Following Tweets Around
Informetric methodology for the Twittersphere

David Gunnarsson Lorentzen
# Table of contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>i</td>
</tr>
<tr>
<td>Svensk sammanfattning</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>viii</td>
</tr>
<tr>
<td>Definitions</td>
<td>x</td>
</tr>
<tr>
<td><strong>Part I</strong></td>
<td></td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Twitter: features and API</td>
<td>6</td>
</tr>
<tr>
<td>1.2 The research tradition: information science, webometrics and informetrics</td>
<td>10</td>
</tr>
<tr>
<td>1.3 Motivation and problem description</td>
<td>12</td>
</tr>
<tr>
<td>1.4 Purpose and research questions</td>
<td>13</td>
</tr>
<tr>
<td>1.5 The case studies and their context</td>
<td>14</td>
</tr>
<tr>
<td>1.6 Demarcations</td>
<td>17</td>
</tr>
<tr>
<td>1.7 Relevance for research and society</td>
<td>20</td>
</tr>
<tr>
<td>1.8 Overview</td>
<td>21</td>
</tr>
<tr>
<td>2 Background</td>
<td>24</td>
</tr>
<tr>
<td>2.1 What is Twitter?</td>
<td>25</td>
</tr>
<tr>
<td>2.2 How can Twitter be studied?</td>
<td>26</td>
</tr>
<tr>
<td>2.3 How can political Twitter conversations be studied?</td>
<td>29</td>
</tr>
<tr>
<td>2.4 Specific focus of the thesis</td>
<td>36</td>
</tr>
<tr>
<td>2.5 Special considerations of the project</td>
<td>37</td>
</tr>
<tr>
<td>2.6 Summary</td>
<td>38</td>
</tr>
<tr>
<td>3 Research framework</td>
<td>39</td>
</tr>
<tr>
<td>3.1 What is a conversation?</td>
<td>40</td>
</tr>
<tr>
<td>3.2 Twitter as a filter</td>
<td>43</td>
</tr>
<tr>
<td>3.3 Affordances of Twitter</td>
<td>46</td>
</tr>
<tr>
<td>3.4 Theoretical aspects on sampling from Twitter</td>
<td>50</td>
</tr>
<tr>
<td>3.5 Dimensions of Twitter activity and relationships</td>
<td>53</td>
</tr>
<tr>
<td>3.6 Summary</td>
<td>60</td>
</tr>
<tr>
<td>4 Method</td>
<td>61</td>
</tr>
<tr>
<td>4.1 Filters as methodological problems</td>
<td>61</td>
</tr>
</tbody>
</table>
4.2 Digital methods
4.3 Data collection
4.4 Special considerations of data collection
4.5 Preprocessing of collected data
4.6 Summary

5 Ethical consideration with regards to user provided information
  5.1 “You are what you tweet”
  5.2 Consent or not?
  5.3 Public or private?
  5.4 An ethical position on Twitter data
  5.5 Summary

Part II
  6 Summaries of the studies
    6.1 Webometrics and web mining
    6.2 Polarised conversations
    6.3 All about politics?
    6.4 Approaching completeness
    6.5 Twitter conversation dynamics of political controversies
    6.6 Users in the studies
    6.7 Further testing of the composite method
    6.8 Summary

Part III
  7 Discussion
    7.1 The conversation and its boundaries
    7.2 The filtering platform and its affordances
    7.3 Sampling, bias and completeness
    7.4 What insights can be gained?
    7.5 Which data collection method and when?

8 Conclusions
  8.1 Not reaching completeness
  8.2 Limitations
  8.3 The next step

References
Appendix

Preprocessing Data set 1
  - Coding of participants
  - Converting tweets to mention graph
  - Converting tweets to retweet graph
  - Converting tweets to descriptive statistics

Preprocessing Data set 2
  - Converting tweets to descriptive statistics
  - Extracting hashtags

Preprocessing Data set 3
  - Coding of participants
  - Building threads
  - Finalising anonymisation
  - Converting tweets to descriptive statistics
  - Creating networks from the data sets

Part IV
  - Studies I-V
Abstract

The purpose of this thesis is to critically discuss methods to collect and analyse data related to the interaction and content on the social platform Twitter. The thesis contains examples of how networked communication can be studied on Twitter, based on the affordances of the platform considering interaction with interfaces and other users. The thesis consists of a summary essay and five articles. The first article compares two areas of research that focuses on web structure, usage and content, while the following four focus on different aspects of Swedish political discussions on Twitter. The Twitter studies focus on communication in the form of tweets (public messages on the platform) of three types: retweets (redistribution of tweets), mentions (inclusion of user names, akin to directed messages) and replies.

The main reason for the focus of this thesis is an identified lack of methodological discussion in relation to analysis of interaction, content and relationships on the platform. Twitter research has been based on easily accessible data without introducing or discussing criteria for collecting appropriate samples for a given research task. Data are available to external stakeholders, including researchers, through the platform’s API (application programming interface). The API has different levels of data availability. All the studies in this thesis have been carried out with the kind of access to the API all registered Twitter accounts have. Researchers have often chosen to collect data through the API, or software that communicates with the API, by specifying a range of keywords or a list of users. As a researcher, it is easy to be seduced by the rich material flowing through the platform, accessible through the API, but there are several challenges involved.

The analyses are based on a view of the Twitter platform as a non-neutral filtering gatekeeper. The filtering works in several ways. On the one hand, all users have the ability to filter forward content to their followers through retweets of other users’ tweets. Twitter can hence be seen as a complex system of interlinked individual recommendation systems. Users can both produce content and filter forward content produced by other users. Since the user chooses who to follow, Twitter can appear neutral, but rather, Twitter is based on a popularity model.
Different content and users are treated asymmetrically. On the other hand, Twitter determines what data are available and how data can be accessed through the API. How Twitter provides access to the data in turn affects the analyses the researcher does. The central problem of the thesis is that researchers do not know what relevant data are not collected. Data collection based on keywords, hashtags or user IDs creates data sets that contain fragments of conversations. Another key problem is that it is not possible to obtain a representative sample through the API. As part of the methodological problems, ethical issues are also discussed in detail in this thesis. All materials that the thesis is based on are anonymised.

The four Twitter articles the thesis is based on make use of hashtag-based and user-based data collection methods. Study II made use of the former by collecting tweets containing #svpol (Swedish politics). The analysis focused on the phenomenon of polarisation, which means that users relate to and communicate with mainly like-minded. In Study III the 985 most prominent users of #svpol were identified after an eight-week pilot study. These users were then followed for three four week long periods during one year. The purpose of the study was to find out what other topics were discussed by making a hashtag analysis, focusing on trends and co-occurrences. The main problem for both collection methods is that they do not collect complete conversations around a topic. To solve the problem, a new method was developed for Studies IV and V. By combining the two methods, replies to collected tweets were stored, regardless if they contained a tracked hashtag or not. Study IV evaluated the method and in Study V, an interaction analysis was made on the conversational threads identified in the data set.

The four studies show a complexity of collecting data and analysing relationships, content and activity on Twitter. Through social network analysis in Study II, it was concluded that Twitter users prefer to follow and retweet like-minded, but they also communicate with others. The studies highlight the different behaviour of the various user groups. In Study II it was noted that the least active group is more focused on retweeting while the most active users are more likely to send messages to others as part of a conversation. In Study III and V, the results indicated that the most active users have a stable level of activity while the least active are most active as reactions to sudden and scheduled events. Study IV showed that different users are prominent in one form of communication compared to other forms. Communication networks based on hashtagged replies were found to be potentially
very different from networks based on replies from a more complete data set, where non-hashtagged replies are also included. A network based on hashtagged communication is thus misleading compared to a complete communication network. In the conversational threads in Study V, hashtags were seldom used, which shows that with the hashtag-based data collection method, the analyst risks missing out on highly relevant content. In Study V, it was also clear that the activity after an event was more focused on spreading information while conversations were emerging after the immediate reaction to the incident. The study also revealed that Twitter can mainly be seen as a source of opinions and reactions, but not as a forum for qualified political discussion. Although there were several examples of threads that involved a relatively large number of participants, a large majority of posts in the threads was comprised of opinions and comments that did not invite further conversation.

In addition to the conclusions drawn from each study, the thesis concludes that the use of Twitter in the studied context is dominated by an elite group of users comprised of roughly 1,000 users. Apart from that it is not entirely trivial to identify the parameters to define what should be studied; tests of the API showed that complete data sets cannot be obtained. Therefore, it is important to reflect on both the data collected and the data excluded, not only as a result of the sampling criteria but also what is not given access to. It is also important to be clear about the affordances for interaction that exist when the study is made, both in the user interface but also what API allows and permits.

This research contributes with knowledge about how Twitter is used in the context being studied, but the main contribution is methodological. With the method developed, collection of more complete data sets is enabled, as is analysis of the conversations that take place on the platform. This results in more accurate measurements of the activity. The discovery of conversations that extends beyond the hashtag questions the results presented in earlier work that only considers hashtagged content. Based on the results of this thesis, there are reasons to suspect that previous studies could differ in terms of, for example, results such as communication network size and shape, as well as the type of users that emerges as prominent in the material, compared to if replies that do not contain the studied hashtag had been collected.
Syftet med den här avhandlingen är att kritiskt diskutera metoder för att samla in och analysera data relaterat till interaktion och innehåll på den sociala plattformen Twitter. Avhandlingen innehåller exempel på hur nätverkad kommunikation kan studeras på Twitter, baserat på applikationens möjligheter och begränsningar för interaktion med gränssnitt och andra användare. Avhandlingen består av en kappa och fem artiklar. Den första artikeln jämför två forskningsområden som fokuserar på webbstruktur, -användning och -innehåll, medan de följande fyra fokuserar på olika aspekter av svenska politiska diskussioner på Twitter. Twitter-studierna fokuserar på kommunikation i form av tweets (publika meddelanden på plattformen) i tre typer: återtwittringar (vidaredistribution av tweets), omnämningar (inkludering av användarnamn, ungefär motsvarande riktade meddelanden) och svar.


Analyserna utgår från en syn på Twitter-plattformen som en icke-neutral, filtrerande gatekeeper. Filtreringen fungerar på flera sätt. Å ena sidan har alla dess användare möjligheten att filtrera fram innehåll till sina följare genom återtwittringar av andra användares tweets. Genom detta kan Twitter ses som ett komplext system av sammankopplade individuella rekommendationssystem. Användare kan både producera innehåll och filtrera fram innehåll producerat av andra användare. Eftersom användaren väljer vem som hen ska följa kan Twitter...


De fyra studierna visar på en komplexitet i att samlar in data och analysera relationer, innehåll och aktivitet på Twitter. Genom sociala nätverksanalyser i Studie II, drogs slutsatsen att Twitters användare föredrar att följa och återtwittra likasinnade, men att de kommunicerar även med andra. Studierna visar på olika beteenden hos olika användargrupper. I Studie II noterades att den minst aktiva gruppen är mer inriktad på att återtwittra medan de mest aktiva användarna är mer

Förutom de slutsatser som dragits i respektive studie kommer avhandlingen fram till att användandet av Twitter i den studerade kontexen domineras av en elitgrupp av användare som är ungefär 1000 till antalet. Förutom att det inte är helt trivialt att identifiera de parametrar som ska definiera vad som ska studeras visade också tester av API:et att kompletta dataset inte kan erhållas. Därför är det viktigt att reflektera kring både vilka data som hämtats och vilka som exkluderats, inte enbart som följd av samplingskriterier men också vad som inte getts åtkomst till. Det är också viktigt att vara klar över vilka egenskaper som råder för interaktion vid studien, dels i användargränssnittet men också vad API:et tillåter och möjliggör.

