to contradicting information which is more convincing.

1.6. Information manipulation in the social media

Literature on social media has overwhelmingly relied on the terms of misinformation and disinformation to engage with the problem of false information. In the following section, I will discuss the limitation of these terms for the purpose of studying social media. I will define the term *information manipulation,* and I will argue that it helps us analyze and explain a broader scope of problems.

In the context of social media, the term *information manipulation* refers to any attempt to deceive or manipulate other users by creating false messages or by manipulating meta-data in the system. This can involve both direct and indirect forms of manipulation, as I will discuss in more detail below.

While social media messages can contain factual information which, in theory, can be falsified, they can also be entirely composed of opinions or ambiguous information. In many situations, the truthfulness of a message cannot be examined or questioned and the message cannot be falsified. Meanwhile, even non-factual messages can change other users' behavior or influence how they think about particular questions. Often, such messages are distributed on a large scale. If a user claims that the President of the United States was not born in the United States, the information can be questioned and there might be reliable sources outside or inside the social media platform which can falsify it. However, if the same user claims that the President is the worst president of the United States, and no one should vote for him, the information in the message cannot be falsified. One can argue against it, but it is not a piece of fact-based information. On the other hand, manipulators can create hundreds of fake users who spread the same opinion. By this action, manipulators can create a false sense of consensus behind a particular view or opinion, in other words, they can pretend that a large group of actual users think along the same lines.

Earlier, we could see that the term misinformation suggests that the content of a message is false, while disinformation indicates that the person who is spreading the information is trying to purposefully deceive others. Still, neither of these terms covers the cases of manipulation which involve the creation of fake circumstances and related meta-data. This can involve faking the author of a particular message or pretending that a large number of people share an opinion, such as when they "like" a contribution or a source.

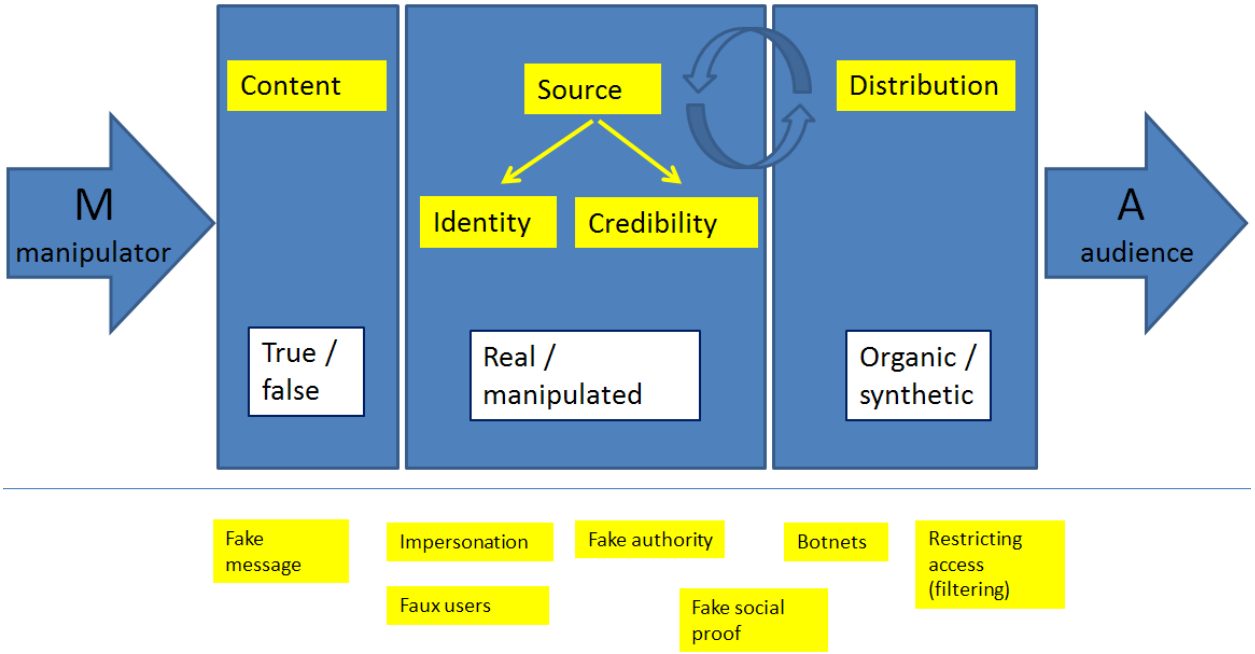
A direct attempt refers to manipulated or false messages and meta-data created to directly manipulate other users. Indirect information manipulation, on the other hand, indirectly manipulates users by attempting to influence how information intermediaries present the information. For example artificial buzz can be created around an existing message, or the search engines may be manipulated toward particular search results.

1.7. A model for understanding information manipulation

Information manipulation is based on the willingness of the recipients to accept the truthfulness of a piece of information. In online and offline situations alike, it takes more than just looking at the content of a message to decide whether it is believable or not. In real life, face-to-face situations, people are "trained" to identify credibility, but the available literature suggests that it is much harder to do the same online (see for example Wallace 1999). One can look at the clothes of the person, the way how he or she speaks, and analyze the other person's non-verbal communication. On the internet, these clues are not available, but users are still not limited to only look at the content of the message. It follows that decisive communication has a different toolbox to manipulate information while Internet users, technology companies, and researcher can also use a different set of cues to detect manipulated information.

To make sense of the creation and detection of false beliefs in the context of online social media, I have created the model in Figure 1.2. The model shows how Internet-based information manipulation can deceive users by fabricating false messages, faking the identity of the source, or giving false credibility to the source of the message. Digital technologies also allow participants to look at how information is being spread, who participates in sharing it, and who can access the specific content. Thus, the model also indicates avenues of detecting information manipulation.

Figure 1.2: A model of information manipulation



In this model, there is a manipulator whose aim is to change what the audience believes or how it acts. The manipulator's message, in theory, can be both true and false. In order to reach the audience in the context of social networks, the manipulator could use faux users, fake authority or impersonate other users by either registering fake users or getting access to existing user accounts. With respect to distribution, it is possible to differentiate between organic distribution (real users share the message) and synthetic distribution, in which case faux users, such as software bots, distribute the message.

CHAPTER 2. REVIEW OF THE TWITTER LITERATURE

2.1. Methodology for collecting articles

Chapter 2 and 3 present a systematic literature review of information manipulation on Twitter and Facebook, respectively. To review the scholarly approaches of information manipulation and its detection in social media, I followed Webster and Watson's method (2002) for creating systematic literature reviews. The most crucial part of the process is identifying the relevant literature.

Webster and Watson suggest a 3 step process to achieve a "structured approach" for identifying the relevant articles for the literature review (ibid.). The process starts with scanning the leading journals and conference proceedings by using a set of keywords in order to identify key sources. The idea of identifying key sources is based on the assumption that a scientific discourse is centered around key journals and conferences which provide a platform for discussion. In this part of the literature review, I followed an iterative process by starting with a relatively limited set of keywords, which I gradually extended based on the vocabulary used in the titles and abstracts of the relevant articles I found through academic search engines. During this phase, I also put together a list of important journals and relevant conferences and systematically searched the archive of these sources. (See both the list of keywords and sources in Table 2.1 below)

The second step suggested by Webster and Watson is going backward by sorting through the citations in the body of literature identified in step one. This step is based on the assumption that both conference publications and journal papers position themselves in the existing literature, and again there is an ongoing discussion in the scientific arena about the specified topic.

The last step is going forward by using electronic scholarly databases to find articles which cite the literature collected in the first two steps. Webster and Watson suggest using Web of Science to find new articles which cite the key papers (ibid.). Google Scholar is also a good way to scan for articles citing the paper. In this step, I used the keywords identified in step one to select articles which not only expanded the scientific discussion, but were also relevant for my research focus.

From the list of articles obtained from search engines and conference and journal archives I selected the relevant and important articles. To determine relevance, I scanned through the abstract of the papers to see if the articles were related to both information manipulation and social media (or the specific platforms of social media I wanted to study). Important articles are reviewed and referenced by other scholars, so in order to determine importance I set up the criteria of having at least five citations for the selected papers. To pick the articles which have more than five citations, I used an open-source reference management software called Zotero and a plugin developed by Anton Beloglazov (Zotero Scholar Citation) which connects Zotero to Google Scholar and extracts the number of citations for the articles in Google Scholar’s database.

Table 2.1: List of keywords and sources used in the literature review of Twitter

Keywords:	Conferences:	Papers:
Twitter Micro blog Spam Manipulation Misinformation Twitter API Credibility Trust, Trustworthy Propaganda Rumor	• International AAAI Conference on Weblogs and Social Media • ACM Web Science Conference • IEEE International Conference on Social Computing • Discovery Science • Computer Security Applications Conference • CSCW - Conference on Computer Supported Cooperative Work • WWW - World Wide Web Conference Series	Communication Quarterly Internet Research First Monday

2.2. Introducing Twitter

Twitter reportedly had 241 million active users at the end of 2013, and these users generated more than 500 million tweets in an average day (Goel, 2014). It is not surprising that almost all the reviewed papers' introduction section mentioned the high number of users and the consequences of the media platform's popularity. For example, Beck points out how Twitter, and other social media sites, have changed the way how Internet users communicate with each other and how the 140 character long messages have started to replace email use and phone calls (Beck, 2011). Similarly, Ratkiewicz mentions how the fact that Twitter has millions of users "offer[s] an opportunity to study patterns of social interaction among a population larger than ever before" (Ratkiewicz et al., 2011).

People do not only rely on social media in their personal communication, but they "have moved from scanning traditional mediums such as newspapers and television to using the Internet, in particular social networking sites, to find news" (An et al, 2012). With the changing media consumption patterns, television stations and print media outlets have experienced decreasing viewership and circulation numbers, and the news media itself has turned to social media and started to use these platforms.

By default, Twitter messages are public, accessible online, and searchable. This feature of Twitter in itself opens up many opportunities to do research on microblogging. Furthermore, from the beginning, Twitter has provided some level of access to these public messages to businesses and academia through the company's Application Protocol Interfaces (APIs).[c] Many researchers praise Twitter for its open data access policy. Researchers from the Oxford Internet Institute traced back the "rising attention" the platform received from the academia, and the numerous research projects and

[c] A web Application Programming Interface (API) specifies how computer programs, often third-party applications, can interact with a web service. For more information about Twitter's use of APIs, see appendix A in the end of this document.

publications focused on Twitter to the relative ease of accessing large Twitter data sets (González-Bailón et al, 2012). Also, An et al (2012) suggest that it is much easier to do research on Twitter compared to traditional media because on Twitter "there is no need for required extensive surveys to gather the required data."

At the same time, the increasing popularity and importance of the platform as an information source also attracted various forms of information manipulation from spreading false rumors, selling twitter followers in illegal market places, to faking grassroots movements by mobilizing an army of centrally controlled faux Twitter accounts. The following sections of this review first categorize the relevant research projects by larger, dominant themes, then describe how various research projects collected large datasets and how these were used to arrive at the research findings.

2.3. Themes in the Twitter literature

In order to explore the variety of research projects and to create a concept map, I started my research by entering the combinations of a few general key terms to the Google scholar search engine, such as "information manipulation" "social media", "Twitter", and "API". As I downloaded the relevant contributions, I expanded the list of keywords to collect papers about specific research themes. This approach, combined with looking up the references and the papers cited by the selected papers, helped to have a better coverage of the relevant literature.

The next step was the analysis of the papers. I created a table of the papers containing the general theme of the paper, the method for data collection, the scope of the data and the statistical methods the researcher(s) used to analyze their data. The following section describes the major themes covered by this body of literature and gives specific examples for each theme.

The research projects were centered around five distinct research themes. Almost all papers fell under one of the following categories:

- credibility assessment;
- crisis communication;
- information security;
- politics and propaganda; and
- prediction power of the data.

Most individual research papers focused on one specific research problem and only described the literature related to their particular themes. Before getting into the details of these research themes, it is worth pointing out the paper by Bruns and Stieglitz, which surveys a wider range of themes in its related studies section (Bruns and Stieglitz, 2012). The paper, which focuses on general communication patterns on Twitter, does not attempt to provide a systematic literature review, but it presents a useful overview of Twitter-related literature. Bruns and Stieglitz describe the literature on Twitter-related communication and group the previous research findings around the following three distinct topics: politics, natural and human disasters, and brand-related communication and entertainment.

In the following, I will present a description of each of the five themes introduced earlier, and discuss why these topics have special relevance for Twitter.

2.3.1. Credibility assessment

Today, Twitter is available on various mobile devices including smartphones and tablets, so users can search for locally relevant, recent information even if they are away from their computer. Even without internet connection, users can send updates (tweets) via SMS. As a consequence, Twitter has evolved from an “opinion sharing medium among friends” to a platform where people can share information as they experience it or disseminate current news in real time (Gupta and Kumaraguru, 2012). On this account, a wide consensus has been emerging on the increasing importance of Twitter as a news source. It has also been pointed out that Twitter and other social media platforms have influenced the way traditional media cover certain events. For example, breaking news has been typically reported by the news media after the social media had already reported about the events. In many cases, newsworthy events are first picked up by Twitter users, simply because non-professional eyewitnesses start sharing reports in 140 character long tweets or in photos or videos before the mainstream media reach the scene (Diakopoulos et al., 2012). For example, soon after the Boston marathon bombing happened, eyewitnesses started to tweet about the event from the scene; both professional media and the police later used these messages to investigate what happened (Rogers and Luckle, 2013).

The larger base of users and the "information overload" on the platform, however, also make it hard to select relevant and trustworthy information. As I mentioned earlier, there are about 500 million tweets shared on an average day on the platform. However, the sheer volume of information available through Twitter is not the only challenge as we will see. Credibility assessment, or the trustworthiness of the information on Twitter, has been heavily researched and presented in many papers. The term credibility has been used in the context of the trustworthiness of a particular source, i.e. a Twitter user, or to describe the information published on Twitter itself. Some scholars have examined credibility in the context of finding reliable news sources on Twitter (Castillo et al, 2011; Benevenuto et al., 2010; Diakopoulos et al., 2012), while others have discussed ways of modeling the credibility of messages by classifying the content of tweets (e.g. Kang, O'Donovan, and Höllerer, 2012; and Qazvinian et al., 2011). Credibility has also often been discussed in the context of high impact events, such as human and natural disasters where the truthfulness of information is especially important (Gupta and Kumaraguru, 2012; Mendoza et al, 2010; Thomson et. al., 2012).

The literature also mentions two aspects of Twitter which make the platform more sensitive to information manipulation. Twitter users, especially inexperienced users, are missing clues to assess credibility that they are used to in real world contexts (Castillo et al, 2011). The authors do not specify these clues, but the social psychology literature mentions numerous clues which help to establish credibility in face-to-face inter-personal communication (Wallace, 1999). Another paper compares Twitter with traditional electronic and print media and claims that it is harder to evaluate credibility on Twitter. The main reason is the decentralized nature of Twitter; while traditional media has been offering only a few and usually known sources, the information on Twitter comes from many different sources, which are also often anonymous (Gupta and Kumaraguru, 2012).

2.3.2. Crisis communication

During emergency situations, microblogging can have an important role in both disseminating information to people who are affected by the situation and receiving live reports from users on the field (eyewitnesses). If the local infrastructure is functioning and people have access to Twitter on their mobile devices, information can reach people faster compared to traditional news media.

Castillo and his colleagues focused on the “time-sensitive” use of social media and studied the role of Twitter communication after the 2010 earthquake in Chile. The research focused on how Twitter was used to distribute time-critical information right after a natural disaster (Castillo et al, 2013). According to the report, Twitter was used to share practical information, such as current road conditions or the location of the functioning gas stations. The authorities also tweeted tsunami alerts for the people who could be affected by the water waves caused by the earthquakes (ibid.).