Det här forskningsprojektet bidrar med kunskap om hur Twitter används i den studerade kontexten, men främsta bidraget är metodologiskt. Med metoden som utvecklats möjliggörs insamling av mer kompletta dataset och analys av de konversationer som utspelas på plattformen. Det resulterar i mer exakta mätningar av aktiviteten. Upptäckten av konversationer som sträcker sig utanför mängden av hashtaggade tweets ifrågasätter resultat som presenterats av tidigare studier som enbart beaktar hashtaggat innehåll. Baserat på resultatet från den här avhandlingen
finns det anledning att misstänka att tidigare studier hade kunnat skilja sig åt vad gäller exempelvis resultat såsom kommunikationsnätverkets utsträckning och form, samt vilken typ av användare som träder fram i materialet, jämfört med om även svar som inte innehåller den studerade hashtaggen samlats in.
Acknowledgments

This project has been a long and exciting journey which has finally ended. A large number of people have been influential and supportive during these five years. I cannot name all of you here, but some of you deserve a special mention.

My supervisor, Jan Nolin, was the father of this initially webometric project. I could not have asked for a better supervisor. Your understanding of social media is second to none. Your optimism is always infectious. I would also like to thank my co-supervisors, at first Katriina Byström and then Gustaf Nelhans, for great support.

My colleagues and fellow PhD students at Swedish School of Library and Information Science who have all contributed to a nice and stimulating work environment.

It is always a pleasure to talk about nerdy sci-fi issues with Amira Sofie Sandin, and my friend Ty Nilsson has always made me smile, even through hard times. I really miss you being around at work, and our table hockey battles.

Emma Forsgren and Maria Lindh, who both started their projects pretty much at the same time as I. We have shared the joy of progression, and the anxiety that arrives with such a daunting project.

Anna Hampson Lundh, who not only brought my wife and me together, but has always had sensible comments on anything related to thesis work.

Other people who warrant special mention here are Frances Hultgren and my mother Tina for proofreading, Helena Francke and Björn Hammarfelt for greenreading, Anders Olof Larsson for great opponent work on my final seminar and Daniel Birgersson for designing this amazing cover.

All of my family, but especially my parents who have always supported me, and who encouraged me to continue studying.

My late grandfather Staffan who still inspires me through his never ending endeavour of learning new things.
Finally, I wish to thank my dear wife Susan. You have been and are my biggest supporter, best critic and, most importantly, the love of my life. You made this thesis possible. This is for you.

Last but not least our two wonderful girls Agnes and Laura, who both were born during this project. You are the meaning of life, you are the inspiration. This is for you too.
Definitions

The following are concepts relevant for understanding the terminology used in and around the Twitter Application Programming Interface (API).

- API: Application Programming Interface. This interface enables data to be mined from an application. In this project, Twitter API v1.0\(^1\) and v1.1\(^2\) have been used.
- Edge (arc): The connection (followership, mention or retweet) between two actors in a network graph. Edge is more generally used for both directed and undirected networks, with arc being used only for directed networks.
- Endpoint: A connection to the API, specifying a certain method for accessing data. Two APIs have been utilised to collect data. From the REST API, endpoints for searching for tweets, collecting profile data and list of friends have been used. From the streaming API, the endpoint for streaming by user IDs, keywords/hashtags and locations has been used.
- Firehose (API): A continuous stream of data from a platform. In the case of Twitter this means paid access to 100% of the tweets posted at any given time.
- Follow-on conversation: Tweets not matching the search criteria but related to the collected tweets as replies.
- Followers: Users following other Twitter users.
- Friends: Other users that a Twitter user follows.
- Hashtag: Any word used in a tweet starting with the hash symbol (#), e.g. #svpol.

---

\(^1\) https://dev.twitter.com/docs/api/1

\(^2\) https://dev.twitter.com/docs/api/1.1
• Mention (@mention/at-mention): The mentioning of another Twitter actor by including its username (also called screenname), following the @ character. Mentioning another actor can be seen as a directed message to that actor, and can be used to invoke the actor into the conversations. A type of a mention is the reply, which is a direct reply to a tweet, often starting with a mention of the user replied-to.

• Node: An actor in a network graph.

• REST API: Provides access to different types of data, such as tweets, profiles, friends and followers. This API can also be used to post data on Twitter.

• Retweet: A tweet that redistributes another tweet.

• Singleton: A tweet that is not directed to anyone in particular and is not a retweet.

• Streaming: Making use of the streaming API, which pushes a stream of tweets in real-time to the data collection software.

• Streaming API: Gives access to tweets as they are posted. Public streams, suitable for data mining around a topic or a set of users, have been used here.

• User/Actor: A term that refers to a Twitter account.

• Thread (threaded conversation): A chain or tree structure of a tweet and its replies, and the replies to the replies. The term thread is borrowed from Internet discussion forums.

• Tweet: A message sent on Twitter, but not a private directed message (DM).

• yTK (your Twapper Keeper): The software whose basic architecture was used to build the tool for data collecting purposes. A thorough description of the software is given by Bruns and Liang (2012).
Part I

1 Introduction

A feed flooded with opinions, elements of serious discussions side by side with banality and nonsense. A lot of people are posting, some of them far too often, others very rarely. Too many people are not there at all. Why would we want to analyse social web data? Are not other social science data sources good enough? Behind the potential of analysing social web data is the idea that something different might emerge that can replace or complement other types of data (e.g. González-Bailón, 2013).

This thesis is an investigation of both specific methods and overarching methodological issues regarding the collection and analysis of social web data from the social media application Twitter, by using examples from the context of Swedish political communication. Social media applications are labelled as social media platforms by media professor José van Dijck (2013). A characterising trait for web platforms is that people and organisations can be both producers and consumers of content (produsers, see Bruns, 2007). Communication on the social web is different from its offline counterpart but can also serve as extension of the minds of its users; for some they are a natural part of the everyday life. New media and digital culture professor Richard Rogers states that “software is running social life” (2013b, p. 3) and inquires if the usage of a hashtag or a set of highly retweeted tweets would be able to represent an event.

A constant challenge for a researcher is to find the data that represent the event in an appropriate way. While offline communication is not digitally remembered, the social media activity can be recorded, and assuming that it is similar to offline communication, it can be used to study what people think about various phenomena. The approach in the current thesis is to follow this notion with some caution. Still, on a minimal level, the idea is that social web data represent the
thoughts of individuals about something, even though it does not necessarily mean that these representations equal actual intentions. The approach in the current thesis is to critically investigate methods applied to Twitter with the aim to study political conversations. Although social web data can be seen as representing public opinion to some extent, the current thesis emphasises that such research results also can be seen as effects of methodological choice. An incomplete data set or an unrepresentative sample involves a bias that could be avoided with sophisticated sampling.

Various social web platforms have different characteristics which invite and allow separate forms of usage and expressions. However, in one way they are all similar as they share the networking features, in which the users are related to other users in various ways. The connections in the network can be interest-based as well as relational, and provide context for the content the users produce. Social web platforms connect people with each other in several ways, from the more explicit two-way friend relation and the unidirectional follower relation to the more implicit automated connection through the likelihood of in-between similarity based upon preferences and actions in a social application. van Dijck (2013) labelled these as human connectedness (explicit) and automated connectivity (implicit) and pointed out that connections between people, things and ideas are coded into algorithms by the social media platforms.

Most social media research utilises social web platforms as a resource for making knowledge claims about social phenomena. Contrary to this, the current thesis is about the methodological wrangle involved in collecting, interpreting and processing data harvested from such platforms. It also concerns the underlying methodological assumptions of these kinds of studies, primarily prerequisites for implementing various data collection techniques. In this thesis, the term methodology refers to the study of the techniques used for collecting and analysing data, which in turn are encompassed in the term method. Here we find new and interesting forms of empirical work which, so far, has concentrated on doing research rather than reflecting on the characteristics of social media data collection. A fundamental problem is that different platforms permit various forms of empirical investigations. Therefore, methodological discussions need to be tied to the features attached to individual platforms. In the current thesis, empirical investigations involving Twitter are in focus.
With regard to the number of registered users, Twitter is one of the most popular social media platforms. While other tools can be considered as walled gardens, Twitter is open (e.g. Rogers, 2013b, p. 159). What this means is that Twitter is open in the sense that anyone can access public tweets, not only registered users, and the registered user does not have to be logged in to do so. Partly for this reason, it has the potential of acting as a major player in meetings and interactions between people. When we use it to communicate, it acts as a layer between us which affects how we interact with each other. For researchers, the platform is convenient as data can be collected for free. An example of this is the observation made by Tufekci (2014), that social media big data analyses are dominated by Twitter studies. An indication of how attractive Twitter is for researchers is the number of articles and conference proceedings dealing with Twitter (excluding research areas such as zoology and veterinary sciences). 3,446 such papers published between 2007 and 2015 were found in Web of Science (topic search), and in Scopus the corresponding figure is 8,615 (title, abstract, keyword search). Granted, not all of these papers are studies of Twitter activity, but nevertheless, Twitter has been studied frequently. A striking gap in Twitter research in recent years has been the lack of attempts to bridge the algorithmic development in computer science with the empirical work made in the social sciences. Hence, a need has arisen for developing methods suitable for social science research goals as well as methodological reflections in relation to social science research. This is the general problem area for the current work.

Because Twitter is much studied and has a fairly large number of users compared to other social media web sites, it is important to understand the usage of the platform and what knowledge can be derived from interactions and relationships. Moreover, it appears crucial to investigate current methods and develop these to understand the complexities of what happens on the platform. A central argument in this thesis is that the pioneering Twitter research published so far has been performed with data sets that are limited and fragmented and that a solution to collect and analyse complete conversations has yet to be presented. Twitter makes

---

3 On Twitter a user account can be set to private, which means that the user need to accept those who want to follow the account. Tweets sent by a private account are only visible to the followers of that account.
use of hashtags (# followed by a text string) for aggregating tweets into a topical stream. Collecting tweets including a given hashtag is a common approach to Twitter research but has a major drawback in that replies without the chosen hashtags cannot be collected. The drawback was acknowledged by the media and culture professor Axel Bruns (2012), but only a few solutions to the problem have been presented (Cogan et al., 2012; Zubiaga et al., 2015). For a researcher, such tracking restricts analysis to the usage of a particular hashtag with potentially significant parts of a communication, the follow-on communication\textsuperscript{4}, not being covered at all (e.g. Bruns & Moe, 2013).

The issue of the incompleteness of such methods of acquiring data is the most important focus in this thesis. The thesis acknowledges the importance and usefulness of data-driven and exploratory methods while also taking the same position as Gaffney and Puschmann (2013) in that research on Twitter should not be tailored to easily accessible data. Pioneering Twitter research has tended to sample through easy accessibility without introducing or discussing criteria for sampling, or strategies for acquiring a sample that is appropriate for the given research task. Instead, there is a frequently taken for granted sense of capturing complete data sets, i.e. “N = all”.

As Twitter is based on a followership model, any user can filter forward (Weinberger, 2011, p. 10-11) information. This means that Twitter can be seen as a complex system of interconnected individual recommendation systems. Users have therefore a dual function of both producing information and filtering forward other user generated content deemed valuable for their network. By following other users, Twitter users create their own filters, resulting in a social network in which the information flow can be calibrated to some extent by the users themselves, depending on their positions within the network. Seemingly, Twitter is a neutral platform, i.e. users are in control of their filters. However, van Dijck (2013), who views Twitter as a mediator, argues that this is not the case. For example, the filters are partly steered by programmability and popularity (van Dijck & Poell, 2013), i.e. information flows are subtly steered by a hierarchical star system.

\textsuperscript{4} Different terms have been used for this, for example follow-on tweets and follow-on communication. In the remainder of the text, the term follow-on conversation is used.
Another interesting feature is that information flows are community-based. In order to discuss community formations on Internet-based platforms, researchers have developed numerous concepts such as **homophily**, **polarisation**, **echo chambers** and **filter bubbles**. These concepts, developed within domains such as sociology, psychology and political science, are used to describe possible consequences of filters that tend to cluster together according to principles of like-mindedness. Homophily, polarisation and echo chambers are all possible consequences of user created filters that tend to cluster together according to principles of like-mindedness. Homophily is the degree to which people in a given context are similar in one or more aspects (e.g. Rogers, 2003, p. 305). Polarisation is the effect of echo chambers, communities or groups which are created as people choose to interact with like-minded (e.g. Sunstein, 2009, p. 60). Filter bubbles are similar, but these are the results of filters created by recommendation systems (e.g. Pariser, 2012, p. 9). These concepts will be discussed in depth in Chapter 3, **Research framework**.