Misinformation can spread faster during a crisis, and it is harder to detect than in normal situations. Studies have shown how local Twitter traffic in general peaked after significant events, and how certain keywords became widely used, such as the

“#terremotochile” hashtag in the case of the 2010 earthquake in Chile (Mendoza, Poblete, and Castillo, 2010). This phenomenon is related to two characteristics of crisis situations that I will call low visibility and high information needs. High traffic makes it harder to find both relevant and true information. Furthermore, information often comes from unverified, private sources as the literature about credibility pointed out. I call these two aspects together low visibility. On the other hand, quick access to information is a must for people directly affected by the disaster, as well as for their loved ones who are worried about them. People at the scene of the disaster need practical information, as they are worried about their immediate safety, while Twitter users who are not at the field feel an urge to help and inform others by sharing news coming from the survivors. I call this aspect of the situation high information need. The combination of a sense of low visibility and high information need can contribute to the quick spread of unverified information.

There is a substantive body of literature studying the spread of rumors on Twitter after natural disasters. Rumour has been widely studied in social psychology since the Gordon Allport and Joseph Postman’s path breaking research in the late 1940s. Their The Psychology of Rumor has since become a classic (Allport and Postman, 1947). Interestingly, many of these articles did not include a proper definition of the term of rumor. Qazvinian and his colleagues on the other hand define rumor in connection with Twitter as a “statement whose truth-value is unverifiable or deliberately false.” The authors also cite DiFonzo and Bordia’s definition:

> “*unverified and instrumentally relevant information statements in circulation that arise in contexts of ambiguity, danger or potential threat, and that function to help people make sense and manage risk*”
>
> (DiFonzo and Bordia, 2007: 13).

2.3.3. Information security

The information security literature describes the approach of using large Twitter datasets to develop methods to automatically detect various forms of manipulative Twitter feeds, such as spam (e.g., Benevenuto et al., 2010; Chu, Widjaja, and Wang, 2012; Irani et al., 2010; Lee et al., 2012; Song, Lee, and Kim, 2011; Thomas et al., 2011), phishing (Beck, 2011), automation of Twitter accounts and social bots (Wagner et al., 2012; Chu et al. 2012), fake grassroots movement (Ratkiewicz et al., 2010; Ratkiewicz et al., 2011).

Spam takes many forms on Twitter. The most well-known forms are tricking users to follow a Twitter account and distributing messages in this manner, or simply using hashtags or keywords which are popular on Twitter, but unrelated to the content of the message. Thus, the Twitter spam research has used various clues to detect spam. Some researchers have focused on the text of the message and employed sophisticated text analysis to detect spam. For example, Shekar has experimented with classifying pharmaceutical spam in a Twitter corpus based on a set of manually labeled tweets (Shekar, 2010). Other researchers have focused on how messages spread and using Twitter's APIs to collect information about the sender and its social network (e.g. data of registration, number of followers). This line of research has found that spammers often utilize social relationships to build trust and distribute messages through a network of "friends" or followers (Chu, Widjaja, and Wang, 2012). The same research project has shown that diffusion of spam is often characterized by an early burst of traffic which is different from the distribution patterns of most legitimate marketing campaigns where users slowly pick up a topic or a message (ibid.).

Benevenuto and his colleagues point out that when a major event occurs, Twitter has a high peak in related tweets, and this sudden increased traffic can be used to distribute unsolicited messages or spams (Benevenuto et al., 2010). Spammers simply have to attach trending keywords and hashtags to the message to attract traffic (Irani et al, 2010).

Stringhini and his colleagues studied the underground economy of buying and selling Twitter followers (Stringhini et al, 2012). The authors introduced the concept of Twitter Account Markets where users can buy followers measured in the thousands of users. They also described the method of getting access to victims' Twitter accounts, as well as various ways that clients of these illegal businesses can spread promoted messages or spams on Twitter. In many cases victims are tricked into trying out a service, and they themselves give access to their account. In return, they get some new followers for free "recruited" from fellow victims. An interesting aspect of the research's findings is the fact that the Twitter APIs can be equally used to support research projects, to create valuable new interfaces and Twitter mash-ups, as well as to access victims' accounts and automatically follow other users.

2.3.4. Politics and propaganda

As Ratkiewicz has pointed out, social media is a battlefield for modern politics (Ratkiewicz, 2011). To understand more about this battlefield, researchers from the Indiana University developed Truthy, a tool to visualize the spread of information on Twitter, and used it to study the online communication and interaction between the communities of different political groups during the 2010 US midterm election (Conover et al, 2011). Based on the same dataset, they studied how political memes spread on Twitter. By using various hashtags, information about mentioning particular users (mentions) and a list of manually picked keywords, the researchers identified the themes which generated large traffic, and visualized how the information contained in these memes spread through the network. The research focused on how legitimate memes can be separated from memes which are propagated using astroturfing, which is the creation of a large number of faux Twitter accounts controlled by one user.

While Truthy was used to detect astroturfing in a political context, researchers from Georgia Tech studied another form of propaganda, the extremely active users on Twitter who “push the agenda” of a single political group (Lumezanu, Feamster, and Klein, 2012). This study is based on a dataset of Twitter messages posted with hashtags related to two current US political events. Based on Herman and Chomsky, the researchers described users who spread propaganda on Twitter as repeaters and amplifiers. Repeating, in the terminology of Twitter, is retweeting and sending messages with more or less the same content over and over again. The study has shown that propagandists, also called hyperadvocates in the paper, acted as repeaters and posted a large amount of retweeted content. Also, these users retweeted messages quicker than “natural” users and collided more often which means they posted the same content multiple times with only slightly modified formats.

Researchers from the University of Cambridge designed a tool for visualizing political bias on Twitter (An et al, 2012). Most major traditional media have a strong presence on Twitter and users can follow their account. From a researcher's standpoint, it is not only possible to look up overlaps between the “readership” of the different media sources, but Twitter also records how these groups of users interact with each other. In other words, researchers can tell how much the different political camps talk to each other, e.g. how often subscribers of a conservative source follow a person who is subscribed to a source which leans toward the political left.

It has been suggested that Twitter and social media in general have changed how Internet users access political news and have discussions about politics. During the Arab springs, for example, internet users reported about the brutality of the police and other atrocities even when state-controlled mainstream media did not cover the events (Howard and Hussain, 2011). Social media, as a free online platform, also has an important role in organizing civil movements. It has shaped how new movements occupied urban spaces, and led to the creation of a new “hybrid space” (Castells, 2012). A good example for this is Egypt where protesters used Twitter to coordinate meeting times or discussed what kind of supplies the people on the streets needed, and also reported about the violence on Tahrir Square (Starbird and Palen, 2012).

However, social media can also be used to spread propaganda or misinformation on a massive scale. The literature on Twitter has studied information manipulation in the context of civil uprisings. For example, a study focusing on the Egyptian and Tunisian revolution pointed out how difficult it was to tell whether a piece of information was truthful or a simple rumor (Lotan et al., 2011). It was often impossible to tell the source of the information or determine whether the information had been modified during retweeting (ibid.). Starbird and Palen also studied information spread on Twitter during the 2011 civil uprising or revolution in Egypt. The research focused on the retweeting mechanism of the platform, and used tweeting and retweeting behavior to determine which tweeters were local users, producing “original information” and who were only making firsthand reports (Starbird and Palen, 2011).

2.3.5. Prediction power of the data

Kalampokis and his colleagues give an extensive overview of a special segment of the social media-related literature, which focuses on the predictive power of various social media platforms (Kalampokis et al, 2013). As their review points out, expressing personal opinions and thoughts are important aspects of social media. Researchers and the business community are equally interested in exploring the opportunities in the analysis of this rich text, hoping to learn about the wider context by studying user behavior on these online platforms. Microblogs, including Twitter are one of the most popular platforms for studying the predictive power of social media. In fact, the systematic literature review did identify more research interest in this platform than any other source (ibid). For example, Lampos and Cristianini used Twitter data to track and predict flu pandemics and estimate the levels of rainfall in a specific location based on geo-coded tweets (Lampos and Cristianini, 2010a, 2010b, 2012). The review also lists more than 15 papers where the authors used Twitter data for predicting the outcome of political events, such as elections, as well as monitoring natural disasters and the outbreak of pandemics (Kalampokis et al, 2013).

2.4. Data collection on Twitter

2.4.1. Open access to data

Twitter opened up its platform to developers by releasing its API right after starting its service back in 2006 (Stone, 2006). Kevin Makice, a computer scientist and the author of O'Reilly's handbook on the Twitter APIs, called this move brilliant, and he claims that the APIs greatly contributed to the success of the company, especially in the early times (Makice, 2009).

In practice, the Twitter API facilitates access to traffic (data) of the site by supporting reading and writing functions. The API also allows third party developers to build tools based on Twitter's platform. Makice calls this latter phenomenon cocreation. In this case, Twitter provides access to some of its data and a platform to interact with its servers, and other companies develop new ways to interact with the service. According to Makice, this is the most important property of the Twitter APIs.

> *"…Twitter API is not found in the nuances of its syntax, but rather in the imaginative and prolific cocreation it inspires."*
>
> (Makice, 2009).

Indeed, Twitter has created a large ecosystem around its microblogging service by providing access to its platform. In five years, Twitter has officially registered over a million Twitter based applications developed by a community of 750,000 developers worldwide (Twitter, 2011). Also, in 2011, Twitter announced its new developer site where programmers can find all relevant Twitter API documentation as well as examples to use the APIs. Furthermore, the company expressed its commitment to support third party developers by connecting them to its internal developer team, and let them create new tools and services together.

A good example for expanding Twitter's functionality and having access to the platform is the development of Twitter's own search function. Makice wrote the following description about it:

> *"In spring 2008, a third-party company named Summize started tracking the public timeline of Twitter, gathering content to make the tweet corpus searchable. It was widely successful, often up and useful even when Twitter was down."* (Makice, 2009).

Later, Twitter decided to buy the company, Summize being its first acquisition back in 2008. Twitter also decided to hire most of Summize's employees and offered the position of Director of Engineering and Operations to its CTO, Greg Pass (Arrington, 2008). The service built by Summize has become Twitter's internal search solution and got the search.twitter.com subdomain.

Based on the reviewed articles, there are three major ways to get Twitter data. Technically, data collection is almost exclusively happening through Twitter's Application Programming Interface (API), but while some projects develop their own tools to access Twitter data, other projects either use third party software solutions to access the data or use data sets previously collected. I will label the first approach as direct access and label the other two as indirect (or mediated) access. This latter approach can also be described as using a corpus. Twitter's set of APIs both provide real-time access to tweets and historic access to data. However, both real time and historic data is limited by Twitter as I will describe later.

2.4.2. The three types of Twitter data

Twitter's APIs have been designed to accept a set of commands (e.g. requests) which rely on certain protocols (e.g. authentication) to interact with Twitter. The website of Twitter (Twitter.com) provides an interface where users can access basic functionalities of Twitter, such as posting a message (tweeting), read messages from other users (e.g. getting tweets from other users or searching), as well as use special twitter-specific actions such as sharing (re-tweeting) or following other users. Twitter's APIs are designed to support the same activities in the context of using a computer software to access the data, in other words, Twitter created a tool for third party applications to read (get) or write (post) messages in the name of authenticated user(s). The same APIs can be used to gather data from Twitter.

After carefully studying Twitter's documentation of its Application Programming Interfaces (APIs) and sorting through the scientific publications about relevant research projects, I distilled three distinct types of access to Twitter's data (see Table 2.2). All three type of access uses Twitter's API to get the data from Twitter, but while researchers using "Direct Access" program the API by themselves, the other two forms have a third party involved in the process. In the case of "Mediated access" the third party actor provides indirect or mediated access to Twitters data in the form of a Twitter application. A good example for this type of access is Topsy which provides real time search as well as analysis for Twitter and Google+ data. The company behind Topsy can extract information from Twitter and sell the data the company collected to commercial businesses or researchers. The third type of access is using data collected during previous research. In some case, researchers make complete datasets available to download and use for non-profit purposes. Recently, Twitter has started to restrict sharing the downloaded data and changed its rules for using the API. As a consequence, the University of Edinburgh had to remove the first large public Twitter data set from its website. Sasa Petrovic, one of the researchers who participated in the original data collection, has published the following announcement on his official website.

> *"We are unfortunately unable to continue distributing our Twitter dataset mentioned in the paper The Edinburgh Twitter corpus, due to a request by Twitter. Please do not contact me asking for the dataset as I cannot give it to you. Consider instead downloading the new Twitter FSD corpus."*

The new dataset is an interesting way to bypass the restriction, because researchers published only a list of unique IDs for the Twitter messages in the original Edinburgh corpus. Using these IDs, researchers can download the dataset from Twitter by using its API. Researchers at the University of Edinburgh published a detailed guideline to download the data.

Each of the forms of data collection discussed above have their comparative advantages and disadvantages. Comparing research projects using the three approaches, we can identify differences in the following dimensions: speed, cost, flexibility, skills needed to access data, "quality" of the data. Table 2.2 below summarizes some of the most prominent differences.

Table 2.2: Types of access to Twitter's data

Data source type	Advantages	Disadvantages
Direct access	flexibility, full access to metadata, "research credit" for collecting it, open source programs to store	needs CS skills, slow, large raw data
Mediated access	no (or less) need to code, quicker, extra services (e.g. visualization)	less metadata, less flexibility
Corpus	cleaned, description, extra information / analysis, readily available	only historic, anonymized, already published, legal problems

I suggest differentiating between historic data and real-time data. Even though Twitter is designed to publish short updates (tweets) about current happenings, tweets posted in the past are also accessible to a certain degree. For example, user profiles contain previously posted tweets, but these are only accessible through Twitter's API in a limited fashion. Twitter provides access to its historic data through both its web based search function and its Search API. Again, this only gives you a limited number of messages, and according to Twitter, you can access the most relevant ones.

Finally, Twitter corpuses should be categorized as historic data, even if the researchers were collecting real-time data at the time of the data collection. The reason behind this is that usually corpuses are only shared with other researchers after the collector finished data collection, cleaned the data, and often only after having published on the basis of the corpus, which means it is only accessible after many months.

Real time data in the context of Twitter means low latency access to recently posted messages. Both REST API and the Streaming API can provide access to real time data. However, for large scale, low latency data collection, the Streaming API works usually better. From the three type of data source, direct access and mediated access can both provide access to real time data. For example, by using Twitter's API, a website can filter and collect all the messages related to computer and video games. As I explained earlier Twitter corpuses cannot provide access to real time data by definition.

2.4.3. Scope of the data

This sections overviews the practice of extracting data about microblogging from Twitter for the purpose of academic research, more specifically, it surveys the timescale and the size of such data collections.

The number of tweets collected highly depend of the research topic and the method of data collection, however, it worth to note that about 80 percent of the projects I reviewed collected and analyzed more than a 100,000 tweets, and more than half of them analyzed millions of Twitter messages. There were three research projects where the number of collected tweets was larger than one billion; two of them studied information diffusion on Twitter, while the third one focused on Twitter spam.

There might be at least three factors behind the large numbers. First, the collection of tweets and the storage of these 140 character long messages are relatively cheap. Secondly, the statistical methods employed by the researchers need large datasets, or at least they can provide more robust results with more data. Finally, there seems to be a common practice within the research community to collect large number of messages.

On the other hand, there are a few factors which limits the data collection. There are two basic limiting factors of data collection: the time data collection needs and the size of the data collected. The data collection itself, partly due to the limits of the Twitter APIs, is time consuming, and research projects do not have unlimited time for data collection. Furthermore, even though the statistical analysis can become more robust with more data, there is a point where additional data does not change the outcomes largely.

Finally, the hardware and software used to analysis the data can limit the size of the data collected.

Furthermore, in some cases there are “natural” limits of data collection which follows from the focus of the research in a self-explanatory manner. For example, Mendoza and his colleagues studied the use of Twitter in crisis situation and they focused on the communication happened in the aftermath of the 2010 Earthquake and tsunami in Chile (Mendoza et al., 2010). The earthquake which reached a magnitude of 8.8 on the Richter scale happened on the February 27 2010, so it made sense to limit the analysis to data between the earthquake and March 2 2010. In another case, researchers from the Indiana University focused on Twitter communication before a political event, the 2010 U.S. congressional midterm elections, and collected data from mid-September to the November election (Conover et al., 2010). During this time they collected 355 million tweets, and there was no reason to continue data collection after the election. Finally, Starbird and Palen studied retweet behavior during the 2011 Egyptian Uprising (Starbird and Palen, 2012). A few days after the protest started in Cairo and throughout the country, on January 27, the Egyptian government shut down the Internet and restricted SMS traffic (Tsotsis, 2011). The researchers started to save Twitter data from February 2, “hours before Internet access was restored in Egypt”, and they collected all local tweets until a few days after president Mubarak stepped down (Starbird and Palen, 2012). In all of the above examples, researchers decided to limit their data collection to a period around a specific historic event.

2.5. Data about data - metadata on Twitter

Twitter data is much more than a collection 140 character (or shorter) messages. There are multiple metadata fields associated with each and every tweet, and most of them are accessible through the Twitter APIs. A typical tweet saved by Twitter has for example a timestamp which contains the time of posting the message, a user id which indicate the user who created the message, and it often has geo-location to identify the place of posting. Furthermore, Twitter save the hashtags used in the tweet, the IDs of other users who shared (re-tweeted) the message, information about replies, and so forth. A list of selected meta-data is available in Table 2.3.

Table 2.3: Selected list of Twitter meta-data (based on Twitter API parameters)

status	The text of your status update, typically up to 140 characters
place_id	A place in the world
lat / long	The latitude / longitude of the location this tweet refers to.
id	The numerical ID of the tweet
lang	language of the tweet
follow	number of users follow the creator of the tweet
track	Phrases of keywords
retweets	Tweets which are retweeted by the user
followers_count	Number of followers
time_zone	Time zone
friends_count	Number of friends
source	Source of posting (web, specific Twitter app, mobile phone, etc.)
favorite	Marked as favorite by other users
created_at	Time of creation
URL	The URLs in a message
user mentioning	User names after @ character
hashtags	Words after # character
sender / recipient	ID of the sender and receiver in the case of direct messages
sentiment	Positive :) / negative :(/ Question ? content
reply	The message is a response to another user

It is possible to access all this meta data through a third party application (mediated access), but in general, direct access gives users the most flexibility to choose which metadata they want to access from Twitter’s database. In other words, indirect access to Twitter’s data often limit the amount and type of information a researcher can access.

Furthermore, due to legal constraints and research ethics, Twitter corpuses often contain only anonymized data. For example, Bifet and Frank decided to use the then publically available Edinburgh corpus for their sentiment analysis, and they wrote about the limited data included in the corpus. The corpus only contained anonymized user names, and for each tweets it had a time stamp and the method the message was posted on Twitter (Bifet and Frank, 2010). Similarly, when Ghosh and his colleagues collected data about 41,352 suspended spammer Twitter accounts and about users who were connected to these accounts, they decided to share the data set with the research community only after they anonymized the data (Ghosh et al., 2012).

Using data collected by other researchers or applications, however, can also add extra information to the data. For example, a corpus can contain sentiment analysis or a network analysis which is not readily available if a researcher decides to access Twitter data directly by using Twitter’s own API. Furthermore, researchers usually share cleaned data, which makes easier to carry on further data analysis.

Even though we could see that using third party applications to access Twitter’s data and Twitter corpuses shared by other research projects have their own advantages, the overwhelming majority of the research papers describe the analysis of large Twitter data sets are based on data collected by the authors themselves. From the 60 research papers I reviewed, almost 40 used Twitter’s API to gather data.

The data collection is interesting for at least two reasons. One is the technical aspect. The authors collected tweets through Twitter Search API which limits the number of queries per hour, thus only gives a limited access to the history of the public Twitter stream. The researchers, however, solved this technical problem by collecting matching tweets for a longer period of time and filtered out the duplicates later. The other interesting aspect of the project is the use of external sources to collect and analyze Twitter data.

Real-time data collection is usually more effective for creating a relatively complete dataset than trying to reconstruct the Twitter stream post hoc. In real time data collection researchers have to simply follow the right hashtag or keyword with the Streaming API. This technical aspect of the platforms, as Bruns and Stieglitz point out, forces researchers to react on current event quickly (Bruns and Stieglitz, 2012).

Users of Twitter developed certain conventions and adapted particular ways of communication to maintain online conversations, such as retweeting hashtags followed by keywords to keep track of a topic. After these conventions were widely accepted by users, the central infrastructure of Twitter started to support them, and today users can search for specific hashtags to find conversations about certain topics (Bruns and Stieglitz, 2012). The same hashtags are similarly useful to researchers who study online (Twitter based) communication around specific topics. Bruns and Stieglitz differentiate between hashtags associated with breaking news or crisis like events and hashtags used to described previously planned events. While in the first case it is hard to pick the right keywords, the latter case helps researchers to follow certain hashtags. For example, Bruns and Stieglitz mention the #royalwedding and the #eurovision hashtags.

Furthermore, the authors compared the timeframes of “lifetime” of hashtags and the activity of various user groups. Bruns and Stieglitz contrasted long term hashtags with a big group of active users discussing the topic with short timeframe hashtags with a relatively diverse. In the latter case, there is a more even distribution of messages sent by the most active users and by the less active users. Accordign to the paper, twitter is used in a different way when users respond to a crisis or a breaking news story and when users use Twitter for “backchannel” information which sometimes supports the “shard experiencing”.

2.6. Data cleaning and data analysis

The review on predicting power of the social media mentioned earlier surveys data analysis methods and suggests a framework for data analysis (Kalampokis et al, 2013). This framework, with some adjustment, can be applied to other fields of social media data analysis. Thus, it is worth to go over the main steps of the process. The authors break down the analysis into two distinctive phases; First, every research project have to focus on the data conditioning phase, where researcher take raw and noisy data and transform it to cleaned data which can be analyzed by statistical programs. The second phase, the predictive analysis phase, is about building a predictive model and use the data to evaluate the model. Obviously, this is less relevant for our project, however, the steps involved in this phase show how complex can be the analysis of the social media. Kalampokis and his colleagues also note that this complexity is the reason why research effort focusing on prediction often needs cross-disciplinary skills; while the first phase, mostly dealing with databases, is more suited for someone having a computer science background; the second phase needs high level statistical knowledge.

How can we transform raw data to high-quality data? High quality data requires both accurate and relevant data, so researchers first have to remove data which is not relevant. The second part is much harder, as we will later see. Also, high quality data should be complete, so during the data conditioning phase, researchers have to address missing values and incomplete data. In order to select relevant data, according to the framework, researchers have to be able to answer the when, where, who and what questions about the data. Table 2.4 was created to summarize these dimensions, based on Kalampokis et al (2013).

Table 2.4: Guidelines for selecting relevant Twitter data

when	choose the right time window for data collection
where	determine the location of the user based on the metadata (e.g. geo-tagging) or the text or published content itself
who	identify the source of a message
what	set the search terms for data collection

The problems listed in Table 2.4 above were mentioned multiple times in the reviewed information manipulation papers and had to be addressed in the project described in the second part of my thesis.

The next stage in the framework is the computation of the predictor variables, which will be later used to predict the outcome of an event during the evaluation of the predicting model. The authors group these variables around three categories. (1) Volume-related variables include the number of tweets sent during a period or the number of reviews written. (2) Sentiment-related variables focus on the emotional aspect of the data. (3) Profile characteristics focus on the meta-data which can be used to describe a user, e.g. number of friends or the ratio of followers and followees.

2.7. Finding a ground truth

Regardless of the amount of the data accessed from Twitter, there is no easy way to tell whether information is true or not. In order to establish a ground truth most research projects looked at outside of Twitter and find sources which were established and reliable. For example, Qazvinian and his colleagues used the Urban Legends website as an outside source to create a list of false rumors (Qazvinian et al, 2011). Urban

Legends is a website hosted by about.com, and it is basically a reference page dedicated to online hoaxes and rumors. The site also includes a collection of fake news reports which is a good source of online political memes and rumors.

According to Mendoza, Twitter data proved to be a rich source to describe communication patterns after a natural disaster, however, in order to research the propagation of false rumors an outside source has to be involved (Mendoza et al, 2010). Similarly to the research project by Qazvinian, Mendoza and his colleagues first had to identify messages with a confirmed truth, tweets with information which have been proven to be reliable by third party sources. Mendoza and his colleagues put together a list of false rumors, which have been proven to be untrue at some point by either the media or personal sources. Besides the availability reliable information from external sources, the tweets had to reach a certain volume to be considered true information or false rumors. After selecting relevant messages based on the above mentioned two criteria, the research group created a list of 7 confirmed truth and 7 false rumors. For every topic, researchers sorted out between 42 to 700 tweets, and these messages were classified in three categories based on their interpretation of the news. Affirmative messages propagated the information with or without adding any detail to the story, denies challenged or disallowed the message, while questions asked for further details.

False rumors, in average, lead to a higher number of messages. Furthermore, the ratio of unique denies and questions were much higher for (later proved to be false) rumors. While 95 percent of the tweets affirmed or strengthened the tweets with real or factual information (truths), less than every third Twitter messages did so for false rumors. Almost every second message denied the false information and another 13 percent of the messages contained questioned or asked for clarification. This finding suggests that the social media itself react strongly to rumors and misinformation and this reaction is discernible by a simple analysis of the content of relevant tweets.

Danah Boyd, an ethnographer at Microsoft Research, studied retweeting practices and asked about the motivations behind retweeting in a series of questions distributed through a popular Twitter account (Boyd et al, 2010). Based on the responses from Twitter users, the researchers claim that users often retweet for “validating others’ thoughts” or “publicly agree with someone”. However, users can also disagree with others and still share the original content or an edited (shortened) version of it as Boyd points out.

2.8. Data as a research output

I deliberately restricted the literature review on papers which are based on large data sets collected from Twitter, and research projects often indicate the data collection itself as an important scientific collaboration to the field. A good example for this kind of strong passion for large datasets is the research paper written by Qazvinian and his colleagues, where the authors argue that Twitter is suitable for research purposes, because it has a large number of messages created by the users.

> *“As September 2010, Twitter reports that its users publish nearly 95 million tweets per day. This makes Twitter an excellent case to analyze misinformation in social media.”*
>
> (Qazvinian et al., 2011)

As I mentioned above, there are projects where the collection of data itself is considered as a significant part of the research’s contribution to the field. Vahed Qazvinian, who was the Ph.D. student of Dragomir Radev at the University of Michigan in that time, claimed that his Twitter dataset focusing on memes is the first large scale dataset collected for rumor detection (Qazvinian et al, 2011). Qazvinian describes the difficulties of rumor related data collection; it is not enough to collect rumor related content from Twitter, the researchers have to identify whether the poster was endorsed with the rumor, or to put it in another way, did the user believed the content of the rumor. Similarly, a research project on studying information credibility identified the label dataset of events collected from Twitter as one of the main contribution of the paper (Castillo, 2011). Often researchers share these datasets with the academic community or they give a detailed description of the data collecting process if sharing is against the terms of using the data gathered from the particular social media platform.

2.9. Summary

Twitter grants open access to a part of its daily stream of tweets and gives some level of access to its searchable data base of past tweets. In the systematic literature review presented above, I have overviewed those research projects which utilize large Twitter data sets and focus on some aspect of information manipulation.

The research focusing on Twitter communication is usually lacking deeper theoretical background and research questions, as a consequence, research is often data driven. This might be related to the fact that many research papers were published in computer science related conference and related conference proceedings.

On the other hand, a large part of the literature is multidisciplinary, and has been created by a diverse group of researchers having background from computer security to ethnography and communication studies. The reviewed papers focused on a wider range of themes including the analysis of how users and researchers can distinguish between true and false content, how detect spam and astroturfing on Twitter, and how the political scene and the news media have changed due to the popularity of microblogging.

The literature points out numerous challenges about detecting manipulated information on Twitter. Some authors pointed out the huge information overload on Twitter which is especially prevalent in the case of searching for a topic or follow a specific keyword or a hashtag on Twitter related to natural or human disasters. During these events there is also a higher dependency on fresh news and reports from the field. Moreover, the information comes from multiple sources, often from “eyewitnesses” on the field, thus, Twitter users have less clues to identify “trusted sources”.

The platform of Twitter itself offers limited tools to detect manipulated information, so according to the literature, it is often hard to tell the source of information and tell when the information was modified during retweeting. Furthermore, automated accounts or Twitter bots can spread messages quickly and create the sense of consensus and importance by propagating messages with similar content from multiple sources. The fact that the distribution of these messages is often looks similar to natural memes or viral messages make harder detection by software.

One of the interesting results of the literature review is the common practice of using outside sources, at least as a starting point, to distinguish between true and false content on Twitter. Based on the research I reviewed, it looks there is a need for having reliable information or some sort of “ground truth” to start to map the distribution patterns of manipulated or false Twitter massages.

CHAPTER 3 REVIEW OF THE FACEBOOK LITERATURE

This chapter will look at the literature related to Facebook. It will follow the structure that was used for discussing Twitter research. After a short introduction of the site, I will look at the literature on information manipulation in Facebook from two perspectives. First, I will describe the conceptual model that emerges from the body of research and give examples for research projects focusing on specific aspects of information manipulation on Facebook. In the second part of the chapter, I will present a data-centric view of the literature and map the various ways to collect social media data from Facebook. I will also show examples for various data collection methods, including surveys, developing Facebook applications, and crawling data from the site.

3.1. Methodology for collecting articles

For writing a systematic overview of the literature about information manipulation on the largest social network platform, I followed an approach largely similar to that used for Twitter. Here again, the methodology is based on Webster and Watson's guidelines (2002).

Due to the similarities in the studies of information manipulation on these two large social networking sites, Twitter and Facebook, I started with the list of conferences I had earlier identified for the Twitter section. First, I created a list of core articles by scanning through these major conferences and using a set of keywords and phrases to search for academic papers published on Facebook related information manipulation, The keywords included both terms referring to information manipulation and relevant sub-themes, such as credibility or attention spam, as well as terms referring to the platform, such as “Facebook”, “Social media”, and/or “Social network”. For a complete list of keywords and a list of relevant conferences and academic journals, see Table 3.1. The initial list included 20 papers.

After collecting the initial set of papers, I followed those citations which appeared to be related to the focus of my thesis, and then selected the relevant papers. This gave me a larger set of papers to work with. Lastly, I also checked the papers cited in the core papers to see how the discourse has evolved and whether these papers influenced the field. Due to the large number of papers published on Facebook in general, I mostly relied on the abstract of the papers to determine whether the articles were relevant for information manipulation. I will include the list of the abstracts consulted in the Appendix. After major themes and most important findings were identified with the help of abstracts, I selected a limited set of papers. Unlike in the case of Twitter, I will present three projects that may be seen as representative of the themes and approaches characteristic of information manipulation research on Facebook.

During the keyword-based search I could rely on Google Scholar. For the conference specific search, however, I decided to use specific datasets provided by either the organizers of the specific conference or by third party academic databases, because these allowed me a much more focused search in the text of the proceedings. For example, using the search engine of the ACM digital library it was possible to identify all the conference proceedings published since the first International World Wide Web Conference held in 1994 and search for specific combinations of relevant terms.