It is a reasonable assumption that applications controlled by recommendation systems will, over time, filter out more and more of the information perceived to be contrary to the recipient’s opinions. During the time span that data for this thesis were collected, Twitter was not steered by such a system, as the filter was at least partly controlled by the users themselves. However, over time, Twitter introduced features such as recommendations on who to follow and other popularity-based algorithms. As democracy requires that citizens are enabled to see things from different perspectives (e.g. Sunstein, 2009), it is of interest to investigate whether social media such as Twitter offer or even support this, and if so, whether citizens would be willing to communicate with other groups. Having said that, the current text is not directly concerned with issues of Twitter and democracy but rather it is concerned with the methodological issues which underpin research with such a focus.

This compilation thesis is based on one literature review and four empirical Twitter studies. The focus is on method development and it makes use of different studies of political communication on Twitter in a Swedish setting as exemplifying cases. The notions of homophily, polarisation, echo chambers and filter bubbles are here studied with different methods and on separate units of analysis. The units are conversations, relationships, mentions of other users and the redistribution of other users’ messages. The methods used include statistical analysis, content analysis and social network analysis. The character of the research problem sets the stage.
for the choice of method. In this research project, the methodological issues of analysing how different aspects of usage of social web applications appear are in focus. The challenge is to make sense of the big data that flows through the Twitter Application Programming Interface (API) (see Twitter, 2015a).

1.1 Twitter: features and API

As a platform, Twitter has specific characteristics that afford the interplay of actors. At the basic level, it facilitates relationships (follower/friend), undirected messages (singleton), directed messages (replies), mentioning other users (mentions), forwarding of messages (retweets) and adding metadata to the messages (#hashtags). Drawing on this, we can infer who is exposed to what message about a given topic. All undirected messages are intended for the followers of the tweeting user or the followers of any hashtag included in the messages. It is also possible to identify which messages are spreading and how they spread across different user groups and what is done with the message over time. Communication on Twitter is labelled as conversation by the owners of the platform. It seems to be an important concept for the platform which can be seen in one of the later modifications of the service. In June, 2015, Twitter attempted to make conversations easier to follow through a series of modifications such as grouping conversations together and highlighting “the most interesting exchanges” around a tweet (Twitter, 2015c). The label conversation is not necessarily in harmony with definitions of conversation in relation to deliberative democracy or preferred definitions from social science research.

Nevertheless, the conversation concept has been adopted by Twitter researchers (e.g. Bruns, 2012; Bruns & Highfield, 2013; D’heer & Verdegem, 2014; Holmberg et al., 2014; Larsson & Moe, 2014; Renz & Sullivan, 2013; Wang, Wang & Zhu, 2013) and, in some cases, used in conjunction with notions of deliberative democracy or related concepts (e.g. Freelon, 2015; Larsson & Moe, 2013; Pond, 2016; Sæbø, 2011). Specifically, this thesis uses conversation as the label for the interactions that take place through the reply function. If we take the discussion forum as an example, the full discussion thread is arguably a demarcated conversation. The definition of a Twitter conversation used in this thesis is all the connected tweets which are contextually related through the reply metadata field.
Although this thesis is not about evaluating the conversations on Twitter or how well Twitter facilitates conversation given the ideals of deliberative democracy (e.g. Fishkin, 2011) or other similar or related forms of democracy, deliberation is used in relation to discussions of the conversation concept, as a contrasting example to the Twitter conversation. The problem at hand is that if Twitter is to be assessed from a democratic perspective, then the research must be based either on the complete conversations or on a sophisticated strategy involving explicit sampling. While some solutions to the collecting of conversation problem have been proposed previously (Cogan et al., 2012; Zubiaga et al., 2015), complete conversations in the sense defined above have not been analysed. Moreover, the methods proposed by Cogan et al. (2012) and Zubiaga et al. (2015) have drawbacks which are pointed out in Study IV.

1.1.1 Twitter as big data

Arguably, when it comes to the size of data processed, Twitter fulfils all the four main characteristics of the notion of big data. Although different definitions of this concept have been proposed, it is characterised, according to one widely accepted notion by four aspects; velocity, volume, variety and veracity (e.g. Ning et al., 2015). On Twitter, there is a large and continuous flow of tweets on varying topics and from different types of users. Questions can also be raised regarding the truthfulness of tweets, whether they are being posted by software created for producing artificial content or not, or the possibility of sarcastic and ironic content. A definition that appeals to this project is that big data is too big for manual methods, hence the need for automated methods. Social media researcher danah boyd and her associate Katie Crawford, argue that treating big data is a question of technological capacity and the analytical skills required to deal with large data sets in various ways, but big data also has to do with mythology: “the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy” (boyd & Crawford, 2012, p. 663).

A slightly different view on big data was taken by sociologist Sandra González-Bailón (2013), who emphasises the richness of data in terms of details and resolution rather than size. Big data studies certainly have their merits, but their value has been contested in relation to the value of insights made from small data (e.g. boyd, 2010a, 2010b; boyd & Crawford, 2012). As boyd puts it:
Big Data presents new opportunities for understanding social practice. Of course the next statement must begin with a ‘but.’ And that ‘but’ is simple: Just because you see traces of data doesn’t mean you always know the intention or cultural logic behind them. And just because you have a big N doesn’t mean that it’s representative or generalizable (boyd, 2010a).

This quote highlights the lack of the “why” and issues related to sampling. Adding to this, González-Bailón (2013) argued that the ways in which big data can be reduced by filtering and aggregating data makes the data interesting. She highlighted the need for theory to eliminate noise from the data and provide context for interpretation. Hence, if a large stream of tweets matching a hashtag is collected, questions should be raised regarding what the sample represents, how well the users in the sample represent the population and, perhaps most importantly, what relevant data are not collected.

1.1.2 Twitter and conversations

Clearly, it is easy for a social media researcher to be seduced by the rich material flowing through the Twitter platform. Along with data collection several challenges arise, and the straightforward access to large quantities of data has prevented researchers from analysing the tweets in the context of the complete conversation. Arguably, as noted above, most research on Twitter usage has focused on subsets of conversations, either by collecting tweets including a hashtag or other type of keyword, tweets posted by or to a set of users or tweets with geographical information matching some criteria defined by the researcher.

There is a very plausible explanation for this focus on subsets of conversations within earlier research. The Twitter API is a gatekeeper through which a finite set of access points are given. Some of them allow querying for tweets matching a keyword while others require geographical coordinates or usernames/user IDs. None of them can be utilised alone for capturing a threaded conversation as in the discussion forum. This is partly due to the API being real-time centred. There is no archive to retrieve tweets from, even though a complete archive could be created with firehose access to the API (paid access to the full stream of tweets). For most researchers data collection is carried out in real-time; activity is recorded continuously. It is not possible to query the API with an identifier of a tweet to collect possible replies to it, and it would not be feasible either, as replies can be
made at anytime. It is not possible to subscribe to a set of tweets for collecting replies. Understanding the API is critical for understanding the data it returns, and the same applies to software designed to work with the API. Given these API issues, there is a need for the scrutiny and development of existing data collection methods.

Both the API and the software are mediators that determine what data can be collected and how. Seen in this way, they are gatekeepers in the sense that they filter the data accessible to a researcher. It is important to critically assess methods and software involved in the research. It is also important to reflect on what questions can be asked and what knowledge can be gained using the various methods of collecting data from Twitter, as well as the various entities of the platform, such as the mention, the retweet, the hashtag and the URL. If we do not have firehose access, we need different and adapted approaches to investigate the usage of the platform. One of the tasks of this thesis is to investigate the problem of collecting conversations without firehose access.

The API presents researchers with both opportunities and problems. Sudden non-scheduled events might be very interesting but also difficult to study due to the real-time centeredness. It is tantalising for researchers to study events that are anticipated and scheduled, such as elections, so that data collection can be planned in advance. Sudden events are difficult to prepare for, and they can also interrupt and disturb a study of general activity. Such an event might result in a turn in the conversations. Hence, it is very likely that there is a bias towards scheduled events in the Twitter research.

We can understand scheduling on Twitter with the aid of the concept programmability, which in relation to schedules refers to constructed hashtags for the purpose of aggregating opinions into one stream of tweets. This also includes the utilisation of spin doctors and bot programs, instructed to produce content at given hours of the day. In this thesis, the general activity around one such hashtag is covered as well as the sudden event called the December Agreement, an agreement between six Swedish parties which was triggered by the failure of a budget proposed by the government.
1.2 The research tradition: information science, webometrics and informetrics

Information science and library and information science are more than just science about information. The fields are broad and relate to many other fields and disciplines. There is a lack of agreement about what information science is about (Robinson & Karamuftuoglu, 2010). Stock and Stock (2013, p. 3-4) highlighted representation, storage and the supply, search for and retrieval of documents and knowledge. Furner (2015, p. 375) argued that if information science was a science about information then the main objects of study would be “information-as-data and systems of data production, transfer and use”, but within the area, research is also concerned with related objects, activities, practices and people. Information science, and the broader discipline library and information science, are seen as inter- or multidisciplinary fields (e.g. Cronin, 2008; Sugimoto, Ding & Thelwall, 2012). Potential weaknesses of the discipline, such as vaguely defined demarcations, an absence of an agreed core, fragmentation and a diverse range of interdisciplinary collaborations, can also be seen as strengths (Nolin & Åström, 2010).

A sub-field of information science is informetrics. It is defined by Tague-Sutcliffe (1992, p. 1) as “the study of the quantitative aspects of information in any form, not just records or bibliographies, and in any social group, not just scientists” and categorised as empirical information science by Stock and Stock (2013, p. 5), whose categorisation scheme also includes theoretical and applied research. Bar-Ilan’s (2008) review highlights the many (variants of) methods developed within informetrics and its related areas. Webometrics was born as informetric techniques were applied to the web (Almind & Ingwersen, 1997) and has inherited a tradition of combining method development with empirical studies. The field was defined as the “quantitative study of web-related phenomena” (Thelwall, Vaughan & Björneborn, 2005, p. 81). Thelwall (2009) later played down its informetrics heritage and re-focused webometrics towards the social sciences.

The sub-field web mining has emerged within the discipline computer science. Web mining is centred on discovering knowledge from the three web aspects structure, content or usage data (e.g. Liu, 2008), aspects which are used in webometrics as well (e.g. Björneborn & Ingwersen, 2004). The sub-field information retrieval, which is connected to both computer and information science (e.g. Robinson & Karamuftuoglu, 2010) is seemingly successful in inter-
disciplinary collaboration. Could webometrics and web mining do the same? This question led to Study I, which was based on the hypothesis that these two fields could potentially benefit from closer collaboration. Webometrics and web mining differ in several ways. Web mining appears to be more experimental and instrumental and the volume of web mining research is much larger than webometric research. Web mining is heavily dominated by methodological studies, whereas webometrics is equally dominated by exploratory and empirical case studies. However, it is not uncommon for webometric research to combine the development of methods with exploration.

Webometrics and web mining are not the only fields devoted to studies of the web. There are many examples of research of different aspects and usages of the web that have been carried out under other labels than these two. These researchers often use of software created by others as data collecting tools. Black boxes are artefacts taken for granted (e.g. Sismondo, 2004). If a piece of software is used as a black box there is a risk that the software effectively restricts the research questions. Study I concluded that programming skills are needed to make use of big data for social science research goals. The lack of overlaps between webometrics and web mining, and the lack of method development and methodological reflections in social science Twitter research (see section 2.3 of this thesis) leave method questions in a vacuum. Computer scientists that are proficient in Twitter methods are seldom interested in exploring the research questions posed by social scientists. Social scientists, on the other hand, often lack the knowledge required to collect or analyse the data needed in order to answer social science research questions. To fully understand social media we need insights and perspectives from different disciplines, but combining these is sometimes hindered by the disciplinary boundaries and the lack of compatible vocabulary and methodology (van Dijck, 2013, p. 43). Thelwall and Wouters’ (2005) view of the information scientist as a data evaluator, method developer and a broker of social science methods in a metadisciplinary context is fruitful for this project.