As I pointed out in the chapter on Twitter, the quality of the research could be measured by the research impact of the published papers. I followed the criteria of having at least five citations for the final selection of the papers. Here again, I relied on the citation numbers in Google Scholar and the plugin developed for Zotero, which could automatically download the information needed for a large number of citations. It is important to note that the literature on the quality of various academic search engines is critical for judging the reliability of the citation numbers used by Google Scholar's database. However, with all the limitation, I did find using the minimum 5 Google Scholar citations a good way to filter out weak papers.

Table 3.1: List of keywords and sources used in the literature review of Facebook

Keywords:	Conferences:	Papers:
Facebook Social network Social media Spam Manipulation Misinformation Facebook API Credibility Trust, Trustworthy	• International AAAI Conference on Weblogs and Social Media • ACM Web Science Conference • IEEE International Conference on Social Computing • CSCW - Conference on Computer Supported Cooperative Work • WWW - World Wide Web Conference Series • CHI - Proceedings of the SIGCHI Conference on Human Factors in Computing Systems	Communication Quarterly Internet Research First Monday Journal of Computer-Mediated Communication

Although Twitter and Facebook can both be characterized as online social media, the literature on the two platforms is certainly different, both in the scientific background of the researchers and in the studied themes. This goes back to significant differences in the platforms themselves, which I will shortly survey in the following.

The first difference is the type of content that is available on the two platforms. While Twitter is designed to share short, predominantly text based messages, Facebook allows its users to post a variety of longer text-based contributions. Besides text, there is a strong presence of multimedia content on Facebook, and users can also install and engage with interactive applications: check in to places, play games, or share travel destinations on maps with their friends.

The next group of differences is related to the control over the access to the content available on Twitter and Facebook. First and foremost, the default setting for tweets is public, while comments or posts published on Facebook are private by default. Besides the difference in these basic default settings, Twitter only allows two levels of access control: the tweets can be available to the entire public of the web or restricted to one or more Twitter users, with whom the message has been purposively shared. Facebook, on the other hand, operates with multiple forms of publicity, which range from sharing a message with the entire social network of the user, with a smaller group within the network, or the entire public of the web, including people who are not even Facebook users or have no connection whatsoever to the user. This latter form of publicity parallels the public access of tweets. These architectural dissimilarities also create a different culture around sharing content on the two platforms.

These differences are somewhat mirrored in the design of the APIs developed for Twitter and Facebook. While Twitter is one of the most open social media platforms in terms of accessing the data stored on its systems, Facebook has much stronger protection against using its data, partly because of the higher levels of privacy expected by its users.

Finally, Twitter is mostly about sharing public information, or sharing emotions or attitudes toward certain issues which have some public interest. On Facebook, however, there is a high concentration of personal information, some of which may be only accessible to research upon special request, and this makes it more difficult to design a good research protocol to access the data. These differences can be expected to make the design of data collection on Facebook more challenging, especially when we consider the requirements of obtaining an IRB approval for human subjects research. Because of this, the review of Facebook research will also include a discussion of ethical questions.

3.2. Introducing Facebook

Facebook is the largest social networking site with more than 1.23 billion monthly active users (Facebook, 2014). Partly due to this enormous popularity, there is a substantial body of literature focusing on Facebook. Also, Facebook is perceived as the prime example of the online social network, and findings from Facebook research are often presented as applying more broadly to social media and online social networks (OSNs).

Online social networks use different models for managing relations and information flow between registered users. In Twitter, each user can decide if he or she wants to follow another user, and receive the feed of their tweets. This type of connection is often referred as unilateral relationship. On the other hand, the friendship mechanism applied on Facebook is bilateral, because both users have to acknowledge the connection (e.g. Guille et al., 2013). At the same time, Facebook also has other, unilateral forms of relationships, such as liking a public figure, an organization, or a company. It is also possible to follow another user, which explicitly indicates intention to get “updates” from the particular Facebook user.

Facebook not only connects users and allows them to communicate with each other, but they can play an important role in spreading information coming from the members of the networks or originating in an outside source. Facebook allows users to simultaneously share information with any number of users. In other words, the platform can both act as a one-to-one and one-to-many communication platform (Bahsky et al., 2012).

From the perspective of information manipulation, Facebook has at least four distinct attributes, based on social, technical and economic characteristics of the platform, which make it interesting and worthwhile to study.

First, Facebook has a high user engagement. According to the Pew Research, two thirds of Facebook users visit the site or its mobile version at least once a day, and about 40 percent check their profile more than once a day (Chen, 2014). Also, Facebook has a large mobile user base. (Compare the 1.23 billion overall monthly active users with the number of 945 million monthly mobile users at the end of 2013 (Facebook, 2014)). The high user engagement and mobile use almost makes Facebook omnipresent in the life of its active user base. It follows from this that Facebook has a very detailed imprint of the personal life of its users, a characteristic that makes the data associated with a user profile extremely valuable. Secondly, Facebook users share personal, often highly sensitive information about themselves on a regular basis. Furthermore, it is possible to connect pieces of information with socioeconomic and demographic data about users, which makes the information shared in the course of personal communication more valuable. This also creates security and privacy risks for the users; the risks include surveillance, identity theft, and cyberbullying. (For more information on these, see Aïmeur, Gambs, and Ho, 2010).

Besides the social dimensions of Facebook use, there are technical aspects of the platform which makes it a potential target for information manipulation. This is the motivation for an important segment of Facebook-research. Although Facebook content is mostly private, applications and the Facebook API can access user content on a large scale if users intentionally place their trust in an application or they are not careful in choosing their privacy settings. Furthermore, security holes in Facebook's system can also make some of the user content accessible beyond the circle that it was originally intended for.

Finally, the last reason to study Facebook is the business potential in its applications. From early on, Facebook allowed application developers to make money on their apps (Graham, 2008). In 2008, the revenue sources for developers include advertisement (based on Google AdSense), Amazon Associates Web Service, and other retail affiliate programs (ibid.). Besides legal and supported ways of monetizing Facebook apps, spammers can build apps which use the social networks of users to distribute advertisement or even malicious links. Studies have shown that "context-aware" spam works very effectively within Facebook, and by using the social connections of the users, spammers can achieve a higher click-through rate compared to traditional email-based methods (Brown et al., 2008).

3.3. Themes in the Facebook literature

3.3.1. Spreading information on Facebook

There are strong motivations behind trying to understand the large scale spread of information in Facebook. Applications include the following of specific real time events (such as political uprisings or natural disasters), as well as marketing and other business. Based on this strong interest, researchers have studied the rising of popular topics, the diffusion of information, and the role of highly influential nodes in the network (Guille et al., 2013).

Researchers from Facebook and the University of Michigan looked at the role of Facebook in the diffusion of everyday information (Bahsky et al., 2012). They applied Granovetter's famous strength of weak ties concept to social media. Instead of looking at strong ties which bond together "small, well-defined" groups, Granovetter had focused on the role of weak-ties between casual acquaintances and how they can act as "local bridges" and help the spread of important information (Granovetter, 1973). Job offers, for example, were commonly shared in these channels. However, the special role of weak ties had not been established in the case of "everyday information", which is available to a large number of individuals. Bahsky and his colleagues used Granovetter's model to study the diffusion of this type of information on Facebook. Based on a systematic analysis of the spread of comments, messages, photos, and threads within Facebook, the authors claim that weak ties contribute significantly to the diffusion of novel and diverse information (Bahsky et al., 2012). It may be added that an important reason behind the spread of information is the relatively low cost of information sharing in Facebook (ibid.).

Findings about information diffusion on Facebook can be applied to the spread of misinformation. A good example is the research studying the role of certain nodes in propagating (mis)information. It has been shown that the diffusion of information can be used to develop counter-strategies, using for example a strategy which relies on highly influential nodes that can effectively spread information to balance out misinformation (Budak et al., 2011).

3.3.2. The motivation behind misinformation

Research has also looked at the motivations behind intentionally creating or distributing false information. It seems that this problem can only be researched by "looking into" the head of users, in other words one has to ask actual Facebook users about their reasons in order to understand their motifs.

Chen and Sin (2013) surveyed more than 170 university students in order to map the possible reasons behind spreading misinformation. Their research has come up with controversial findings. They could identify a strong norm about sharing truthful and reliable information, but people did not always follow these norms. In other words, "knowledge does not always translate into action". The authors used the term "knowledge-behavior gap" to describe this special phenomenon.

Interestingly, their research has also shown that differences in personality could affect the way and more importantly the reasons behind sharing misinformation. According to the above mentioned paper, introverted individuals share misinformation in order to socialize or be part of the community while people with high self-esteem share for self-expression.

Facebook provides various ways of self-presentation from real-time status updates to various forms of expressing preferences or engagements like picking favorite movies or books, and post photos with friends. Naturally researchers with an interest in both psychology and online communication find Facebook a rich platform to study. In an exploratory study based on surveying 477 users, Rosenberg and Egbert looked into the self-presentation and impression management and their relationship to the users' goals on Facebook (Rosenberg and Egbert, 2011). They point out that self-presentation tactics often involve manipulation and various forms of self-promotion. This is especially true for Machiavellian personalities who are, according to the authors, "manipulative and willing to fabricate impressions of themselves". The paper also describes goals directly related to the desire to influence others and cause some sort of behavior changes. Even though these are relevant for information manipulation, the paper stays on an abstract level and the findings cannot be directly used to either identify or mitigate information manipulation on Facebook.

3.3.2. Automating the distribution of misinformation on a large scale

Spam on Facebook is a good example for distributing messages on a large scale, which can count as information manipulation. Facebook is both a highly effective communication channel and a network of users who trust each other. Both of these attributes makes Facebook a valuable target for spammers. Gao and colleagues studied spam in the context of Facebook (Gao et al., 2010). Their research has shown that users are more likely to respond to a message from a member of their Facebook social network (a friend) than from a stranger. Because of this, Facebook can work more effectively for spreading unsolicited messages, and users are more likely to click on the links that are distributed in this manner than those in traditional email-based spam. Besides unsolicited advertisement, Facebook spam campaigns can be used for phishing and spreading malware. Gao et al have found that the overwhelming majority of Facebook accounts that started or contributed to these spam campaigns were compromised or faux accounts.

Spamming on Facebook often relies on networks of socialbots, which represent a case of intentional manipulation of information, as it consists in registering faux users which can be used as proxies for distributing messages on a large scale. The approach is described in detail by Boshaf et al (2013). A socialbot is a software which mimics the actions of real users of a particular OSN, sometimes using artificial intelligence to determine when and what to do. This is not a highly developed area within the field of artificial intelligence and the bots would not have the level of sophistication for passing the Turing test, but they may be good enough to avoid suspicion. Some bots, for example, are programmed to contact only a limited number of users in a single day. It is much harder to detect a socialbot exhibiting this type of modest, human-like behavior than a spambot, which sends out a massive amount of unsolicited messages, even if some users, for example companies and large non-governmental organizations, might also send large amount of messages in a relatively small time. Another group of social bots can be described as "self-declared" bots which are not trying to hide or cover their activity. Boshmaf and his colleagues (2013) cite the Twitter account of a weather forecasting service as an example for sending a large number of automated messages on a regular basis.

Similarly to traditional botnets, a Socialbot Network (SbN) consists of several, sometimes 1000s of socialbots and a botmaster which is a special software designed to control and coordinate the actions of the socialbots. The botnets are controlled by an individual or adversary, often referred as botherder.

Based on the existing literature, Boshmaf et al also describe how adversaries can circumvent the current software-based self-defense mechanisms of social media. For most social media platforms users only need to present a valid email address, create a simple profile, and sometimes use the Captcha method, which requires typing in text presented as an image. For example, email addresses can be created relatively easily with some email providers who don't try to prevent multiple registrations from the same user. Similarly, Captchas can be tricked by a combination of optical character recognition and machine learning or simply by hiring cheap human labor over the internet to manually enter the codes.

According to the authors, socialbots can be used to spread misinformation on a large scale in order to "invisibly" shape public opinion. Socialbots can also collect personal data or perform surveillance of social media contributions, and spread malicious content. In some case, the malicious content provides access to the victim's computer or to their social media profile which can be used for further attacks.

The Canadian researchers designed an interesting, but somewhat disturbing experiment to show how vulnerable OSNs are to the attack of socialbots (Boshmaf, 2013). They created an SbN as described above and deployed it over Facebook. The networks consisted of 102 socialbots and it was active for 8 weeks. The socialbots followed the commands of the researchers and during the time of the software experiment they contacted all together 8570 users with a simple friendship request. Interestingly, almost half of the users, 3055, accepted the request and established a "friendship" with the faux socialbot user. Having a good looking picture, especially displaying the image of a young and attractive woman helped the bots to find new “friends”.

Some might find problematic this research method where faux users deceive real users who are not aware of the experiment, and as it follows, have not signed a consent form. The socialbots not only befriended the users unaware of the experiment, but they also queried their personal data, including their address, phone number, and relationship information. Regardless of the ethical concerns, the experiment has shown how malicious users can exploit this method and how easy it can be to access personal information about users.

Another important finding of the research was the fact that the socialbots achieved a much higher acceptance ratio for their friend requests when they contacted users who with whom they already had common friends. This suggests that users were less careful once they thought they had a connection to the faux user, and also means that being part of someone’s social network provides easier access to other members of the network. Finally, users who had a larger number of contacts were also more ready to accept the friendship from a socialbot compared to users with a lower number of friendships.

3.4. Data collection on Facebook

Wilson, Gosling, and Graham (2012) have surveyed the data collection approaches used in research about Facebook, and cite surveys, applications and data crawling as the three major forms of data collection. They suggest that Facebook is appealing as a research tool not only because of the large number of users, especially in the younger generations, but also due to the “measurable traces” they leave in the platform. My own survey of the literature suggests that survey is the predominant approach for studying Facebook, and only a small number of research project collect traces from Facebook. In this respect, Facebook differs significantly from Twitter, which was characterized by a predominance of original content and data in research. This is most probably due to the challenges of accessing raw data from Facebook.

Facebook has an own team of researchers who have special access to user data, and are thus able to produce in depth data analysis, but researchers outside of Facebook have no direct access to the data stored on Facebook's servers.

While Twitter sells its user data through third party sellers for both research and marketing purposes, Facebook only provides indirect access to its data about users, which has become a major source of revenue for the platform. This knowledge, including rich details about the mobile users of the platform, is used to target advertisement which is the main revenue source for Facebook. The mobile platform in itself has generated more than a billion dollar revenue in the last quarter of 2013 (Edwards, 2014).

At the same time, the Facebook platform offers third-party developers the opportunity to develop applications designed for data collection (Wilson, Gosling, and Graham, 2012). As I have mentioned before, users may allow these Facebook applications to have access to a subset of their personal information for providing their services. These small programs can be used for the purposes of data collection and analysis in the same way.

A third type of data collection method is to crawl personal data from user profiles publicly available on the web without the "active participation" of the users (Wilson, Gosling, and Graham, 2012). Gjoka et al (2011) suggest that crawling data from online social networks is important because a complete dataset, even in anonymized form, is rarely available from any of the online social networks. It should be added that since March 2011, Facebook has had a very strong control over data crawling (Facebook's privacy terms interpreted by Wilson, Gosling, and Graham, 2012). It is against the terms of use to carry on any automated data collection, including the use of scrapers and harvesting bots, without the explicit approval of Facebook.

The prevalence of surveys may further be explained by the fact that the other data collection methods require special, mostly programming skills. Facebook applications need to be created, or web content needs to be parsed. Cleaning and analyzing larger data sets extracted from social media may require further programming. It is not surprising that the only paper mentioned in the review which used large data set from Facebook was not written by social scientist. The project analyzed Facebook relationships across multiple countries and used data about millions of users (Backstrom et al., 2012). Two of the five authors, Lars Backstrom and Johan Ugander worked for Facebook as data scientist, and a quick Google search showed that they studied computer science and applied mathematics at the Cornell University, while the other members of the project also had strong backgrounds in mathematics and computer science at the University of Milano.