As well as informetrics, webometrics has had its fair share of method development, with information scientist Mike Thelwall as an important contributor (e.g. Thelwall, 2001; 2011; Thelwall et al., 2010; Thelwall, Vann & Fairclough, 2006). Inspired by the works of Thelwall, this thesis builds on and extends the pioneering work by Bruns and associates. By making available a series of scripts for processing Twitter data, Bruns and Burgess (2011b) opened a research area for
many non-technical researchers. However, there is a need to further develop the research area by method developers. To transfer the methods to other settings, the method developers also need to act as method brokers, hence the method focus in this thesis. With a combination of previous experiences in web usage mining and programming knowledge I have found a strong and solid methodological foundation, thereby avoiding the black box problem which arises when researchers use software they do not understand how it works in relation to the data source.

The title of the thesis suggests that I position this work within informetrics and not within webometrics. The definition used here is similar to the one by Tague-Sutcliffe (1992) that informetrics concern mainly quantitative aspects of information, but in this thesis it is restricted to networked communication on a commercial platform. Twitter is not only a web application as it is accessible through apps, SMS and the web, and from different devices. Also, while some of the work performed here is clearly in line with Thelwall’s definition from 2009 and can be considered as the application of (mainly) quantitative methods on a social science problem, the focus of the thesis is on method development and methodological discussions. Considering the pure-applied dimensions of the taxonomy presented by Becher and Trowler (2001), which was used in Study I, I find informetrics less applied than webometrics, and more focused on methodology. Whereas applied knowledge is more inclined towards results and techniques, pure knowledge is oriented towards understanding and interpretation. Both of these fields arguably move between pure and applied, however.

1.3 Motivation and problem description

If we accept that Twitter is an important platform for communication there are good reasons for focusing research on the activities, interactions and relationships on Twitter to better understand them. From an information science perspective, it is important to study what the users make of this forum, and from an informetric perspective, it is important to develop methods that are suitable for the research questions rather than posing questions adapted to the available data. To be able to perform analyses, the understanding of the nature of the platform and its usage is crucial. To be able to do accurate measurements, we need access to as complete data as possible, which are collected based on well-defined sampling criteria. When we talk about Twitter activity in different contexts, we need a
comprehensive rather than a fragmented picture. This research project provides a comprehensive picture of the activities within a given context, as well as the methods to acquire such a picture.

The vast majority of Twitter studies that have been published have relied on a limited and biased sampling method. There are several data collection issues in relation to Twitter. Firstly, we are unaware of data withheld by Twitter (e.g. boyd & Crawford, 2012). Secondly, collecting only those tweets matching a hashtag or a keyword, or are restricted to a set of user accounts, collects partial conversations. Thirdly, a representative sample in the traditional social science sense cannot be collected from Twitter for various reasons. We cannot be sure that a sample is random (boyd & Crawford, 2012; González-Bailón, 2013) and the user base on Twitter is not representative in relation to the whole population (e.g. Barberá & Rivero, 2015). The three data collection issues act as basis for the problem studied in the thesis and are further discussed in Chapter 7.

As it is easy to be misguided by the straightforward access to data, this project seeks to explore what insights can be made through the use of different methods. The overarching problem can be summarised as how to develop and apply methods to capture the activities within the clustering around political hashtags and how such investigation can be understood. The problem involves both epistemological and ontological opportunities and constraints within political Twitter communication.

1.4 Purpose and research questions

The purpose of this thesis is to involve critically in methodological discussions about specific methods for collecting and analysing Twitter interactions and content, based on the identified affordances and limitations of the application. The research questions are:

- RQ1: What kind of problems for collecting and analysing data can be identified within contemporary research on Twitter-based political communication?
- RQ2: In light of existing difficulties, which kinds of approaches can be developed in order to improve on current research practices?
The Twitter API can be used in different ways to collect data. To collect tweets as they are posted, the streaming API (Twitter, 2015f) can be utilised by matching keywords, users, and geographic locations or by sampling 1% of the stream. With the focus on a topic, the two latter are not utilised here. Instead, these questions deal with keyword-based and user-based methods. The endeavour to collect complete data represented by Study IV has triggered further exploration of the Twitter API. The API documentation (Twitter, 2015d) shows how two different parameter types can be combined. In this case, the parameters are hashtags and user IDs. The questions are directed to the investigation of the potential for development with regards to completeness in the chosen setting.

- **RQ3**: What could be the relevance of such methodological contributions for other Twitter research investigating different contexts?

The discussion around sampling and completeness is extended by an additional data collection, comparing the Swedish political setting with the Australian by collecting data with different parameters. This kind of method triangulation makes it possible to assess the completeness of the data set. Moreover, the thesis discusses potential uses of the methods outside the studied cases. An important part of the thesis is related to the knowledge that can be derived from the collected data given the data collection method. There are potential cases where the hashtag-based or user-based methods are sufficient and follow-on conversation is not needed, although such data sets are fragmented. The thesis discusses when and in what context these three methods are most suitable.

### 1.5 The case studies and their context

The studies in this thesis act as examples of method development and as such the thesis differs from what has been done in most of the research cited. I make use of the same type of material as the cited research but from a methodological angle. To put the reader in the picture, the research that makes use of Twitter to study democracy and politics will be shortly presented. The distinction between the empirical research carried out by others and the empirical aspects of the studies in this thesis including its methodological focus is important, partly because the same or similar concepts are used are used in both, even though research interests differ.
For the purposes of this thesis, it is useful to look at social web platforms as technical solutions to social and political problems. One such problem is the distances between policy makers and the public. With the emergence of the Internet and its platforms a new kind of public sphere was created, an arena for discussion that did not exist before. This arena has over the past half-century become increasingly dominated by commercial interests. Numerous works on the promise of the social web and its democratic potential have been published (e.g. Brown, 2009; Jenkins, 2008; Shirky, 2009; 2011; Tapscott, 2008), but as the applications have emerged and grown, this democratic potential has been much debated and questioned (e.g. Ellison & Hardey, 2014; Fuchs, 2014; Morozov, 2012).

Twitter’s democratic potential has been questioned too (e.g. LaMarre & Suzuki-Lambrecht, 2013; Larsson & Moe, 2013; Larsson & Moe, 2012; Sæbø 2011; Yardi & boyd, 2010), and several studies have found that a minority of users account for the largest share of messages (e.g. Barberá & Rivero, 2015; Bruns & Highfield, 2013; Bruns & Stieglitz 2013b; Tumasjan et al., 2011). However, the fact that many more users that post something, implies the existence of larger discussions. It is difficult to assess the size of the group of “lurkers” that only use Twitter as an information source, but it is very likely to be much larger than the group of users posting at least once. In 2011, Twitter estimated that up to 40% of its visitors were lurkers (Steen-Johnsen & Enjolras, 2015, p. 128).

Despite this estimation and the openness of the platform, only 5% of the Swedish population consulted Twitter on a daily basis during 2014 (Nordicom-Sveriges mediebarometer: 2014, 2015). Hence, the setting in which all the studies in this research project are carried out can best be described as an interest-based, elite-centred, digital social network produced on a commercial platform. Such a setting can be labelled as a political Twittersphere, which is a sphere of communication (Ausserhofer & Maireder, 2013). The political Twittersphere can be artificially demarcated from the entire Twittersphere either by a set of political users (Ausserhofer & Maireder, 2013), or by a political hashtag (Larsson & Moe, 2012). Prominent actors within the sphere can be seen as opinion leaders (Rogers, 2003) or perhaps cognitive authorities (Wilson, 1983). However, such artificial demarcations make it difficult to understand how they are situated in a larger context, and without asking their followers how they are influenced, both opinion leadership and cognitive authority are outside the scope of this thesis. Instead, the thesis makes use of the concept elite users, which includes traditional elites within
the setting, such as politicians and mass media actors, and early adopters (e.g. Larsson & Moe, 2012). Twitter was initially more of a communication utility, but has since moved towards a celebrity focus, where popularity is considered important (van Dijck, 2013). This entails that elite users are more visible. It is the clustering around these elite users within the political Twittersphere that are investigated in this thesis.

As noted above, this thesis is based on five articles. The first article compares two research fields devoted to web content, structure and usage. The other four are performed in one setting, making use of three different data collection techniques. The setting chosen is communication on Twitter around the topic of Swedish politics. Political usage, contents and topics have been much studied on Twitter so far (see Chapter 2). The choice of using political communication as case studies in this research is primarily motivated by the abundance of data, the expected diversity of viewpoints, and the presence of different user types, such as ordinary citizens, journalists and politicians. Twitter has among other things been used as a tool for a political discussion, although this is not what it is originally shaped for. What makes it particularly interesting for the purposes here is that it is comprised of communities of interests. Its members create hashtags and gather around these to participate in discussions. The discussions can be easily followed by both members and non-members of Twitter.

A number of studies have been used as stepping stones in this project. Initially, when the project was designed, Twitter research on Swedish and Scandinavian political conversations had focused on how politicians interact with the public (e.g. Grusell & Nord, 2012; Sæbø, 2011), but Larsson and Moe (2012) were also interested in the “high-end” users of Twitter. Grusell and Nord (2012) and Larsson and Moe (2012) focused on election times. This project puts focus on the communication patterns on a general political topic outside elections (Study II), the topics discussed apart from politics (Study III) and the conversations following a major political event (Studies IV and V). A political topic is here defined as tweets including a hashtag used for political and politics related talk, and the follow-on conversation of these tweets. The empirical material used for this project was collected as three different data sets by taking the hashtag as a starting point. A set of tweets containing a hashtag is perhaps best described as an ad-hoc public (Bruns & Burgess, 2012). The use of a hashtag could be a conscious attempt to tie the message to a conversation, although there is no guarantee that all users of hashtags follow the conversations (Bruns & Moe, 2013). Dormehl (2015, p. 247) underlines
that the point of using hashtags is more for the search engine algorithms than other people.

Conversations can evolve around hashtags. The hashtags act both as topical and non-topical keywords that aggregate a stream of tweets. Arguably, the ad-hoc public this aggregation creates differs from more intuitive conversation solutions such as the discussion forum. The research project this thesis is based on started with a data-driven and exploratory analysis of the relationship and communication patterns of the general Swedish politics hashtag #svpol (Study II, Data set 1). Since then, the methods have been developed continuously during the research project. When analysing the patterns among relationships, conversational tweets and retweets the question of to what extent conversation beyond the hashtag (follow-on conversation) exists in this setting was raised. To be able to collect this follow-on conversation a modification of the data collection software was needed, leading to an eight week long data collection aiming to aggregate a set including tweets matching #svpol and its follow-on conversation. The data set was then used to identify the most prominent participants in the conversations, which were followed through the API during three four week periods over one year in Study III (Data sets 2a, 2b and 2c). The initial aim was to also analyse conversational threads within the captured data set, but some issues with the data collection method (discussed in Chapter 4) required refining of the method. The new method was tested and evaluated on a new data set in Study IV, and the thread analysis was finally made in Study V (Data set 3). The method for collecting conversation is both user and hashtag/keyword based. Hence, I refer to it as the composite method hereafter. All of these approaches, the hashtag-based, the user-based and the composite method have their limitations and affordances. An overview of the studies and data sets is given in Chapter 6.

1.6 Demarcations

Studying Twitter with available tools involves making numerous informed decisions. Arguably, this is more significant than in other types of social science related studies, both qualitative and quantitative. In the study of Twitter, there is an overload of data and it is possible to pursue many alternative tracks and opportunities. A PhD project is a long journey and requires focus. Therefore, an
important part of the project was to develop specific guidelines to support ongoing
decision-making. These limitations are described below.

Data types are restricted to the contents produced within the defined topical space
(tweets) and relationships among the participants within the topic (followership).
Among the tweet metadata, the following fields were used:

- URLs (added after Study II),
- Hashtags (added after Study III, derived from tweet content in Study III),
- RT of tweet (this was derived from tweet content in Studies II-IV),
- Profile data (anonymised),
- Language code (mainly for spam detection),
- Tweet replied to (added after Study II) and
- Lists of friends (users followed by a given user).

An aspect not studied is the usage of lists, which is somewhat similar to friending
with a slightly different meaning. Ideally, lists should be compared with friend lists
as users undoubtedly use them in different ways. Followers have not been collected
as a pilot study gave clear indications of such an exercise being time consuming.
The reason for this is that few user accounts are followed by large numbers of
users, requiring a wealth of requests to the API. Friend numbers are more evenly
distributed and are generally lower than follower numbers, making such data easier
to collect. As followership studies have all been based on a fixed set of user
accounts where relationships between the participants have been outlined, friends
were collected instead. For each user account, all friends were collected but only
those relationships with another user in the fixed set were stored. Arguably, the
considerations taken due to the state of the API is one example of tailoring
research questions to data availability. This appears to be a major problem for a
multitude of approaches exploiting big data and it emerged as a problem for the
project at hand, and is an issue addressed in the thesis.

Data have been collected based on hashtags and users, and not based on
geographic coordinates. Methods that are used to understand the intention of an
action (such as interviews, or questionnaires) are not used here, because intentions
are not the focus of investigation in any of the studies. The focus of investigation
lies on how actors relate to each other (structure), the content the actors produce, and the usage patterns within the conversations.

The method and algorithm development in web mining focuses on efficiency in the collection and analysis of the data, storage or modelling issues and spam detection, as shown in Study I. This project focuses on completeness and how to analyse the collected data, and thus it differs from what is emphasised in web mining. Twitter spam does exist in various ways, for example in the form of fake non-human followers and bot accounts posting tweets that include popular hashtags and usernames. Spam has been dealt with in all three studies, with a combination of manual and automatic methods. In Studies I and II, hashtag spam was identified and removed in the pre-processing steps. In Study III, spammers were also identified through trend analysis based on user segments derived from activity. Sudden peaks in the activity of the least active group indicated the introduction of spammers to the data collection method. However, none of these were included among the elite users the study focused on. In the comparison of the hashtagged set and the follow-on set in Study IV, spam was not removed as the purpose was to compare the two sets. It was found that possible bot accounts did use the hashtag to tag celebrities into the conversations. However, in the threads analysed in Study V, no spam was present. Finally, an advantage with collecting friends lists instead of followers is that spam followers are likely to be excluded from the data (this requires that the selected users are not spammers).

While only 5% of the Swedish population visit Twitter on a daily basis, the content is not only published on this platform. Twitter is a complex platform which is part of a larger ecology, and tweets are referred to in both print and online mass media. It has been engineered to fit other social media platforms (e.g. van Dijck, 2013), so it is possible to publish tweets in other settings, exposing them to non-Twitter users. What happens on Twitter is also visible to some extent in other spaces and settings. The platform is challenging to explore and understand and research of its different aspects involves several methodological problems. A political Twittersphere is woven into other political discussions, however, in this research project, the focus is on one platform.
1.7 Relevance for research and society

The research project is part of the multidisciplinary research programme Social Media Studies at the Swedish School of Library and Information Science (University of Borås, 2015), and is based on a view that emphasises the benefits of multidisciplinary research. Such benefits include collaboration with other related fields (Nolin & Åström, 2010), the possibility to act as brokers of social science methods (Thelwall & Wouters, 2005) and the generalist approach to communication (Robinson & Karamuftuoglu, 2010). The target reader group for this thesis is primarily comprised of information scientists, but the methods should be of interest for other disciplines as well as they are transferable to other contexts. Political researchers, as well as politicians and journalists are reasonable target groups. In this context, there are two major opposing political parties (the Social Democrats and the Moderate Party), and recently the two coalitions of parties (the Red-Greens and the Alliance). A challenger, the Sweden Democrats, a party emphasising tougher immigration policy, has entered the arena during recent years, but the more established parties have actively tried to block their influence. This kind of conflict is transferable to other political contexts, i.e. what does Twitter say about politics and democratic processes?

The methods developed can also be used to investigate and discuss the activities of and around prominent actors, for example potential opinion leaders in beneficial network positions, regardless of context. Actors in beneficial network positions can be information rich, capable of spreading information or both. Knowledge of this is useful for creating relationships based on the purpose of Twitter participation. It might also be relevant to understand what characterises these actors, if they become important because of offline status, early adopting, tweeting behaviour, and so on, and their roles in and relations to other Twitterspheres.

Eveland, Morey and Hutchens (2011) argue that the emphasis in studies of informal political conversation has not been on the process of communication. They stress that focus on the process of communication is the only way to understand such conversation. If the incompleteness issue is dealt with, complete conversations can be collected and studied. Another point of relevance is the possibility of archiving conversations for future, historical studies. For an archive to represent the conversations in the best possible way, methods must be developed to satisfy the completeness requirement.
The research project develops methods for collecting and analysing more complete conversation than is the case in traditional methods, and they can be applied in various settings. The investigation will furthermore demonstrate various ways of working with Twitter data in the analysis of social science related problems and of understanding interactions between its users and the content users produce, by using data sets that in some aspects are too big for manual methods, but in other aspects suitable for manual analysis.

The methods can be used by political researchers to understand echo chambers and cross-boundary conversation, as these are important aspects of democracy discussions. As noted above, several researchers question the democratic potential of Twitter (e.g. LaMarre & Suzuki-Lambrecht, 2013; Larsson & Moe, 2013; Larsson & Moe, 2012; Sæbø 2011; Yardi & boyd, 2010), but due to issues of sampling such conclusions have often been drawn from the analysis of incomplete data sets. What is yet to be done is an analysis of the quality of complete conversations from a democratic perspective, as well as the development of the conversations from year to year, as the characteristics of the conversations might be a matter of maturity in the usage of the social web applications.

1.8 Overview

The thesis starts with a summarising essay which describes the context for the studies and includes summaries of the studies, a discussion and conclusion, thereafter follow four peer-reviewed articles and one manuscript. As previously outlined, the four Twitter studies are based on the usage of hashtags. Based on the findings and methodological reflections of Study II, in which the communication patterns of #svpol were analysed, the project was then refocused towards methodology and the aim of collecting complete conversations. The main reason for writing a compilation thesis was to be able to publish results when they were fairly up-to-date, thus allowing for feedback on the relevance of the topic while progressing to the next study. The thesis consists of four parts:

I. Context for the studies, methods and research ethics (Chapters 1-5),
II. The studies summarised (Chapter 6),
III. Discussion and conclusions (Chapters 7-8) and
IV. The articles.

Chapter 2 supplies a background to the study, starting with a description of Twitter based on previous research. The chapter builds on an extensive literature review of research papers about Twitter studies related to politics, focusing on methods used to access and analyse data. The review also highlights some aspects analysed in the studies and used in this thesis. As computer science is mainly focused on developing and evaluating algorithms and not on applying the methods to social science problems (as shown in Study I), this discipline was excluded from the literature searches. The chapter also includes reflections on how researchers relate to the concept of conversation and how it can be studied. Building on previous research, the specific focus of the thesis is outlined. The chapter closes with the special considerations of the project.

Chapter 3 contains a research framework. This is divided into five main parts: 1) what a conversation is, 2) Twitter as a filtering mediator, 3) affordances of Twitter, 4) theoretical aspects of sampling and 5) dimensions of Twitter activity and relationships. The chapter starts with theoretical notions of conversation in relation to deliberation and deliberative democracy. With the technical definition of conversation as it is used by Twitter and in this thesis, it is relevant to relate such a definition and the findings of the thesis to other theoretical notions of the concept. As noted above, Twitter cannot be seen as neutral. In this chapter, the different ways in which Twitter filters data, both through its API and through its users, are discussed. Twitter is here seen as a gatekeeper that has certain affordances which shape research on Twitter activity as well as the activity itself. Then, a discussion about sampling follows, focusing on data complexity issues, black boxing through software and convenience sampling, which is the usage of the API in a way that is straightforward but results in incomplete data. Finally, the chapter outlines important aspects related to social network theory, how the underlying followership model shapes communication on the platform and boundary issues of conversations.

In Chapter 4, methodological issues are discussed, taking its point of departure from the notion of digital methods (Rogers, 2013b). This leads to the introduction of two types of results, one related to method development and the other to analysis of the actual activity and relationships on the platform. Specific attention is given to data collection issues, and what consequences different choices have. The chapter also includes notes on preprocessing of collected data.
Chapter 5 consists of a discussion on ethical aspects. In this discussion, legal and ethical issues relevant for the data collection and analysis are outlined and discussed. The chapter also includes different views on ethical aspects on social media data from researchers, as well as the *Recommendations from the AoIR Ethics Working Committee* (Markham & Buchanan, 2012). The chapter closes with the ethical position taken for this research project.

In Part II, Chapter 6 is a summary of the articles in this thesis. Additional analyses and data which were not fitted into the limited format of articles are included alongside some of the articles. The four Twitter studies are all summarised through description of context, methods and data, results and how each study has contributed to the thesis project. Also included is an analysis of overlaps of prominent users across the three data sets and further testing of the composite method. This test is based on a data collection made outside of the published articles. Here, #svp0l and the Australian counterpart #auspol are used with different parameters for evaluating how complete a data set including follow-on conversation is.

Part III includes discussion and conclusions as Chapter 7 and 8. The discussion takes off from the empirical results of the four Twitter studies. Then the research questions are discussed from the perspectives of three main themes: *the conversation and its boundaries, the filtering platform and its affordances, and sampling, bias and completeness*. Following this, the focus turns to what insights can be gained by using the methods of this thesis and applying them to some different contexts and settings. The essay part of the thesis ends with conclusions in Chapter 8, in which I return to the issue of completeness. Finally, limitations and ideas for future research are outlined.
2 Background

The thesis takes its starting point in a number of different traditions. For background and the research framework, research in the form of articles, conference proceedings and books from different disciplines have been utilised. Here, we find social media researchers such as José van Dijck, specific Twitter research publications by media scholars such as Axel Bruns, Anders Larsson and Hallvard Moe as well as researchers with a broader approach such as Richard Rogers.

van Dijck’s book *The culture of connectivity* (2013) takes an actor-network theory and macroeconomic view on social web platforms. The view taken is holistic in that the author believes that back-end analysis cannot be separated from front-end analysis. Naturally, it is the chapter about Twitter that is most relevant for this thesis. More generally, Mike Thelwall’s book about webometrics (2009) and *Digital Methods* (2013b) by Richard Rogers have been important publications for this project. For the research framework, books about how new media affect democracy and the society, such as Cass Sunstein’s *Republic 2.0* (2009) and David Weinberger’s *Too Big to Know* (2011), give a theoretical ground and provide context for the studies. Big data literature by Mayer-Schönberger and Cukier (2013) and Dormehl (2015) and a book about *Gatekeeping theory* written by Shoemaker and Vos (2009) have been other important publications for this thesis.

Some publications have been useful for understanding the Twitter API and for an overview of software. Matthew A. Russell’s book *Mining the Social Web* (2011) was used as a starting point for the project and *Analyzing Social Media Networks with NodeXL* (Hansen, Shneiderman & Smith, 2011) provided a solid description of networks, network analysis and the software NodeXL, which, however, was later discarded as a data collection tool.5

This chapter starts with a description of Twitter and how it is a part of a social media ecosystem. The following two sections discuss how Twitter can be studied

---

5 A good overview of data collection tools is provided by Gaffney and Puschmann (2013). The interested reader is also recommended to follow PhD student Wasim Ahmed’s blog at https://wasimahmed.org/about/.
generally and political conversations specifically. The chapter ends with an articulation of the specific focus of the thesis and some special considerations for this research project. The largest part of the chapter, section 2.3, builds on three literature searches in Web of Science and Scopus. From an extensive literature search carried out in several databases for Study IV it was clear that reasonable and scalable solutions to the problem of collecting conversational threads had not been previously published. The study presented and discussed the two possible solutions (Cogan et al., 2012; Zubiaga et al., 2015) referred to in the introduction of this thesis.