Wilson and his colleagues also point out that social scientists often turn to problems in which they have had a long term interest or extensive experience, such as the study of real life social networks or how people communicate and form their identities (Wilson, Gosling, and Graham, 2012), which imply the use of conventional social science research instruments.

The Wilson review also provided a limited survey of the major recruitment strategies for Facebook research (Wilson, Gosling, and Graham, 2012). The first common strategy identified was recruitment of participants in an offline context. In some cases, as we could see, recruiting volunteers from schools proved to be a relatively easy way to get a sizeable number of respondents for a Facebook related survey project. Wilson points out that this method is particularly relevant if the researchers are interested in comparing online and offline behavior. Furthermore, if researchers would like to study non-users, or they want to have a representative sample, offline recruitment is the best way to achieve it. Caers et al (2013) also identified offline recruitment as the most common method behind the social science papers.

Online recruitment is more challenging because of the lack of a complete list of the users (population), which makes sampling difficult. The literature suggests to rely on smaller, but representative samples, (Gjoka et al., 2011). For example, the paper compares various forms of random walks; some of these lead to completely biased samples while others end up producing relatively uniform samples (ibid.).

Recruiting participants via a Facebook application could lead to non-random, large scale data. It is generally true that in order to get volunteers, researchers have to build applications which provide some sort of feedback to the users (e.g. instant analysis of the results), must be interesting and ideally are able to spread virally. For example, MyPersonality, a survey application on Facebook designed to measure personality, gave personalized feedback to the respondents, and it allowed users to rate the personality of their friends (Stillwell and Kosinski, 2012). This application was developed by a Ph.D. student at Cambridge University, and it has reached more than 6 million Facebook users.

The app was used to study various domains, ranging from the connection between personality and web browsing to the privacy awareness of Facebook users. Interestingly, respondents did not only provide answers to the online questions, but approximately 40 percent of them also gave access to their personal profile data on Facebook, so the application could collect a rich dataset of personal information (ibid).

In the following section I will show further examples for the three major types of data collection from the literature I surveyed. I will start with a project where data was collected with offline methods, and respondents were also recruited outside of the social media. Next, I will describe a series of cases where researchers either developed a special application or accessed personal data that was already being collected through an already deployed and popularized application. The last section will show various ways to access data by crawling Facebook.

Besides these three general ways of accessing data, a small number of researchers have access to basically all information, including communication data stored on Facebook's servers. These researchers belong to the so called in-house research team of Facebook, which at times also includes short-term research interns from mainly US universities.

3.1.1. Survey data

In research projects conducted by social scientists, using survey was the predominant way to collect data about Facebook users. This was especially true for research projects focusing on attitudes or the difference between online and offline "friendships".

What do researchers gain and lose on choosing surveys instead of other methods of data collection? One clear advantage of a survey is the ability of collecting a wider range of socio-demographic data in a standardized and comparable format. Eszter Hargittai and Danah Boyd (2010) used survey to study attitudes and practices around privacy on Facebook. Facebook has been widely criticized for not having enough protection over the personal data of its users and having a rather complicate Facebook privacy setting interface. In December 2009, Facebook changed its privacy policy and made users' content publicly available within Facebook as the default privacy option. Consequently, users who did not pay close attention to these details shared their personal information, including family and relationship status, work and education, as well as posts they create, with all users on Facebook. Besides other users, this personal information was available for software developers who build applications for the Facebook Platform.

The researchers tried to understand how users privacy practices have changed due to this and some other new privacy related changes over Facebook. Their choice of data collection was surveying 18- and 19-year-old students. For example, they asked questions about the highest level of education and ethnicity. Besides seeking a wider range of demographic information, which was also probably more reliable, they asked questions about the respondents' Internet skills to answer specific research questions. Similarly, they asked questions about the users' attitudes toward privacy on Facebook.

Even though the researchers could have asked questions about the social networks of the users, for example the number of their friends, these answers might be limited or even false. Similarly, self-reporting privacy settings or the frequency of use of various features on Facebook might be less precise than using log data or crawling information.

Caers et al reviewed the scientific papers published between 2006 and 2012 on Facebook, and they provide a useful, systematic overview of the data collection methods used in social science for studying Facebook (Caers et al., 2013). Interestingly, the paper almost exclusively cited research projects which used survey as a tool to collect data about Facebook users. A few papers, mostly in the category of building and maintaining Facebook relationships, looked at profile data; however, these projects employed a qualitative approach and analyzed only a small number of profiles.

The review also formulated some critical observations about the research projects. Caers and his colleagues were critical about the size and the quality of the data collection, and suggested more robust data collection.

> *"It is striking how many articles are based on samples of US students, often with low numbers of respondents. With millions of users worldwide, research on Facebook should be taken one step further, expanding research to multiple countries and settings and integrating research findings."* (Caers et al., 2013)

They pointed out how researchers use their own classroom of students as a "convenience sample." The limited scope of data collection and the bias in sampling seriously limit the generalizability and the relevance of the findings reviewed. The reviewers also pointed out that employing a more sophisticated method, such as random walks, could strengthen the research findings.

There is no excuse for laziness and going for a "convenience sample", but the picture might be more complex in some cases. There might be at least two reasons for why survey results were based on a relatively low number of respondents. First, conducting surveys is probably the most common quantitative methods in social sciences; so it can be considered as the traditional research method. At the same time, it is difficult to create representative samples of a large online user population if one doesn't have access to the list of all users.

3.3.2. Facebook applications

The Facebook Platform is a software environment which supports developers in building "applications with deep integration into Facebook" (Blizzard, 2007). The applications cannot overwrite the privacy settings of the users, but they can access public data and data that the users have agreed to share with the app. For example, users can allow the application to access their friend list or some profile information such as the date of birth, sex, or religious and political views.

Researchers from the University of California created three Facebook applications in order to gather a "rich dataset on the usage of social network applications" (Nazir et al., 2008). The project was highly successful and together, the applications could reach about 8 million users. At the time of writing, some of the apps were actually positioned in the top one percent of all Facebook applications. The most successful application, with more than 4 million users, was called Got Love, and it simply allowed Facebook users to select some of their friends and display them as loved friends. They also developed a game called Fighter's Club which allowed users to participate in teams and get into long virtual fights with other teams. During the fights users had to get support from their friends to defeat the other team, so they invited other people to join the game. Users installed the application on a voluntary basis and the application did not trick users into sharing their personal data. These kinds of applications might still raise some ethical questions. Especially if users are not clearly aware of the fact that the data from the game is used for research purposes or if they are not fully aware of the nature or value of the personal data they have shared with the application.

Even if the main purpose of the application is data collection, in order to have a large user base the application has to be interesting or useful to the users, otherwise they simply won't install it. Sometimes the purpose of the application development is not motivated by data collection, instead, it aims for the practical implementation of previous research findings. This is especially true for privacy or security applications, where there might be a conflict between data collection and reaching a large number of users. In other words, the possible users of these applications might be extremely sensitive to privacy and would not install the application if it was to collect and store their personal data. A good example for this type of application is the Privacy Awareness App developed by the Austrian Wirtschaftsuniversität Wien. This Facebook application, after accessing all of the user's personal information, creates a list of all the data stored on Facebook's servers about the user. The app also includes a visualization of the user's social network. The terms of use for the application explicitly say that the app won't store any user data.

Similarly, the X-Vine plug-in is designed to detect Sybil attacks and enhance a user's privacy; it uses information about its users and their friends in order to detect Sybils, but it deletes all collected data on a regular basis (Mittal et al., 2011).

There are applications which both support research purposes and help to protect users' security. MyPageKeeper, for example, is designed to help protect Facebook users from spam and malware (Wang, 2012). The application uses social context to differentiate between legit and suspicious activities. The application has been installed by more than 12000 users, and it collects and analyzes wall feeds from it users on a regular basis. The application also provides feedback to the users and warns if it finds questionable posts.

Fire et al developed a Facebook application to research a problem. The application has three main functions. The first module of the application is focusing on detecting faux users by analyzing various network information available for the users without any mean to aggregate or analyze the information. The second function of the application helps users to make informed decisions about privacy setting over Facebook, so it while the first one is an active tool designed to mitigate the possible security risks created by the faux users, this one is a passive, preventive tool. The last component scans all the application installed by the users and look for malicious applications which might pose a security risk to the user. The application looks for various features of the users and their relationship. For example, it can identify friends with high risk of being faux users by looking into the numbers of common friends, co-membership in user groups, and the application even searches for photos and videos where the particular friend appears to be together (tagged) with other real users, including the user of application.

Due to extensive media coverage (for example in the popular technical magazine Wired) the application developed by Fire and his colleagues has been installed by more than 3000 Facebook users from numerous countries. This extensive user base has contributed to the building of a large dataset on Facebook friendships between real and faux users. This knowledge can be used to make the platform more reliable. The researchers used this “unique dataset” for supervised machine learning and developed classifiers for detecting faux users in the system.

What are the advantages and the disadvantages of using a Facebook application for data collection? A Facebook application can potentially reach a larger number of users than most survey based research does. For the survey projects described in the literature reviews or surveyed for the thesis, the number of respondents ranged between 75 and 2000. In contrast, we could see that the three applications developed by the University of California reached 8 million users. Beyond the number of users, it is also important that applications can log the interaction between users and connect some of the information by using the social network data available on Facebook. A possible drawback of this approach is the need for programming skills. Finally, it is not always easy to tell the time scale of data collection if the researchers want to reach a minimum number of users. The literature has reported about applications designed for data collection being much less popular than the developers had previously anticipated.

It is worth noting that these project often place active intervention in front of research in the sense that they develop an application to deal with a problem which has not been well researched. On the other hand, it often appears to be the case that developing an app might be the only feasible way to collect data on the problematic phenomenon in question. Furthermore, even without strong and systematically collected empirical evidence, one can get a good sense of the underlying issues.

3.3.3. Crawling data

The following section provides examples for data crawling where researchers use automated ways, usually Facebook's APIs, to download information about a large number of users. Both the terms of use for Facebook's APIs and Facebook's privacy policy have changed quite often, so some of the methods described might not be available today. Researchers often use an actual Facebook account (either their own profile or new profiles created for the sake of research), and download the data accessible from that specific account.

For example, researchers from the University of Florida and the Los Alamos National Laboratory studied the spread of misinformation by using a dataset about Facebook social networks (Nguyen et al., 2012). In that time users could join to regional networks, based on their city, region, or country, which made easier to find other users from the same region. Relying on this possibility, the researchers created new accounts and downloaded data by accessing the New Orleans regional network. The same regional groups allowed researchers to download information about a large number of users once they also joined the network. During the research project mentioned above, Nguyen and his colleagues crawled network information about more than 63,000 users and their 1.5 million friendship links (ibid.).

Another mostly US-based group of researchers studied social spam by crawling Facebook for three months (Gao et al., 2010). Instead of focusing on one geographic region, this research project chose 8 regional networks in the US, as well as Egypt, Russia and two further countries from Europe. They seeded 50 random users from each region and searched for other users with visible profiles in the same network based on a special algorithm. The regions have various sizes of users, but in the end of the data collection period the researcher collected 187 million wall posts from 3.5 million Facebook users (ibid.).

Kushin and Kitchener (2009) studied political discussion on Facebook. They used a different approach and focused on one discussion group with 800 members from various regions of the world. This dataset was smaller, but it focused on a specific discussion which was public, so it could collect a complete log of the conversation.

Just as other data collection methods, crawling has its advantages and disadvantages. One advantage of this model compared to the other two approaches is the lack of users' action. Although this aspect of data crawling raises some ethical question that I will discuss in the next section, it is a more reliable way of collecting large datasets. However, the lack of users' consent also means that the scope of data is limited to the pieces of information that a user makes generally accessible in his or her privacy settings. Usually using the API needs less programming skills than building a complete Facebook application.

3.5. Ethical questions

Facebook, poses more challenges from an ethical perspective compared to other forms of social media due to three major reasons. 1. Facebook contains rich, often highly sensitive, personal data. 2. The research on the use of privacy settings on Facebook has shown that users often don't understand the stakes of sharing personal data with strangers or third party applications on Facebook. 3. In connection with the second point, it is hard to decide what is considered public and what is considered private on Facebook.

If we look at the three major types of data collection methods on Facebook, there are different ethical questions with regard to each of them. Probably, the easiest case is the offline recruiting of subjects, because surveys and interviews are well covered by the IRB process, and in most cases ethical review boards have to approve the research project. Using Facebook applications or crawling publicly available data are much more challenging to assess from an ethical perspective and the traditional research protocols are not prepared for evaluating the risks of data collection. On the one hand, someone could argue that the general principles of doing human subject research should be applied, and participants of a research project should be informed about the procedure, they should fill out a consent form and the researchers have to prove that they will adequately protect the information they collect. On the other hand, collecting large scale data that is publicly available in social networks would become almost impossible. As Wilson points out, the situation gets more complicated when researchers access private information, data which users decided to not share with the general public in their privacy settings, or when the researcher interacts with the user, for example, in the form of a Facebook application (Wilson, Gosling, and Graham, 2012).

Sharing the data collected from social networks can contribute to future research. However, it has to be done in a way that protects the highly personal information of the users and ensures that data is only shared in anonymized forms. There are successful examples of sharing social media related data with the wider research community both in the case of offline data collection and examples where the researchers used an application to gather data. An example for the latter is the myPersonality Project which allowed fellow researcher to use some of their data after they registered as collaborators. The available data included demographic and location information, Facebook network data, and sensitive data such as political views or religion as well as psychological profiles (Stillwell and Kosinski, 2012).

A smaller, survey based dataset is available from the University of Illinois at Urbana-Champaign. The researchers behind the project decided to share some of the data with the research community. From the IRB documentation of the project, we can learn that the research was based on a survey, the data collection was voluntary and anonymous, and the researchers did not collect any personally identifiable information (University of Illinois at Urbana-Champaign IRB, 2009).

However, sharing the data and protecting the identity of the research subjects is not always successful, as Michael Zimmer (2010) describes in a paper about a 2006 large scale data collection. The ‘‘Tastes, Ties, and Time’’ (T3) project collected and made public information about students of an undisclosed US university. However, just in few days after publishing the dataset, people were able to find out the name of the university. The anonymous dataset was first narrowed down to a set of possible locations, and then it was quickly traced back to the prestigious Harvard College in Massachusetts.

In order to protect the privacy of the 1700 students involved in the study, the T3 researchers, besides not disclosing the name of the university, removed all student names and ID numbers from the data base. The code book, however, gave enough information to identify the university as Zimmer described in great details. This case study revealed the challenges and technical problems of anonymizing data. Also, it raised important ethical questions about publicly available anonymous personal data. Zimmer concluded that the fact that data was already available in the social media due to low privacy settings doesn’t mean that it is acceptable to collect the information on a massive scale and later share it with others.