2.1 What is Twitter?

Sociologist Dhiraj Murthy (2012) has described Twitter as a microblog that is event related. His book *Twitter: Social Communication in the Twitter Age* (2013) puts the platform in a historical perspective and provides examples from Twitter usage that relate to journalism, disasters, activism and health. In this book, microblog is defined as:

an Internet-based service in which: (1) users have a public profile where they broadcast short public messages/updates whether they are directed to specific user(s) or not; (2) messages become publicly aggregated together across users; and (3) users can decide whose messages they wish to receive, but not necessarily who can receive their messages; this is in distinction from most social networks where following each other is bidirectional (mutual). (Murthy, 2013, p. 10)

This definition arguably positions Twitter as the prototype for a microblog. Other applications sharing some or all three of these characteristics are *Tumblr, identi.ca*, the peer to peer application *twister* and the Chinese platform *Weibo*. Apart from facilitating the function of writing short messages of a maximal length of 140 characters, the Twitter member can, for instance, follow individual accounts, mark messages as favourites, direct messages to others (@mentions) and forward messages (retweeting). A central feature is the hashtag, which can be used to tag the tweet topic-wise and by doing so involve it in a larger conversation (e.g. Bruns & Moe, 2013). Twitter is an “echo chamber of random chatter”, and a place of collective opinions and trends that quickly appear and disappear (van Dijck, 2013,
Twitter has also been described as a medium where the banal exists alongside discussion of current events (e.g. Murthy, 2013; Rogers, 2013a). Murthy (2013) has argued that if researchers ignore this banality they “risk applying an elite bias to understanding Twitter” (p. 149). Rogers (2013a) identified a continuous shift away from the banal during the evolution of Twitter, but also conceded that the trivial still thrives.

As Twitter has evolved, functionality has been added, and features changed. Trending topics were added in late 2008 together with the hashtag, and retweet functionality was added in 2009. When these, alongside @replies and @mentions, became central to the member, Twitter moved to be more of a utility for ordinary users and not only for the tech-savvy (van Dijck, 2013, p. 72-73). As Twitter is a commercial platform that needs to generate profit, it has evolved over time. These changes have both been applied to the user features and to the API rules. As a company for which advertising revenue is important, the more users the better, but a problem for the platform has been the stagnation of user growth (Goel, 2015). Changes of the user interface and the user features are strategies to attract new users. These include the favourite button, which was once a star, but was replaced by a heart in late 2015, indicating a semantic shift from bookmarking to liking.

Twitter has moved from initial ideals of serving as neutral, public utility (as electricity or water) to uplifting its popular users, such as celebrities, politicians and influential users. Neutrality is hindered by hyperconnected users with the ability to influence information flows, the savouring of influential users and algorithmic bias (van Dijck, 2013). A clear indication of this is changes made to the timeline, which tends to emphasise “important tweets” (Twitter, 2016c), undoubtedly a move towards the popular. However, this function can be disabled.

### 2.2 How can Twitter be studied?

Following van Dijck (2013), the perspective chosen not is to view Twitter as an intermediary, but rather as a mediator. The platform can be studied in multiple ways. In this thesis, two main approaches are used; hashtag oriented, in which one or several hashtags are tracked, and user oriented, in which a set of users are followed. Among the areas and approaches that are not within the scope of this thesis are:
• Business economics (how business models are made in social media),
• Political economy (how power relations are created in social media),
• Political science (how the political processes have the potential of being transformed through social media),
• Media and communication (Twitter as part of a larger media ecology),
• Pure user studies (followership relations or analysis of profiles only),
• Linguistics (the language of Twitter, sentiment analysis and opinion mining) and
• Computer science (refining the technology of Twitter and development of algorithms for processing and analysing Twitter data).

From the beginning, Twitter was very open with its data and free access was given. Later however, Twitter has a history of changing its API rules as well as restricting access (Burgess & Bruns, 2012). The earlier practice of allowing generous access to the API was called whitelisting, which was utilised by both scholarly and commercial third party actors (van Dijck, 2013). However, this practice was discontinued early in 2011, thereby forcing researchers to start from scratch. Previously, it was possible to query the Twitter API up to 20,000 times per hour (Barash & Golder, 2011), but this functionality has been withdrawn (Read Write Web, 2011). In 2010, Twitter started to charge external developers for using its data (van Dijck, 2013, p. 81), and by the time of the second data collection for this project the API v. 1.0 was deprecated and replaced by API v. 1.1, further limiting access to the API. Recently, Twitter has introduced its “data grants” project (Twitter, 2014a) through which the modest number of six research institutions were granted access to Twitter’s archive (Twitter, 2014b). With this in mind, researchers need to be innovative and develop methodology for approaching completeness regarding data associated with a conversation.

The Twitter API enables many aspects to be studied, but its restrictions have obvious implications for researchers. In a review by Zimmer and Proferes (2014), the use of larger data sets in research was found to decrease during 2012, possibly as a result of the new API restrictions. They also found that researchers until 2012 had focused on less than 1,000 users and on more than 100,000 tweets. 17 research areas were identified in their review, of which computer science, information science and communications were dominant. Another review by Williams, Terras
and Warwick (2013), found that most research between 2007 and 2011 had focused on messages or users, some of them combining both entities.

An important point in this thesis is that tweets should be analysed in their context. Hence, the conversation itself is the optimal sample, but there is no single API method for collecting these. One option in collect tweets is to use the GET search/tweets endpoint of the REST API (Twitter, 2015e) which offers a way to post a detailed query to the API up to 180 times per 15 minutes. The problem with this approach is that there is a restriction of 500 characters and that not all tweets are indexed or available. The search API can only be used to make repetitive searches with, for example, a small set of hashtags, keywords or usernames. With the restricted amount of calls to the API, it is not feasible to use the search API to search for potential replies to collected tweets.

A more suitable option for research purposes is to use the streaming API (Twitter, 2015f) to stream tweets in real-time. The streaming API endpoint must be used with at least one of the predicate parameters follow, track or locations. Follow collects tweets sent to or by a maximum of 5,000 users, including retweets, track collects tweets matching up to 400 keywords or hashtags and locations collects tweets from up to 25 locations, defined by a set of geographical coordinates. There is a streaming cap which applies to this by limiting the amount of tweets to 1% of the total volume of tweets at any given time (e.g. Gaffney & Puschmann, 2013; Morstatter et al., 2013). With the current state of the API, workarounds are needed to collect complete conversations.

As previously stated in this thesis, a problem with Twitter research is the exclusion of follow-on conversation. When only one streaming parameter is used, the data set is restricted to the tweets that include the keyword/hashtag, are posted by or to a set of users, or are tagged with matching geographical coordinates. A major ambition, developed through the current project, is to navigate toward complete data. This is important as tweets matching a hashtag do not seem to constitute a sufficient set for answering research questions. Collecting tweets that match a hashtag or are posted by a set of users effectively limits the possible research questions that can be posed which in turn results in research questions that are tailored for easily available data. The organisation of Twitter streams invites us to investigate conversations on the basis of initiatives to conversations, or at best, fragments of conversations. Given a method where more complete conversations are collected, theory-based questions can be posed. For example, in
communication networks where follow-on conversation is also captured, we can be more confident of identifying those who are influential than in a communication network created in the traditional way. Basically, a communication network includes the users as nodes and the messages as edges. If A retweets or replies to a tweet posted by B then an edge from A is drawn to B. A mention network is slightly different, as a tweet can contain more than one mention. If A mentions B and C, edges are drawn from A to B and from A to C.

2.3 How can political Twitter conversations be studied?

In this section, an overview of literature relevant for the thesis is outlined and discussed. Studies considered relevant have a descriptive, an exploratory or an empirical focus. Relevant literature was continuously sought and used during work with the different studies. This was particularly the case for the first study which aimed to review and identify linkages and differences between two different research approaches. However, in closing the project and in working with the summary essay the review was updated systematically as a methodological contribution to the emerging research area of political Twitter communication. The main purpose of this review is therefore to identify variations in methodological approaches and discussions. This provides a context within which the original aspects of the current work can be identified.

The literature review summarises how Twitter has been used as a data source and research within the realm of political communication. It starts with an overview of politics-related research of tweeting in which a sample of articles were coded with the aspects communication type studied, if any (e.g. followings, mentions, retweets, URLs), approach (e.g. hashtag, user, geographical), method (e.g. statistics, social network analysis, content analysis), data collection tools (e.g. direct use of API, software), discipline (based on first author) and how the authors label the user interactions (e.g. conversation, communication, discourse). The review continues with a thematic division of research where some examples are highlighted.

All articles were retrieved from the two sources Web of Science and Scopus by searching for research within social sciences. In Web of Science, a topic search was made using the word twitter. In Scopus, the word twitter was directed towards the fields title, abstract and keywords. Searches were made at three separate dates.
On February 8 2016, publications from 2015 were collected. Items published until the end of 2014 were collected on June 5 2015. On the 1\textsuperscript{st} of March 2016, additional material from 2016 was collected. The search results were filtered manually by removing all non-empirical papers and all papers not studying political Twitter usage or political Twitter communication. A summary of approach and method is given in Table 1.

Table 1. Cross-tabulation of data collection approach and data analysis methods. * interviews, observations, trend analysis, cluster analysis, machine learning, engagement analysis.

<table>
<thead>
<tr>
<th>Data analysis method</th>
<th>Hashtag (N = 57)</th>
<th>Keyword (N = 19)</th>
<th>User (N = 48)</th>
<th>Other (N = 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Statistics</strong></td>
<td>35</td>
<td>15</td>
<td>33</td>
<td>4</td>
</tr>
<tr>
<td><strong>Social network analysis</strong></td>
<td>27</td>
<td>8</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td><strong>Content/discourse analysis</strong></td>
<td>24</td>
<td>3</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td><strong>Sentiment/semantic analysis</strong></td>
<td>6</td>
<td>5</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td><strong>Other</strong>*</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

The most common data collecting approaches for studies of tweeting are hashtag or user-based. In nine studies, these two were combined in various ways. In one study, a sample from the streaming API was used, and one study used geographical coordinates for collecting data. Most of the studies were from the areas of media and communication (58) followed by political science (28). Notably, few examples from information science were found, which is somewhat surprising considering the many articles from this discipline identified by Zimmer and Proferes (2014).

Given attention to how data are collected in the current text, it is of particular interest to survey the various approaches in the material reviewed. However, far from all declared how data were collected, but direct access to the API was most common (41 cases). your TwapperKeeper (yTK) or its precedent TwapperKeeper was used in 17 studies, and NodeXL in six. These findings can be compared with the findings of Zimmer and Proferes (2014), who identified a heavy domination of API usage and few examples of software such as TwapperKeeper. Finally,
reflections on limitations regarding data collection were not common. In 31 cases, the authors were clear about this, but in 55 cases, the authors failed to reflect on such issues. These figures do not include the 22 papers where data collection was unclear. Overall, it is apparent that methodological issues regarding data collection within the area of political Twitter communication are often neglected.

Twitter studies are often distinctly marked by the choice of approach, falling into one of the following three categories: hashtag-based (e.g. D’heer & Verdegem, 2014; Larsson & Moe; 2014; Larsson & Moe, 2012; Maireder & Schlögl, 2014; Tinati et al., 2014; Xu et al., 2014), keyword-based (e.g. Himelboim, Smith & Shneidman, 2013; Tomasjan et al., 2011) or user-based (e.g. Ausserhofer & Maireder, 2013; Graham et al., 2014; Kruikemeier, 2014; Renz & Sullivan, 2013; Yoon & Park, 2014). Some studies utilised a combination of two categories. Jungherr and Jürgens (2014) identified a set of most followed users from whose tweets the ones matching a given hashtag were sampled. Dubois and Gaffney (2014) and Dyagilev and Yom-Tov (2014) turned this around and selected users of a given set of hashtags. Aragón et al. (2013) tracked both hashtags and selected users, and a similar approach was utilised by Bruns and Highfield (2013). Barberá et al. (2015) combined keywords and hashtags in their data collection.

Twitter data have been combined with other sources in some studies. Croeser and Highfield (2014) used tweets matching hashtags, field observations and interviews with Twitter users to study the Occupy Oakland movement. Kruikemeier (2014) followed the activity of politicians during a Dutch election campaign. From a quantitative content analysis which was compared to voting data, it was concluded that politicians might benefit from Twitter usage in a two-way communication style.