CHAPTER 4 THE RESEARCH PROBLEM AND RELATED INSIGHTS FROM THE LITERATURE REVIEW

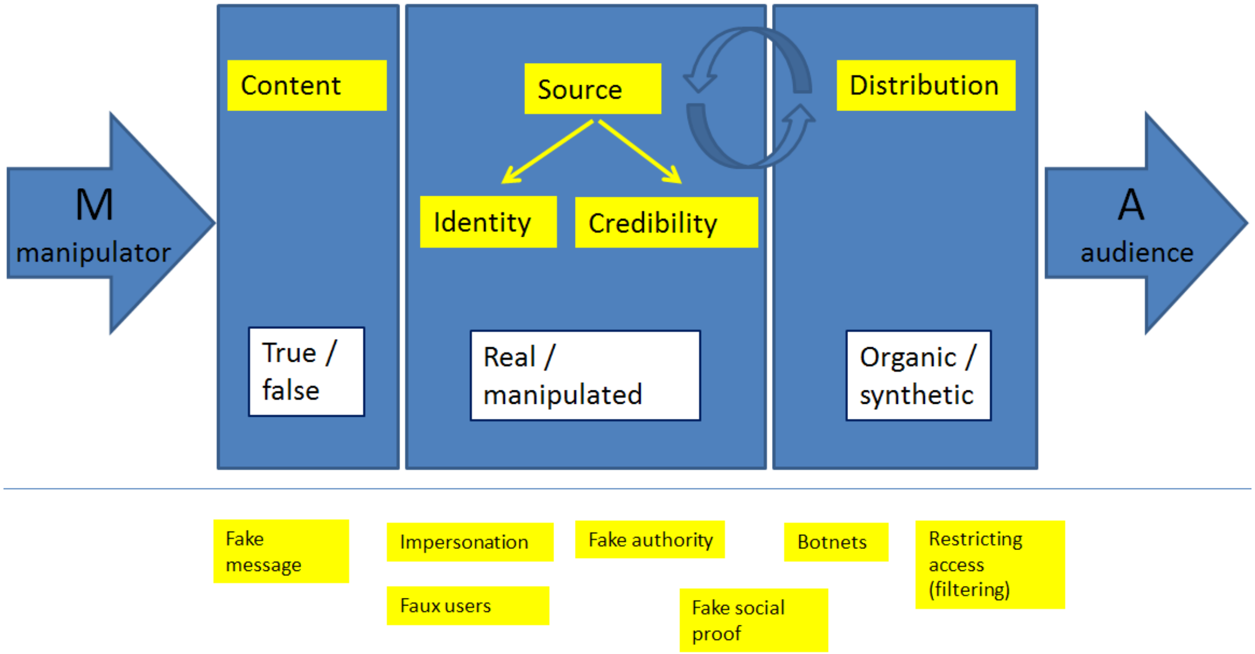
Michael Best and his research team outlined a research agenda to investigate the possibilities of using social media for election monitoring in the context of African national elections. They created the Social Media Tracking Center to provide real-time response to cases of violence and election irregularities detected in social media. This research has indicated the existence of information manipulation on social media outlets and the possibilities for citizens to engage in surveying their own election process with the help of online tools. My case studies explore this problem.

More specifically, my research problem has been the following: How does misinformation and disinformation come about in online social media, like Twitter in a high-stakes situation, and how could this be mitigated? Can social media self-correct the processes behind misinformation and disinformation? What are the heuristics of users for detecting and countering possible misinformation and disinformation? How can these insights be used to enhance the self-correction mechanisms inherent in social media?

This chapter defines the research problem for the case studies in the last chapter of the thesis, and summarizes the most important insights from the literature review that have contributed to structuring the analysis of the cases, and they also serve to situate the findings obtained from the case studies.

My analysis was relying on the model of information manipulation that I have outlined, which outlines the various layers of clues for detecting the truthfulness of information (see Figure 4.1).

Figure 4.1: Model of information manipulation in online social media, reproduced from Chapter 1.



The model suggests that information manipulation may concern the content, the source and the distribution mechanisms. In actual contexts, the various forms of manipulation appear in parallel. Thus, a fake message may be rendered more credible by manipulating the identity of the source or creating fake distribution processes. In exploring how users detect and counter misinformation and disinformation, all these layers should be taken into account.

The architecture of social media becomes relevant for the analysis on this account. A basic insight is that Twitter is much less complex than Facebook. The literature review also provides a rich set of examples for information manipulation on these platforms, but does not typically address the problem of detection from the perspective of the user. This is going to be my task in the analysis. At the same time, the studies give insight about possible forms of manipulation.

The literature also claims that misinformation spread faster in a crisis situation. In my analysis I refer to these situations as low visibility situations with high information needs, because in these settings it is increasingly difficult to check the reliability of the sources, but on the other hand, there are high stakes and a strong need for real time information. The Twitter literature also covers civil uprisings, and research projects has shown that the truthfulness of a message is especially hard to determine in these settings.

The research project by Lotan et al (2011) on Twitter communication during civil uprisings also shown that the „low visisbilaty" also makes harder to detect if a source of a message was modified or faked, especially when a message was retweeted. Based on this finding I found the last research question.

As the literature pointed out, the credibility of a rumor doesn't necessary come from its truthfulness or any extra information which is supporting the claim, instead the information is often only considered to be credible, because many people believe in it.

Both the research on Twitter and Facebook describe how metadata about the information can be used to determine the credibility of a source or a particular message, as well as other possible clues which can help to make that distinction.

CHAPTER 5 DATA ON ELECTION MONITORING

5.1. Introduction

Election monitoring traditionally relies on information collected by trained, formal election monitoring crew. Meanwhile, reports from social media have an increasingly important role in overseeing the fairness and the freedom of democratic elections. The Social Media Tracking Center (SMTC) can be described as an alternative election monitoring process designed to monitor election-related reports from social media, verify information, as well as support real-time response to the verified incidents in the form of contacting the local authorities or other relevant stakeholders.

Aggie, the software designed to automatically collect, store, aggregate, manage and visualize reports from various online social media platforms was developed by researchers and students from the Technologies and International Development Lab at Georgia Tech led by Dr. Michael L. Best (Smyth and Best, 2013). Aggie is capable of automatically collecting election related social media content, including Tweets and Facebook posts based on a list of predefined sources and search terms, and save them in a database. Beside Twitter and Facebook, Aggie has been used to collect data from Google+ and Ushahidi. Aggie also aggregates the social media content and provides real-time trend analysis and visualization to support quick reaction in the SMTCs where a trained staff is constantly monitoring the incoming election related reports.

Aggie and the SMTCs were deployed in four African elections in collaboration with local civil organizations. Although Aggie is geared toward real time data aggregation, the system is saving all the collected social media data in a database creating an archive of the election related social media conversations and supporting post-hoc data analysis.

The following three case studies are using data collected during the 2011 Nigerian Presidential Election and the 2012 election in Ghana. The aim of the analysis is to illustrate how misinformation and disinformation arise in this situation, and understand the possibilities of self-correction in social media. The first case study outlines the reactions to an occurrence of hate speech and threats of physical violence by an unidentified user and suggests that truthfulness may not be at the center of diffusion of messages in social media. The second case study is about faking an authoritative user identity in order to spread misinformation and indicates a range of clues and approach which could be seen at work in countering information manipulation. The third case study describes the rise of wide-spread rumors of ballot box snatching, and addresses the difficulty of establishing ground truth with respect social media messages.

5.2. Reaction to hate speech during the Nigerian election

On 17 April, a day after the 2011 Nigerian Presidential election, a user named "abuabdallah92" posted a hateful tweet calling for resistance and aggression in response to the election results and a claimed "stolen election". The post generated at least 218 responses on the Nigerian Twitter sphere. The original post quoted below reported about protests and claimed that a group was willing to attack the local police station.

> *"we succesfully start our protest in Rigasa kaduna! We want to burnt the police station of our reign"*

Tweet from abuabdallah92

The reaction to the tweet from other Twitter users was quick and furious, and almost exclusively expressed disagreement. The responses ranged from being shocked to calling for the arrest of "abuabdallah92" or wishing his death. Numerous messages called the twitter user mad, crazy, and uneducated, which was another way to express disagreement with the message of the tweet.

Due to the nature of the medium, the reactions (retweets) contained the original message and were mostly restricted to a short, one or two-word comment. The following quotes are two prime examples for this type of reaction.

> *"I pray u get arrested RT @abuabdallah92: we succesfully start our protest in Rigasa kaduna! We want to burnt the police station of our reign"*

Tweet from PlaybackGenius

> *"This guy is crazy @abuabdallah92: we succesfully start our protest in Rigasa kaduna! We want to burnt the police station of our reign"*

Tweet from Roshio2gbaski

Not every tweet had a comment attached to the original message which made it harder to tell how other users related to the original message. 55 tweets out of the 218 responses collected by Aggie had no extra comment added to the original message, in other words, the users simply retweeted the original message. Even though there was no sign of other Twitter users agreeing with the original poster, the opinion of the users behind these retweets cannot be established. Nevertheless, these retweets had contributed to the diffusion of the original hateful message.

Interestingly, there were only four messages that asked for the possible identity of the original poster. The following quote from a user called ‘aderibs’ gives a good example for this type of tweet.

> *“who b dis guy sef?RT @obylepatohbad: Geez!!!RT @abuabdallah92: we succesfully start our protest in Rigasa kaduna! ... http://tmi.me/91sbj”*
>
> Tweet from aderibs

Considering the fact that the second half of the original message is describing a plan for a serious crime which could lead to the death of innocent people, property damage, chaos and possibly more unrest, it is strange that no tweets indicated that people were taking action, such as reporting the case to the relevant authorities. At least, none of the users mentioned this in their tweets. It seems, however, that many users shared the original message for the sake of pointing it out to a community or bringing a particular user’s attention to the problem. At least this is what can be gleaned about their motivation based on the comments added to their retweets messages, such as “*Twitfam re u all seeing dis?*” or “*Dis guy was tweetin dis 15 hours ago*”.

Aggie, the software used to collect the relevant tweets during the election, has saved all the messages which have contained the original message (or its retweeted version), because it contained keywords relevant to either the election in Nigeria (the name of the location in the tweet) or closely connected with a possible violent event (such as burning down the police station). Interestingly, many of the original content has disappeared from the public Internet since 2011. Before I started to analyze the content of the messages, I tried to collect more information about the user, abuabdallah92. At the time of my research, none of the original messages were available on Twitter. The user has deleted his original profile (or the profile was banned), but someone (and this might have been the same person) registered with the same user name a year later (the registration data indicates 2012). When searching for the text of the reactions from other Twitter users, I also had difficulties finding the original messages. Twitter's public search engine has no relevant results by searching neither for the user's name, abuabdullah92, nor the content of the message.

The only tweets I could find online were the ones which used Twitter message shorteners, online tools designed to publishing longer than 140 character tweets. These tools allow users to write longer messages and by simply clicking on a web link (an URL) other users can access the whole message. See the following message which has an address at the end of the tweet to access the full text of the original message.

> *"I can't wait till u bcome a christian @abuabdallah92: we succesfully start our protest in Rigasa kaduna! We want (cont) http://tl.gd/9u0att"*
>
> Tweet from mz_essay

The URL reveals the remaining part of the tweet as presented below.

> *"I can't wait till u bcome a christian @abuabdallah92: we succesfully start our protest in Rigasa kaduna! We want to burnt the police station of ou"*
>
> Tweet from mz_essay

The fact that these messages are not available makes it much harder to reconstruct the discussions and understand who is the person behind the username, abuabdallah92. While some people suspected one of the violent protesters behind the user named, and only few others questioned this assumption.

This case study shows how difficult it is to evaluate the source of a message. Furthermore, it tells us that almost no one questioned the authenticity of the original message, and, based on the reactions, they have taken the message seriously and accepted it is as true.

The same author posted another hateful message on the next day which even caused a larger scandal on Twitter and generated close to 400 responses in less than 24 hours.

> *"what we vote is not what we will be given, so we cant stay at home, we must come out and slaught every xtian in kaduna."*

Tweet from abuabdallah92

Beside reteweets and discussion, it has generated messages directly addressed to the original poster of the hateful tweet.

> *"it's obvious u r ignorant stupid + brainwashed. How do u think a riot & killings of xtians will help justify a win 4 Buhari?"*

Tweet from Ms2ra

Also, this generated a deeper conversation where people, besides expressing their frustration, tried to start a dialog with abuabdallah92.

> *"@abuabdallah92 don't desecrate d name of Gen Buhari in d name of violence. Do u know how many xtians voted for him cos we believed in him"*

Tweet from Midewaju

> *"u shouldn't make this out into a xtian vs muslim issue. Even Buhari wouldn't and doeant want that"*

Tweet from deleneye

Due to the large number of retweets and responses, this message was saved on an online service called The Tweet Watch, which is a site archiving trending messages with some background information about the user. This website indicated that the user, abuabdallah92 had 468 followers and 8 friends in the time of posting this tweet.

While in the two other case studies presented later the traffic generated by the original post was concentrated in the first two hours after posting the tweet, in this case the traffic was distributed over a 10 hour period peaking after the second hour.

To sum it up, this case study about a pair of hateful tweets showed that people can contribute to the distribution of a piece of content even if they don't agree with the content. Interestingly, the users did not try to identify the source of the message or tried to question the truthfulness of the message itself. In this process, communicating their own opinion on the topic seemed to be what guided their actions.

5.3. Impersonating BBCAfrica during the 2011 Ghana election

The next case study is a simple and clear example for manipulating the source (the identity of the sender) on Twitter without causing too much damage. It is a prime example for information manipulation and shows how Twitter as a technological platform makes it relatively easy to create false messages with a seemingly legitimate sender, and how other users can be tricked into believing that the information is coming from a particular source.

In the night of the 2011 Ghana presidential election, a user named kwame_sika fabricated the following faux message.

> *"RT @BBCAFRICA: NPP leading in the Ghanaian election EC edges every one to stay calm"*

Tweet from kwame_sika

In this case, the user manipulated the source of the message to spread unfounded information and it caused a great deal of confusion and took almost an hour to BBC to react on the situation.

However, it was only 40 minutes after the original message appeared that a user posted the following tweet:

> *"ReneAsante BBC never reported this...smh"@BBCAFRICA: NPP leading in the Ghanaian election EC edges every one to stay calm""*

Tweet from ReneAsante

Another 10 minutes later BBC finally debunked the message and posted the following update:

> *"Please ignore the tweet saying the NPP has won Ghana's election. It was not sent by @BBCAfrica."*

Tweet from BBCAfrica

Interestingly, this message caused an even greater confusion, because users had started to guess how someone could possibly post a message in the name of BBC Africa. Even the famous blogger, Egghead Odewale, Eggheader on Twitter, suspected a hacking attack against BBC and posted the following tweet.

"Hacking? RT @BBCAfrica: Please ignore the tweet saying the NPP has won Ghana's election. It was not sent by @BBCAfrica."

Tweet from Eggheader

The solution to the problem is much more simple, however. Users can manually create retweet messages by creating a message that looks like a retweet which looks like a real Twitter retweet: "RT @Username:" In this structure RT stands for retweet while @ is used for mentioning the username of the "original" poster.

The initial reaction to the faux tweet was to question the truthfulness of the message without questioning the authenticity of the sender. While many people simply retweeted the message, some added comments to the original in the first half an hour as shown in the following two examples:

"ReAlly??? RT @kwame_sika: RT @BBCAFRICA: NPP leading in the Ghanaian election EC edges every one to stay calm"

Tweet from KofiYankey

"Lyk Seriously? "@kwame_sika: RT @BBCAFRICA: NPP leading in the Ghanaian election EC edges every one to stay calm""

Tweet from Mzswtpam

In theory, a user could check the authenticity of a tweet by simply looking up the profile of the user to see if that person has published the content. With this method, one could "manually" compare the original message with the retweeted message and decide whether the retweeted message is authentic or not. However, Twitter as a platform does not do this comparison automatically. Partly because users can edit the retweeted message to save space for their own comments. For example, it is common to either share only part of the retweeted message or simply use abbreviations to reduce the character count of the original message.

Another way of checking the message is looking for clues of manipulation. For example, in the BBCAfrica case in Ghana, the message had spelling and grammatical errors, indicating that the source might not be a journalist from the British media company. In fact a Twitter user, Ebenezer Gwumah, who is a communication professional according to his profile, responded in about 30 minutes (faster than anybody else) and indicated that the message was fabricated.