The study of conversational threads, pioneered by the project at hand, can be seen as quite unique, given this literature review. The only example of studying conversational threads found was provided by Kim and Yoo (2012). However, this study focused on to the extent to which a message was retweeted or replied-to, and the threads themselves were not analysed. Among the findings was that if a tweet included a URL or had a more positive tone it was more likely to be retweeted and less likely to be replied-to, and if a tweet was more personal it was more likely to generate replies than retweets.

A common approach to analyse Twitter communication is social network analysis (e.g. Ausserhofer & Maireder, 2013; Barberá et al., 2015; Conover et al., 2011;
D’heer & Verdegem, 2014; Dang-Xuan et al., 2013; Larsson & Moe, 2014; Larsson & Moe, 2012; Maireder & Schlögl, 2014; Yoon & Park, 2014; Xu et al., 2014). This approach allows followership, mentions and retweets to be studied as networks. In some cases, these are all merged into the same network (e.g. Choi, Sand & Park, 2014). Other ways of analysing Twitter activity and data include descriptive statistics and content analysis. Content analysis has been used both quantitatively (e.g. Kruikemeier, 2014; Meraz & Papacharissi, 2013; Small, 2011) and, less commonly, qualitatively (Himelboim, Hansen & Bowser, 2013).

Given this literature study, it would be fair to say that the concept of conversation lacks the theoretical discussion upon which Twitter research can stand. It is quite common for researchers to refer to conversations when they are actually studying partial conversations. Articles in which a set of hashtagged tweets constitute a conversation (i.e. the hashtag defines the conversation) have been quite common (e.g. Bruns, 2012; D’heer & Verdegem, 2014; Larsson & Moe, 2014; Pond, 2016). Tinati et al. (2014) viewed the retweets within the hashtagged space as conversation and Christensen (2013) used conversational ecology via retweets, while others have defined mentions as conversational (e.g. Barberá & Rivero, 2015; Bruns & Highfield, 2013; Holmberg et al., 2014; Renz & Sullivan, 2013; Wang, Wang & Zhu, 2013). In Meraz and Papacharissi (2013) the addressivity markers retweets, VIA and mentions, and hashtags constituted the conversation. Other concepts used are communication (e.g. Croeser & Highfield, 2014; Kruikemeier, 2014), communication space (e.g. Jungherr & Jürgens, 2014) and discourse (e.g. Drüeke & Zobl, 2016; Dyagilev & Yom-Tov, 2014; Trilling, 2015; Xu et al., 2014).

Some general points seem to apply to political Twitter communication. Apart from the aforementioned note on a lesser number of people accounting for a large share of the tweets, there is also an imbalance in initiating conversation versus receiving tweets (e.g. González-Bailón et al., 2014; Wang, Wang & Zhu, 2013). However, there is also an example of more reciprocal communication in the case of “high-end” users (Larsson & Moe, 2012). Another noteworthy point is that activity is event-related; increasing during or immediately after events such as Election Day (e.g. Bruns & Highfield, 2013; Larsson & Moe, 2012), televised political debates (e.g. Larsson & Moe, 2013), and protests (e.g. Jungherr & Jürgens, 2014). During protests, activity was more centred on spreading the news rather than on conversation, compared to activity before and afterwards. The use of a hashtag from the user perspective has been recognised as an ad-hoc public by Bruns and
Burgess (2012), in which conversations around a hashtag suddenly appear and disappear. One example is the analysis of the hashtag #aufschrei (German for “outcry”) by Maireder and Schlögl (2014). Their study showed how the usage of a hashtag can grow significantly in a short time space.

With the real-time orientation of the API it is not surprising that many studies have focused on scheduled events. There are quite a few examples of predicting election results (e.g. Franch, 2013; Soler, Cuartero & Roblizo, 2012; Tumasjan et al., 2011; Xu et al., 2014). Some of the studies have even found that counting references to political parties could be used to predict the election results (Tumasjan et al., 2011; Soler, Cuartero & Roblizo, 2012). This research have been criticised and results have been disproved in other studies (e.g. Metaxas, Mustafaray & Gayo-Avello, 2011; Jungherr, Jürgens & Schoen, 2012).

Other studies of elections have analysed patterns of activity in conjunction with elections by tracking hashtags (e.g. Bruns & Burgess, 2011a; D’heer & Verdegem, 2014; Larsson & Moe, 2012). In one of few examples of longitudinal studies of political Twitter usage, Larsson and Moe (2014) compared Twitter activity during two elections, separated by two years. Bruns and Highfield (2013) utilised a combined approach as they followed politicians and compared their activity with the overall activity of an election-related hashtag. A different method was used by Christensen (2013) who manually collected tweets posted by minor party presidential candidates. Twitter usage in conjunction with other political events has been studied as well. Jungherr and Jürgens (2014) selected users of a hashtag among the 80,000 most followed German tweeters during a protest. There are more examples of studies of Twitter usage during protests (e.g. Bastos, Mercea & Charpentier, 2015; Jackson & Welles, 2016; Tinati et al., 2014; Wang, Wang & Zhu, 2013) and similar events, such as debates (e.g. Mascaro & Goggins, 2015; Pond, 2016; Trilling, 2015).

Several studies have investigated politicians’ use of Twitter. Graham et al. (2014) focused on Twitter usage by political actors during an election campaign. Using quantitative content analysis, they coded the types functions and topics of tweets, as well as the actors in interactions. Himelboim, Hansen and Bowser (2013) similarly followed politicians during a campaign, and applied social network analysis and a grounded theory based content analysis of relationships and tweets. Political actors’ Twitter usage outside elections has also been studied in several papers (e.g. LaMarre & Suzuki-Lambrecht, 2013; Larsson & Ihlen, 2015; Renz &
Sullivan, 2013; Yoon & Park, 2014). There are also examples of political conversations not in conjunction with elections or other events. Small’s (2011) content analysis of tweets containing #cdnpoli (Canadian politics) is one example. Other examples include the user-based study of Austrian political conversations by Ausserhofer and Maireder (2013), and Larsson’s (2014) study of Norwegian and Swedish hashtags.

Polarisation and similar aspects have been previously studied within the topic of political Twitter communication. Using content analysis of replies, Yardi and boyd (2010) found evidence of Twitter users being more likely to communicate with like-minded even though they were exposed to different viewpoints. Combining hierarchical cluster analysis with textual analysis, Graham (2016) found that terms used by white extremists on Twitter were not different to those found in the mainstream political discourse. Colleoni, Rozza and Arvidsson (2014) used machine learning and social network analysis to categorise users into the two political camps Democrats and Republicans. Given a definition of homophily based on the number of outbound relations with users sharing the same political viewpoint divided by all outbound relations, they found that if Twitter is viewed from a social aspect, the studied space is more echo chamber-like. But if we take the perspective of Twitter as news medium then it is more like a public sphere. The authors stressed the importance of taking political culture and practices into account.

It seems as if Twitter users are more inclined to spread information confirming their viewpoints but that they are also willing to talk to non like-minded. In a study by Conover et al. (2011), polarisation was found in the retweet network but not in the mention network. Examples of retweet polarisation can also be seen in studies of Dyagilev and Yom-Tov (2014), where linked-to pages with a polarised language were more likely to be retweeted by supporters of the perspective from which the page was written, and in Barberá and Rivero’s study (2015), with the latter looking at transitive retweets, that is, retweets of retweets. Another related example is the study by Barberá et al. (2015), who analysed retweet patterns in twelve events and issues, both political and non-political. They found that echo chambers are more prominent in political topics, but also that while Twitter users seem to prefer to retweet tweets posted by like-minded, they are exposed to ideologically-wise diverse information. Aragón et al. (2013) used social network analysis to study replies and retweets related to political parties. The networks of retweets showed clear indications of balkanisation while networks of mentions did not to the same
extent. Cross-party communication mainly consisted of exchanges between similar parties. It was also concluded that most communication was one-way and that new and minor parties seem to be better connected when looking at intra-group connections.

Typically, a network with distinct clusters reveals echo chamber tendencies, but even if a network is not clustered it can still reveal that people are more inclined to communicate with like-minded. An example of a network without clearly separated clusters can be found in the study by Ausserhofer and Mairerder (2013), in which groups of professionals, activists, marketers and citizens were identified in a mention network including 882 active users.

Choi, Sand and Park (2014) used social network analysis on an aggregate level to analyse followership relations, retweets and mentions, i.e. all in a single network. A similar study was conducted by Himelboim, Smith and Shneiderman (2013) with a focus on echo chambers. In both studies, clustering tendencies could be seen in the networks. The idea of merging different network types to a single network is not likely to be a suitable approach, as it has been shown that followership, mention and retweet networks differ statistically (Overbey et al., 2013). Another good reason to separate the networks from each other is that a combined network will always be skewed towards followings, which seem to be characterised by polarisation (Kim & Park, 2012). It is reasonable that a followership network is denser as users are more likely to communicate with and retweet a subset of users they follow. In addition, replying to someone requires more effort than retweeting someone.

Research on the problem of identifying influential actors and opinion formers on Twitter focuses on activity and relationships. Ye and Wu (2010) found that retweet counts was a better estimator of influence than follower counts and reply counts. Anger and Kittl (2011) suggested using combined metrics that take into consideration both the interaction ratio and the retweet and mention ratio, whilst also considering the number of followers. A different approach was proposed by Rao and Nagpal (2011) who calculated influence based on how fast information from a given user in a specific area is spread. Related to this is the work by Bakshy et al. (2011), which makes use of a narrow definition of influence, i.e. an actor influences its follower if a URL posted by the actor is later posted by the follower. Dubois and Gaffney (2014) based their approach on relationships. Based on a set of users of given hashtags, a follower graph was created, to which six different
network metrics were applied. Up to 200 most recent messages from users in the
data set were collected and a subset of these was coded in a content analysis
focused on finding evidence of knowledge or expertise, based on the usage of
given keywords.

\section*{2.4 Specific focus of the thesis}

Despite the substantial amount of research devoted to this young social media
platform, there is a problematic lack of methodological development with regard to
the collection of complete conversations. Returning to the concepts of method
developer and method broker it becomes clear that the Twitter research community
has hitherto been unable or lacked the ambition to develop innovative data
collection methods. Unawareness of what data might be omitted in data collection
(e.g. follow-on conversation) is problematic and it seems to be generally
acceptable to Twitter researchers to collect data in a too convenient way. In cases
where the size of the follow-on conversation is large, the sentiments and opinions
expressed in the hashtagged tweets are likely to be drowned. It is not unlikely for a
hashtagged tweet to spark reactions of disagreement in the follow-on conversation.
In such cases, the hashtag-based approach thus favours the minority view. Moreover,
networks of relationships might be completely different from the
convenient hashtag, user or geography-based sample compared to the more
complete set including follow-on conversation. It is crucial to reflect on the value
of the knowledge and insights gained from a \textit{convenience sample} (i.e. easily
accessible data collected by specifying just one API parameter, see 3.4) when the
follow-on conversation is unknown.

Using hashtag or keyword-based data collection methods, tweets not including
either are excluded, regardless if they are replies to collected tweets or not.
Similarly, with a user-based data collection, tweets from one user to another user,
both outside the set of followed users, are not captured, even though they might be
part of the conversations involving a followed user. There are two levels of data
availability. The first includes the data that are made available through the API.
The second includes the data that are available through the software using the API.
To the two types of measurement errors identified by González-Bailón et al. (2014)
– the coverage and representativeness of the API and the communication networks
emanating from the sample – we can now add these two levels as aspects.
This thesis is not the first scholarly attempt to solve the problem of collecting complete conversations without firehose access. Methods for extracting conversational threads have been proposed by Cogan et al. (2012) and Zubiaga et al. (2015), but both of these have drawbacks. Cogan et al. (2012) attempted to reconstruct threads by querying the API for usernames detected in the collected data set and Zubiaga et al. (2015) scraped the pages of highly retweeted tweets to collect the threads. The former method is not feasible due to the restricted number of calls to the API. The latter method might work for collecting complete threads but scraping is probably against the Twitter developer rules. Also, when considering Kim and Yoo’s (2012) findings, which indicated that one type of tweet is highly retweeted and another type is more replied-to, it seems likely that a selection based on retweet counts would miss out on the most conversational threads.