> *"If you'll fake BBC tweets, spell correctly. RT @kwame_sika: RT @BBCAFRICA: NPP leads in Ghanaian election EC edges every one to stay calm"*
>
> Tweet from gwumah

Today, BBCAFRICA has more than 440,000 followers, but in the time of the election, in 2011, the account had already had almost 10,000 followers based on the statistics from Wildfire Social Media tracking application owned by Google.

The followers automatically see tweets from BBCAFRICA, so it is not surprising that 94 users shared the message about denying the authenticity of the original message. Also, on the Twitter archive of BBCAFRICA, the discussion about how the fake message was fabricated is still available.

The first lesson from this case study is that people start to retweet messages if it seems that it is coming from a reliable source without questioning the authenticity of the message. Based on the reaction, Twitter users had relatively little knowledge about how easy it is to manipulate the source of a retweeted message, or at least they don't think about it. Finally, small signs like wrong spelling or bad use of grammar could help some users to identify a fabricated message.

5.3. Ballot box snatching and lynching during the 2011 Ghana election

The last case study gives an example for spreading rumor over Twitter. I decided to include this case study to my thesis, because it gives me the opportunity to show how layers of information manipulation become confounded in one single case.

This case study is based on a set of reports about people stealing ballot boxes during the election in the Ashanti region of Ghana, mostly from the city of Kumasi. It is rather difficult to tell what exactly happened during election day simply due to the conflicting reports from the region. For example, while the reports on Twitter refer to multiple cases of ballot box snatching, the Public Relations Officer at the Ashanti Regional Police Command denied these claims:

> *"We have such situations at different places where people allege that a ballot box has been picked, we went down to the place and there was nothing like that. And I can tell you nobody attempted to pick a ballot box from the Ashanti Region and no ballot box has been picked as we speak."*
>
> (GhanaNation, 2012)

Ballot box snatching is a known technique to manipulate the results of the election in Ghana. It was reported to have been used in previous elections. Even before the 2012 election, the police announced preliminary action and preparedness to handle possible attempts to steal ballot boxes with votes inside (Ghanaian Chronicle, 2012).

During the day of the election, multiple sources reported ballot box snatching from the Kumasi region via Twitter. The first tweet was followed by other reports, but in the first 4 hours no confirmation came from eyewitnesses, nor was any mention of a reliable source for the information. During this time, more and more reports surfaced about people being lynched and some people claimed that the ballot box snatchers belonged to the political party NDC, and they were beaten up by supporters of the opposition party. Furthermore, three low resolution images started to circulate on twitter showing a Ghanaian man in a white t-shirt covered with blood, which suggested that he may have been badly beaten.

However, in the social media, there was no mention of media reports of such incidents from the region and after 4 hours Citi news posted the following tweet on its official Twitter account:

> *"Citi News reporters, Betty Agyemang and James Obeeko say reports of ballot box snatching in some areas in Kumasi are not true #ghvotes"*
>
> Tweet from Citi973[d]

Later, some Twitter users started to refer to TV3, a Ghanaian television. They also claimed that the channel reported about ballot box stealing in certain areas of the region. Others responded that the original tweeter simply wanted to create fear, and there was no such report from the TV channel.

[d] According to the Twitter account of the radio, Citi 97.3FM is an English speaking radio channel operating from Accra, Ghana.

I have investigated the particular incidents, and found contradictory reports. I have already cited how police and some media outlets disconfirmed the incidents. I have encountered no reports from authoritative local sources. At the same time, the publication of a formal election monitoring group mentioned violence from the region in its final report, without providing further details (CODEO, 2012).

This case study shows how difficult it is to evaluate the truthfulness of the content posted on Online Social Networks. Even outside sources are difficult to use for establishing some form of ground truth. The lack of eyewitnesses or people who communicate from the field makes this even more challenging. Finally, it seems that during election time, especially if there were precedents for election irregularities in the past, the police has made preparations for dealing with ballot box snatching and there is political tension in the air, people will readily accept the message as truthful and share the content extensively.

APPENDIX A

THE TECHNICAL ASPECTS OF TWITTER'S API

A web Application Programming Interface (API) specifies how computer programs, often third-party applications, can interact with a web service. An API allows other program to access specific features of a web service. For example, based on the API a software can display certain type of information on a customized map, save a file to a cloud storage service, or access the data stored in a database (Proffitt, 2013). Especially in the context of web services, APIs are often used by software developers who develop a new interface to an existing service, or introduce new services using elements of existing web services. In other words, by using APIs programmer do not have to start building a new service from scratch, instead they can build on existing platforms - as Webopedia summarizes it, ideally an API "makes it easier to develop a program by providing all the building blocks." (Webopedia, 2014).

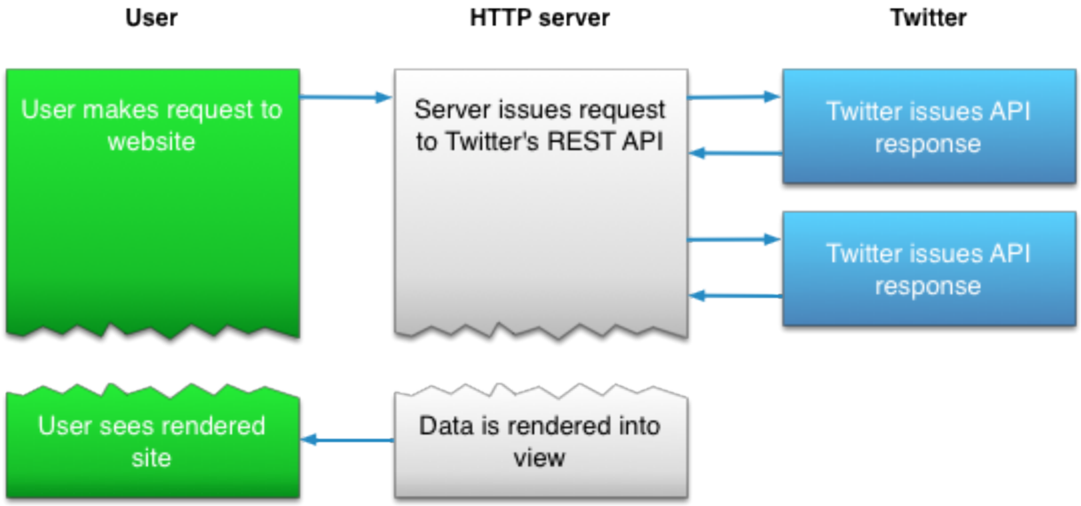
> *"Twitter maintains an open platform that supports the millions of people around the world who are sharing and discovering what's happening now. We want to empower our ecosystem partners to build valuable businesses around the information flowing through Twitter. At the same time, we aim to strike a balance between encouraging interesting development and protecting both Twitter's and users' rights."*
>
> (API Terms of Service by Twitter, 2013)

Twitter provides free but limited access to both real-time and historic data. The data about micro blogging is both available for commercial and research purposes. In general, Twitter is committed to maintain an “open platform”, but the access to its historic data is limited, and the company designed technical limits (caps) in accessing real time data. There are two main reasons behind limiting the access to the otherwise public communication created and shared on Twitter: the company would both like to protect the privacy of its users and secure its own business interest.

The following section of this paper gives an introduction to the technical background of the Application Programming Interfaces offered by Twitter. Without understanding the technical background of the APIs, it would be hard to understand the research focusing on Twitter communication. Twitter’s own website offers a search functionality and direct access to each users’ public profiles, however, for collecting large scale data, it is just much more effective to use Twitter’s APIs.

Twitter has two types of APIs: the REST API and the Streaming API. They are designed for different purposes, and they have different technical model for accessing the data. This section compares the two types of APIs based on Twitter’s documentation designed for programmers who develop new Twitter applications.

The REST API is designed to support a developer who would like to design a new interface to use Twitter’s main functions; Through the API, users can access previously posted tweets, post new tweets, or share them by retweeting. The image below summarizes the technical aspect of the process.

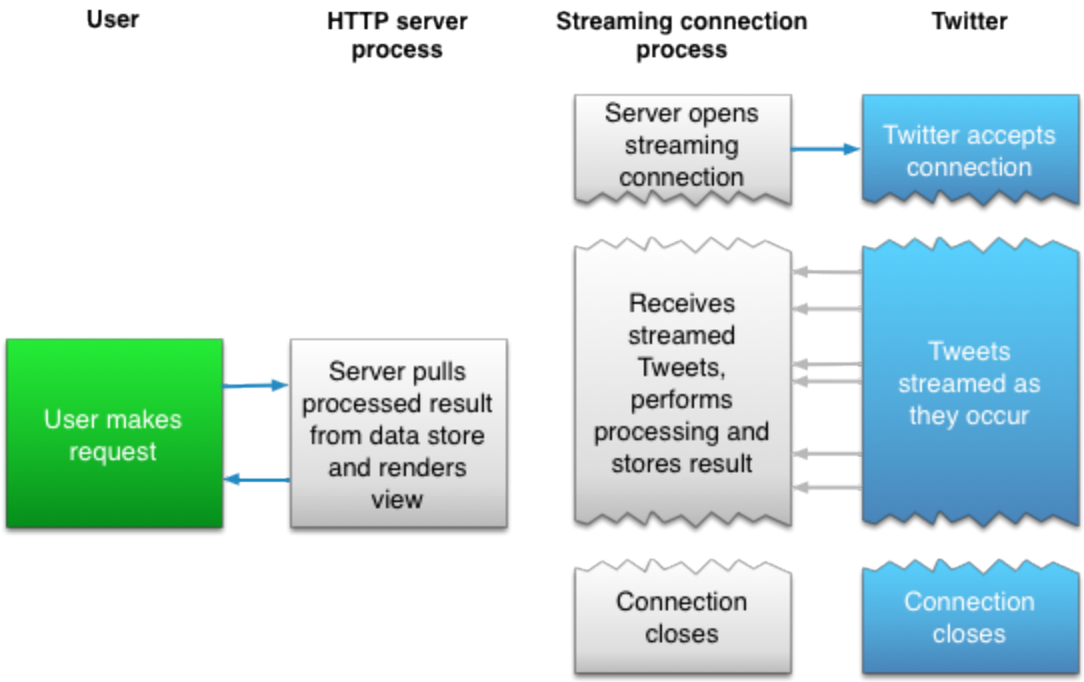

User
HTTP server
Twitter
User makes request to website
Server issues request to Twitter's REST API
Twitter issues API response
Twitter issues API response
User sees rendered site
Data is rendered into view

The REST APIs offer many resources to third party software developers and to their clients. While some of these APIs support looking up previously posted information, for example accessing the last 100 tweets of a user, others allow posting new tweets or retweeting previous messages in the name of registered users.

The Twitter Search API, which is also part of the REST API v1.1, allows users to enter keywords and returns results which are relevant to the specified query. It is important to pinpoint that the Search API doesn't provide access to all the previously posted tweets which would match the specific query. As the API's documentation notes, "the Search API is focused on relevance and not completeness."

Besides limited access to historic data, there are other limits of using the family of REST APIs. For example, Twitter limits the number of new tweets posted through an API in order to control server traffic and limit the distribution of spam. There are similar restrictions on accessing data in order to limit the traffic created by the APIs. Twitter usually determines these limits based on the number of users (or access tokens) an application has. Twitter applies the limits on a 15 minutes interval. For example, an application can have 180 search queries per user in every 15 minutes in the Search API.

According to Twitter's documentation, the Streaming API is designed to "give developers low latency access to Twitter's global stream of Tweet data". Besides the different functionality, the API follows a different technical model for accessing the data as the graph shows below.


User
HTTP server process
Streaming connection process
Twitter
Server opens streaming connection
Twitter accepts connection
Receives streamed Tweets, performs processing and stores result
Tweets streamed as they occur
User makes request
Server pulls processed result from data store and renders view
Connection closes
Connection closes

While a REST API allows access to the data through requests, a streaming API provides access to new tweets as they occur. This gives close to real time, typically within one second latency, access to tweets matching the previously defined parameters. The Streaming APIs accept keywords, hashtags, and geo-locations as parameters. The Streaming APIs can also be used to follow specific users based on their user IDs.

Twitter has three different Streaming APIs, each of them is designed for a different use case as the following table summarize it.

	Focus	**Use cases**
Public streams	public data flowing through Twitter	following a user or a topic /data mining
User streams	single user's view of Twitter traffic	designing a new user interface for twitter
Site streams	multiple user's view of Twitter traffic	a server connects to Twitter on behalf of many users

Twitter's Streaming API has its own limits. Due to the different technical design, these limits or streaming caps are not based on the number of requests, but on the amount of data accessed. The cap of streaming data is based on the whole volume of Twitter traffic in any given point of time and it depends on the level of streaming access a user has. For example, the free version of Twitter's streaming API provides no more than 1 percent of all Twitter traffic, while the access called Firehose gives access to all Twitter data (100 percent). After an access reaches its limits, Twitter's Public Stream API gives only a filtered access of the data. Users can also parameter the streaming API and access a subset of Twitter data which helps to keep traffic low. For most research projects, reaching the one percent of all Twitter traffic is not a real danger. For example, the Streaming API can provide access to data from a specific geographic region, like the United Statas or the city of Atlanta. However, if at any given point of time the traffic created at that specific geo-location gets higher than one percent of all Twitter traffic, the API starts to stream a filtered data set.

Finally, the use of keywords, Twitter users names, and geo-locations are also limited in the free Streaming API. Currently, Twitter implies the following limits: an API user can enter no more than 400 terms at once, while the number of users followed is limited to 5000, and the API can only handle a maximum of 25 geo locations.

APPENDIX B

FACEBOOK API

Facebook has a set of APIs to allow applications to interact with its services, read and write user's data. The most complete access to Facebook is the data available through the Graph API which Facebook describes as the "primary way for apps to read and write to the Facebook social graph" (Facebook, 2014). The Graph API allows third party applications developers or researchers to build tools which can search Facebook posts, limit the results by location and get information about a Facebook user, a page, an event or a group.

Facebook designed a special API to access its public feed, the content which is posted on Facebook with a privacy setting set to public. These public streams include both the individual users' updates and updates posted on pages. However, this API provides only minimal access to metadata, but it is possible to use a combination of APIs and get extra information from the Graph API for the data gathered via the Public Feed API.

In general, access to private communication data on Facebook is limited, however the company's newest API provide access to all communication in an aggregate, thus anonymous form. Facebook calls the Keyword Insights API as an analysis layer on top of its services. Users of the API can search for specific terms, and the API gives back statistical information about certain aspects of the Facebook users who mention the term, including gender, current city and an age range (Facebook, 2014). Facebook launched this API in 2013 September (Facebook, 2013). The announcement of the API caused some controversies, but the company was quick to stress that the API doesn’t provide access to individual messages and user names are not connected to the anonymized data (Constine, 2013).

REFERENCES

Aïmeur, E., Gambs, S., & Ho, A. (2010, February). Towards a privacy-enhanced social networking site. In Availability, Reliability, and Security, 2010. ARES'10 International Conference on (pp. 172-179). IEEE.

Allport, G. W., & Postman, L. (1947). The psychology of rumor. Henry Holt and Company.

An, J., Cha, M., Gummadi, K. P., Crowcroft, J., & Quercia, D. (2012). Visualizing media bias through Twitter. Retrieved from http://www.aaai.org/ocs/index.php/icwsm/icwsm12/paper/download/4775/5075

Arrington, M. (2008). Confirmed: Twitter Acquires Summize Search Engine. TechCrunch. 2008-07-15. Retrieved 2014-01-08. Online available: http://techcrunch.com/2008/07/15/confirmed-twitter-acquires-summize-search-engine/

Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2012). Four degrees of separation. In Proceedings of the 3rd Annual ACM Web Science Conference (pp. 33–42). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2380723

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In Proceedings of the 21st international conference on World Wide Web (pp. 519–528). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2187907

Beck, K. (2011). Analyzing tweets to identify malicious messages. In Electro/Information Technology (EIT), 2011 IEEE International Conference on (pp. 1–5). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5978594

Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). Detecting spammers on twitter. In Collaboration, electronic messaging, anti-abuse and spam conference (CEAS) (Vol. 6). Retrieved from http://ceas.cc/2010/papers/Paper%2021.pdf

Besterman, M. (2013). Someone is Wrong on the Internet. Retrieved from http://mattbesterman.com/blog/wp-content/uploads/2013/12/Someone-is-Wrong-on-the-Internet.pdf

Bifet, A., and Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. In Discovery Science (pp. 1–15). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-16184-1_1

Blizzard, C. (May 27, 2007). Facebook Platform Launches. Retrieved from https://developers.facebook.com/blog/post/21/

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2013). Design and analysis of a social botnet. Computer Networks, 57(2), 556–578.

boyd, d. m., & Ellison, N. B. (2010). Social network sites: definition, history, and scholarship. Engineering Management Review, IEEE, 38(3), 16-31.

boyd, d., Golder, S., and Lotan, G. (2010, January). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In System Sciences (HICSS), 2010 43rd Hawaii International Conference on (pp. 1-10). IEEE.

Brown, G., Howe, T., Ihbe, M., Prakash, A., & Borders, K. (2008, November). Social networks and context-aware spam. In Proceedings of the 2008 ACM conference on Computer supported cooperative work (pp. 403-412). ACM.

Bruns, A., and Stieglitz, S. (2012). Quantitative approaches to comparing communication patterns on Twitter. Journal of Technology in Human Services, 30(3-4), 160-185.

Budak, C., Agrawal, D., & El Abbadi, A. (2011). Limiting the spread of misinformation in social networks. In Proceedings of the 20th international conference on World wide web (pp. 665–674). Retrieved from http://dl.acm.org/citation.cfm?id=1963499

Caers, R., De Feyter, T., De Couck, M., Stough, T., Vigna, C., & Du Bois, C. (2013). Facebook: A literature review. New Media & Society, 15(6), 982–1002.

Castells, M. (2012). Networks of outrage and hope: Social movements in the internet age. Polity.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In Proceedings of the 20th international conference on World wide web (pp. 675–684). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1963500

Castillo, C., Mendoza, M., & Poblete, B. (2013). Predicting information credibility in time-sensitive social media. Internet Research, 23(5), 560–588.

Chen, X., & Sin, S. C. J. (2013). 'Misinformation? What of it?'Motivations and individual differences in misinformation sharing on social media. Proceedings of the American Society for Information Science and Technology, 50(1), 1-4.

Chen, Y. (2014) Facebook Still Dominates in Social Media User Engagement [Study] http://www.clickz.com/clickz/news/2320802/facebook-still-dominates-in-social-media-user-engagement-study

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg? IEEE Transactions on Dependable and Secure Computing, 9(6), 811–824.

Chu, Z., Widjaja, I., & Wang, H. (2012). Detecting social spam campaigns on twitter. In Applied Cryptography and Network Security (pp. 455–472). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-31284-7_27

Cialdini, R. B. (2007). Influence: the psychology of persuasion. New York: Collins.

Conover, M., Ratkiewicz, J., Francisco, M., Gon\ccalves, B., Menczer, F., and Flammini, A. (2011). Political Polarization on Twitter. In ICWSM. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2847/3275

Constine, J. (2013). Facebook Will Give Marketers Access To Conversation Data With Richer Identity Details Than Twitter. TechCrunch. Retrieved from http://techcrunch.com/2013/09/09/your-conversations-are-the-product/

Diakopoulos, N., De Choudhury, M., & Naaman, M. (2012). Finding and assessing social media information sources in the context of journalism. In Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems (pp. 2451–2460). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2208409

DiFonzo, N. and Bordia, P. (2007). Rumor Psychology: Social and Organizational Approaches. Washington, DC: American Psychological Association.

Edwards, J. (2014). Facebook Shares Surge On First Ever $1 Billion Mobile Ad Revenue Quarter. Business Insider. Retrieved from: http://www.businessinsider.com/facebook-q4-2013-earnings-2014-1

Facebook, Inc. (2014). Facebook Reports Fourth Quarter and Full Year 2013 Results. Retrieved from http://investor.fb.com/releasedetail.cfm?ReleaseID=821954

Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., & Zhao, B. Y. (2010). Detecting and characterizing social spam campaigns. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement (pp. 35–47). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1879147

Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Korlam, G., Benevenuto, F., … Gummadi, K. P. (2012). Understanding and combating link farming in the twitter social network. In Proceedings of the 21st international conference on World Wide Web (pp. 61–70). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2187846

Gjoka, M., Kurant, M., Butts, C. T., & Markopoulou, A. (2011). Practical Recommendations on Crawling Online Social Networks. IEEE Journal on Selected Areas in Communications, 29(9), 1872–1892.

Goel, V. (2014, February 5). User Growth for Twitter Starts to Slow, and Stock Dips. The New York Times. Retrieved from http://www.nytimes.com/2014/02/06/technology/twitters-share-price-falls-after-it-reports-4th-quarter-loss.html

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., and Moreno, Y. (2012). Assessing the bias in communication networks sampled from twitter. arXiv preprint arXiv:1212.1684.

Graham, W. (2008). Facebook API Developers Guide. Infobase Publishing.

Granovetter, M. (1973). The strength of weak ties. American Journal of Sociology, 78(6), l.

Guille, A., Hacid, H., Favre, C., & Zighed, D. A. (2013). Information diffusion in online social networks: A survey. ACM SIGMOD Record, 42(1), 17–28.

Gundecha, P., & Liu, H. (2012). Mining Social Media: A Brief Introduction. In 2012 TutORials in Operations Research (pp. 1–17). INFORMS. Retrieved from https://www.informs.org/Pubs/Tutorials-in-OR/2012-TutORials-in-Operations-Research-ONLINE/Chapter-1

Gupta, A., & Kumaraguru, P. (2012). Credibility ranking of tweets during high impact events. In Proceedings of the 1st Workshop on Privacy and Security in Online Social Media (p. 2). Retrieved from http://dl.acm.org/citation.cfm?id=2185356

Hargittai, E., & boyd, d. m. (2010). Facebook privacy settings: Who cares? First Monday, 15(8).

Hermida, A. (2012). Social Journalism: Exploring how Social Media is Shaping Journalism. In E. Siapera & A. Veglis (Eds.), The Handbook of Global Online Journalism (pp. 309–328). Wiley-Blackwell. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118313978.ch17/summary

Howard, P. N., & Hussain, M. M. (2011). The role of digital media. Journal of Democracy, 22(3), 35-48.

Irani, D., Webb, S., Pu, C., & Li, K. (n.d.). Study of Trend-Stuffing on Twitter through Text Classification. In Seventh annual Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS). Redmond, Washington, US.

Kalampokis, E., Tambouris, E., and Tarabanis, K. (2013). Understanding the predictive power of social media. Internet Research, 23(5), 544-559.

Kang, B., O’Donovan, J., & Höllerer, T. (2012). Modeling topic specific credibility on twitter. In Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (pp. 179–188). Retrieved from http://dl.acm.org/citation.cfm?id=2166998

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons, 53(1), 59–68.

Karlova, N. A., & Fisher, K. E. (2013). A social diffusion model of misinformation and disinformation for understanding human information behaviour. Information Research, 18(1).

Karlova, N. A., & Lee, J. H. (2011). Notes from the underground city of disinformation: A conceptual investigation. Proceedings of the American Society for Information Science and Technology, 48(1), 1–9.

Kim, W., Jeong, O.-R., & Lee, S.-W. (2010). On social Web sites. Information Systems, 35(2), 215–236.

Kushin, M. J., & Kitchener, K. (2009). Getting political on social network sites: Exploring online political discourse on Facebook. First Monday, 14(11). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/viewArticle/2645

Lampos, V., and Cristianini, N. (2010, June). Tracking the flu pandemic by monitoring the social web. In Cognitive Information Processing (CIP), 2010 2nd International Workshop on (pp. 411-416). IEEE.

Lampos, V., and Cristianini, N. (2012). Nowcasting events from the social web with statistical learning. ACM Transactions on Intelligent Systems and Technology (TIST), 3(4), 72.

Lampos, V., De Bie, T., and Cristianini, N. (2010). Flu detector-tracking epidemics on Twitter. In Machine Learning and Knowledge Discovery in Databases (pp. 599-602). Springer Berlin Heidelberg.

Lee, K., Caverlee, J., Kamath, K. Y., & Cheng, Z. (2012). Detecting collective attention spam. In Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality (pp. 48–55). Retrieved from http://dl.acm.org/citation.cfm?id=2184316

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. Psychological Science in the Public Interest, 13(3), 106–131.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., & Pearce, I. (2011). The Arab Spring| the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. International Journal of Communication, 5, 31.

Lumezanu, C., Feamster, N., & Klein, H. (2012). #bias: Measuring the Tweeting Behavior of Propagandists. In Sixth International AAAI Conference on Weblogs and Social Media. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4588

Makice, K. (2009). Twitter API: Up and running. Oreilly and Associates Incorporated.

Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter Under Crisis: Can we trust what we RT? In Proceedings of the first workshop on social media analytics (pp. 71–79). Retrieved from http://dl.acm.org/citation.cfm?id=1964869

Mittal, P., Caesar, M., & Borisov, N. (2011). X-Vine: Secure and pseudonymous routing using social networks. arXiv Preprint arXiv:1109.0971. Retrieved from http://arxiv.org/abs/1109.0971

Mittal, P., Caesar, M., & Borisov, N. (2011). X-Vine: Secure and pseudonymous routing using social networks. arXiv Preprint arXiv:1109.0971. Retrieved from http://arxiv.org/abs/1109.0971

Nazir, A., Raza, S., & Chuah, C.-N. (2008). Unveiling facebook: a measurement study of social network based applications. In Proceedings of the 8th ACM SIGCOMM conference on Internet measurement (pp. 43–56). ACM.

Nguyen, N. P., Yan, G., Thai, M. T., & Eidenbenz, S. (2012). Containment of misinformation spread in online social networks. In Proceedings of the 3rd Annual ACM Web Science Conference (pp. 213–222). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2380746

Organisation for Economic Co-operation and Development (OECD). (2007). Participative web: user-created content. Retrieved from http://www.oecd.org/internet/ieconomy/38393115.pdf

Petrovic, S. (2014). Sasa Petrovic's hompage at The University of Edinburg. Online available: http://homepages.inf.ed.ac.uk/s0894589/index.html

Proffitt, B. (2013). What APIs Are And Why They're Important. Online available: http://readwrite.com/2013/09/19/api-defined#awesm=~ouyLxfBlUoNsQT

Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1589–1599). Retrieved from http://dl.acm.org/citation.cfm?id=2145602

Ratkiewicz, J., Conover, M., Meiss, M., Gon\ccalves, B., Patil, S., Flammini, A., & Menczer, F. (2010). Detecting and tracking the spread of astroturf memes in microblog streams. Retrieved from http://arxiv.org/abs/1011.3768

Ratkiewicz, J., Conover, M., Meiss, M., Gon\ccalves, B., Patil, S., Flammini, A., & Menczer, F. (2011). Truthy: mapping the spread of astroturf in microblog streams. In Proceedings of the 20th international conference companion on World wide web (pp. 249–252). Retrieved from http://dl.acm.org/citation.cfm?id=1963301

Rogers, S. and Luckle M. S. (2013, July 10). The Boston Bombing: How journalists used Twitter to tell the story. Twitter Blogs. Retrieved February 17, 2014, from https://blog.twitter.com/2013/the-boston-bombing-how-journalists-used-twitter-to-tell-the-story

Rosenberg, J., & Egbert, N. (2011). Online impression management: personality traits and concerns for secondary goals as predictors of self-presentation tactics on Facebook. Journal of Computer Mediated Communication, 17(1), 1-18.

Russo, A., Watkins, J., Kelly, L., & Chan, S. (2008). Participatory communication with social media. *Curator: The Museum Journal*, *51*(1), 21-31.

Shekar, C., Wakade, S., Liszka, K. J., & Chan, C.-C. (2010). Mining pharmaceutical spam from Twitter. In Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on (pp. 813–817). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5687162

Smyth, T., & Best, M. L. (2013). Tweet to Trust: Social Media and Elections in West Africa. Presented at the Sixth International Conference on Information and Communication Technologies and Development (ICTD2013), Cape Town, South Africa.

Song, J., Lee, S., & Kim, J. (2011). Spam filtering in twitter using sender-receiver relationship. In Recent Advances in Intrusion Detection (pp. 301–317). Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-23644-0_16

Starbird, K., and Palen, L. (2012). (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising. In Proceedings of the acm 2012 conference on computer supported cooperative work (pp. 7–16). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2145212

Stillwell, D.J., Kosinski, M. (2012) myPersonality project: Example of successful utilization of online social networks for large-scale social research. Mobile systems for Computational Social Science (Low Wood Bay)

Stone, B. (2006). Introducing the Twitter API. Online available: https://blog.twitter.com/2006/introducing-twitter-api

Stringhini, G., Egele, M., Kruegel, C., & Vigna, G. (2012). Poultry markets: On the underground economy of Twitter followers. In Proceedings of the 2012 ACM workshop on Workshop on online social networks (pp. 1–6). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2342551

Sunstein, C. R. (2009). On rumors: how falsehoods spread, why we believe them, what can be done. New York: Farrar, Straus and Giroux.

Sunstein, C. R., & Vermeule, A. (2008). Conspiracy Theories (SSRN Scholarly Paper No. ID 1084585). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1084585

Thomas, K., Grier, C., Song, D., & Paxson, V. (2011). Suspended accounts in retrospect: an analysis of twitter spam. In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference (pp. 243–258). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2068840

Thomson, R., Ito, N., Suda, H., Lin, F., Liu, Y., Hayasaka, R., … Wang, Z. (2012). Trusting Tweets: The Fukushima Disaster and Information Source Credibility on Twitter. In Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management. Retrieved from http://www.iscramlive.org/ISCRAM2012/proceedings/112.pdf

Tsotsis, A. (2011). Egypt Situation Gets Worse, People Reporting Internet And SMS Shutdown. Jan 27, 2011 by Alexia Tsotsis. Online available: http://techcrunch.com/2011/01/27/egypt-situation-gets-worse-people-reporting-internet-and-sms-shutdown/

Twitter Official Blog. (2011). One Million Registered Twitter Apps. Online available: https://blog.twitter.com/2011/one-million-registered-twitter-apps

Twitter. (2013). API Terms of Service: Most recent changes. July 1, 2013. https://dev.twitter.com/terms/api-terms/diff

University of Illinois at Urbana-Champaign IRB, 2009

Wagner, C., Mitter, S., Körner, C., & Strohmaier, M. (2012). When social bots attack: Modeling susceptibility of users in online social networks. In Proceedings of the WWW (Vol. 12).

Wallace, P. (1999). *The psychology of the Internet*. Cambridge University Press.

Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M., Zheng, H., & Zhao, B. Y. (2012). Social turing tests: Crowdsourcing sybil detection. arXiv Preprint arXiv:1205.3856. Retrieved from http://arxiv.org/abs/1205.3856

Webopedia (2014). API - application program interface. Online available: http://www.webopedia.com/TERM/A/API.html

Webster, J., and Watson, R. T. (2002). Analyzing the past to prepare for the future: writing a literature review. MIS quarterly, 26(2).

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A Review of Facebook Research in the Social Sciences. Perspectives on Psychological Science, 7(3), 203–220.

Zimmer, M. (2010). “But the data is already public”: on the ethics of research in Facebook. Ethics and Information Technology, 12(4), 313–325.