Given the problems outlined here, the ambition is to make a contribution to the existing gap between the empirical/exploratory and the methodological/algorithmic research. Bruns and Moe’s (2013) claim that covering conversation beyond the hashtag is not an ambition to cover for hashtag-based studies together with Gaffney and Puschmann’s (2013) warning against tailoring studies after data availability, makes it clear that there is more to do in this specific research area. The solution to the problem of not collecting follow-on conversation presented here is the development of the composite method, which combines hashtag/keyword tracking and user following. As became obvious in the literature review, social science has been slow in developing research methods suitable for political topics in Twitter studies. The main issue here is that researchers are unaware of data they do not capture. Hence, there is a need for method development, which can be directed to the arguments of Thelwall and Wouters (2005).

2.5 Special considerations of the project

The project is based on the Swedish political Twittersphere, which is a case believed to be information rich enough for developing methods. Hence, political hashtags have been used, since their usage patterns have proven to involve a larger share of mentions than other hashtags and thus likely to be more conversational. The focus on followings, mentioning and retweeting in this project follows a
tradition developed in previous research. Although Twitter has more functions than these three, they are far more often studied than usage of other functions such as lists and favourites.

Polarisation is here studied with similar methods to those used by Conover et al. (2011), but including the three aspects followership, mentions and retweets (Study II), and as in Yardi and boyd (2010) by using content analysis (Study V). These two methods complement each other, first by analysing the broader patterns of Twitter use and relationships and then by identifying how the interactions are manifested. Longitudinality is studied through the use of hashtags as well as statistics on overall Twitter communication including retweets by other members exposed to the messages of the elite users (Study III).

Method development lies first and foremost in the data collection activities. Study II did not involve follow-on conversation but acknowledged that its absence might affect the structure of the mention networks. Study III implemented a data collection procedure that was able to capture follow-on conversation and Studies IV and V developed this procedure further. With the sets collected, an algorithm for calculating prominence based on activity (number of tweets posted), visibility (number of tweets received and number of repliers) and spreadability (number of tweets retweeted and number of retweeters) was applied on a weekly basis. The algorithm was based on Anger and Kittl (2011).

2.6 Summary

This chapter has summarised research on political conversations on Twitter, the political use of Twitter and the various methods used to collect and analyse data. Access to data is limited and restricted by both the technological aspects of the API and the rules that must be followed in order to use it. Researchers have often chosen to access the API in one particular way, by using hashtags, keywords, user IDs or usernames. As many articles do not discuss limitations related to software or API, this is problematic. The next chapter discusses how Twitter as a filtering mediator shapes both the research and the usage of the platform.
3 Research framework

So far in this thesis the neutrality of Twitter as a platform has been questioned and issues of sampling data from the platform have been briefly discussed. If Twitter is not neutral, it is a mediator rather than an intermediary. The aspects of the mediator focused on here are its affordances in terms of how people and software can interact with its user interface and API, and how data and content are filtered by the platform, its API and its users.

With informetrics dealing with quantitative aspects of information, it is not surprising that the theory section in the review by Bar-Ilan (2008) identified usage of statistical models, and laws of distribution and concentration. Sugimoto (2016) acknowledged that the field makes use of a multitude of theories, often importing these from other disciplines. When informetric techniques are applied, relevant theories for the studied context or domain need to be considered. With the methodological focus here, the starting point is the study of quantitative methods for a given context in relation to a specific platform. If measurements are to be as accurate as possible some aspects need to be observed. Such aspects include potential bias in the API, how the affordances of the platform affect the interactions and how the platform itself emphasises content and users. The research framework has an emphasis on Twitter as a platform and how it affects communication and the access to data from a social media researcher perspective.

This framework consists of five main parts. The first part discusses theoretical notions of conversation in relation to deliberative democracy. The second part introduces a theoretical focus on Twitter as a filter, both regarding information flows between users and how the API filters data from a researcher’s perspective. Filtering makes the study of conversations on Twitter complex and difficult. In the third part, attention is therefore turned to the affordances of Twitter, with a focus on technological aspects affecting how Twitter shapes conversations. Part four outlines an approach to issues of data complexity, which is grounded in notions of big data, software and sampling. The inclusion of this discussion is partly motivated by the identified lack of reflections on data, software and API in previous research on Twitter within the political realm. The main reason for discussing these issues is that researchers need to be aware of the limitations involved when studying activity and relationships on a commercial platform.
Twitter is fundamentally a social networking site and this is an aspect that Twitter researchers need to approach within a critical frame, as it is important to know how to interpret the different network metrics in relation to the different types of connections that are made through followings, mentions, replies and retweets. Therefore, relevant aspects of social network theory are included in part five. The interactions can be seen as a type of signalling behaviour, akin to citations. The various signals (following, retweeting, mentioning, replying and liking) are different types of recognition or acknowledgment of other users. Finally, it is fairly common that previous research on political dimensions defines interactions within a hashtagged space as conversation, even though such a space is very likely to exclude large parts of the conversations. In this research project, the concept of conversation is central. Part five includes a discussion of how the followership graph and the way Twitter presents conversations limits the visibility of conversations, and ends with aspects of the boundaries of the conversations.

3.1 What is a conversation?

While the thesis makes use of a technical definition of conversations similar to the definition that Twitter seems to prefer, it makes sense to discuss this definition in relation to other, more theoretical, definitions, as well as related concepts.

In educational theorist and cybernetician Gordon Pask’s *conversation theory* (1976), the interacting parts strive towards understanding a topic through a conversation. The theory has been reflected upon by numerous scholars. For example, Pangaro (1996) claimed that a conversation is required to agree on the meaning of something, and according to Ford (2005), participants in a conversation search for agreement of both the “why” and the “how” of a concept.

Information scientist Susan C. Herring (2010) referred to the very broad and early definition of conversation in the Online Etymology Dictionary (Conversation, 2016): “to have dealings with others”, but perhaps it is more common to see the conversation as face-to-face talk which happens in turns (e.g. Sacks, Schegloff & Jefferson, 1974). While it is acknowledged that there are similarities between face-to-face conversation and computer mediated conversation in that people take turns, the latter is different as it is shorter, its interactions are multiple and overlapping and turn-taking works differently (Herring, 1999). Conversation on Twitter is asynchronous, although probably with shorter time spans between a
tweet and a reply than other online forums. Conversations in online forums have a tree structure as does those on Twitter.

Magnani, Montesi and Rossi (2012) define conversation technically and in a similar way to Twitter’s own definition. Here, the trees of messages created by replies comprise a conversation, but note that their data collection approach is certain to leave out parts of the interactions as it is only based on keywords. The authors indicated that within the same tree, conversations that are topic-wise distinct from the main conversation can develop. They also viewed the hashtag as a connector of separate threads. Following this, the conversation can be seen to exist on different levels: the aggregated stream of hashtags, the tree consisting of replies and sub-conversation within each tree.

The conversation concept can also be viewed from the perspective of deliberative democracy. Fishkin (2011, p. 245) defined deliberation as “the weighing of reasons under good conditions in shared discussion about what should be done” where good conditions implies that participants have access to “reasonably good information” and are “willing to participate conscientiously” in a balanced discussion. Deliberative democracy is then defined as the “theory that attempts to combine deliberation by the people themselves with an equal consideration of the views that result” (Fishkin, 2011, p. 245). These definitions can be used for determining the quality of deliberation. According to Fishkin (2011), the five conditions information, substantive balance, diversity, conscientiousness and equal consideration apply for such judgments. The conditions highlight aspects such as the need for accurate information, the participants should reflect the wider public, that there exists a balance between arguments from different perspectives and that arguments should be considered on their merits.

For a computer mediated conversation to live up to the ideals of deliberative democracy, the mediator would have to fulfil several conditions. Graham (2008) presented an approach for detecting, describing, and assessing political talk in non-political online discussion forums, highlighting four normative conditions of the process of deliberation:

- *The process of achieving understanding* consists of the four components rational-critical discussion, reciprocity, reflexivity, and empathy. Here, emphasis lies on that claims should be reflected upon and justified as well as on participants should strive for maintaining coherence and continuity.
This in turn requires that participants listen and respond to other participants’ questions, arguments or opinions (reciprocity), reflect on other participants’ arguments or positions compared with their own (reflexivity), and endeavour to put themselves in another participant’s position (empathy).

- **Equality** includes *structural equality* (access, including skills of communication) and *discursive equality* (all participants are considered equal members). This means that one individual (or a group of individuals) cannot be privileged by the system, or should not dominate the conversation so others’ voices cannot be heard.

- **Freedom** includes *structural autonomy* and *discursive freedom*. The discursive space should be free from outside influence, coercion and control, and every participant must have the right to express opinions, raise issues, and criticise other participants.

- **Sincerity** implies that all information that is relevant to the discussion should be made known to other participants. This includes the intensions, motives, desires, needs, and interests of the participants. (Graham, 2008, p. 20-21)

If we view Twitter as a forum through the conditions identified by Fishkin and Graham, it is clear that deliberation in its strict meaning is unreachable. The points that are most troublesome relate to equality, which the platform fails in two ways. First, access to the platform requires Internet connection, second, Twitter’s affordances make it a forum for those who master them and third, Twitter emphasises the popular. Therefore, it seems reasonable to view the political conversations from a slightly different perspective. Arguing that deliberative theory has limited the study of political conversations, Eveland, Morey and Hutchens (2011) suggest putting equal emphasis on the participants’ and the political theorists’ perspectives. They defined *informal political conversation* as “interpersonal and small-group interactions about the broad topic of politics that take place outside of formal deliberation settings” (Eveland, Morey & Hutchens, 2011, p. 1083). This definition seems to encompass activity that is less bound to ideals, focusing on communication rather than on the extent to which the conversations live up to deliberative ideals.
Much of the sophisticated research on Internet-based conversations has been built on investigations of discussion forums. Arguably, Twitter differs from a discussion forum in a couple of ways. Tweets containing a political hashtag are supposed to be content-wise related to it. If we only look at the set of tweets that contain a political hashtag it seems reasonable to view the discussion as a political forum. However, not all Twitter users enter discussions with the aim of talking about political issues. Any user who follows a user posting a tweet with a particular hashtag can be exposed to that tweet. Hence, a hashtagged political discussion on Twitter is a public political discussion held in a non-political forum, open to possible listeners that do not visit Twitter primarily for discussing politics.

3.2 Twitter as a filter

In the following, I will describe a view of Twitter as a non-neutral filtering mediator that has affordances which influence how the platform is used and how it can be studied. van Dijck (2013, p. 28) views Twitter as a mediator as it “shapes the performance of social acts instead of merely facilitating them”. This view of Twitter as a non-neutral mediator is suitable for the data and methodology focus here, but it needs to be extended with affordance theory and gatekeeping theory.

Originally, the concept of affordance was associated with the physical realm. Gibson (1977) introduced the theory of affordances, which refers to how an object allows an organism to perform an action in relation to the object. The theory has been extended to include cognitive constraints in human computer interaction, as a match between human and technology (e.g. Fragoso, Rebs & Barth, 2012; Ware, 2012, p. 356). Fragoso, Rebs and Barth (2012) divided social media affordances into the categories representational, technical and socio-cultural. In the case of Twitter, a representational affordance is the 140 characters limit while technical affordances include the interaction between user and system, for instance, with possibilities to post pictures, sound and video. The openness of the system, the language of tweets and conversational features such as mentions, retweets and followings can be seen as socio-cultural affordances.

By re-grounding affordance theory in socio-cultural theory, Kaptelinin and Nardi (2012) proposed a mediated action perspective, in which the possibilities for human action are influenced by meditational means and environment. There are two main perspectives on interaction with Twitter that are relevant for the purposes
of this thesis. The first perspective concerns how the actions of Twitter users are mediated as human-computer interactions. Following, mentioning, retweeting, linking, information seeking and hashtag usage have been viewed as user affordances (e.g. Haustein et al., 2014; Holmberg et al., 2014; Lawrence et al., 2014; Zappavigna, 2011). Murthy et al. (2015) applied a different level of affordance by investigating how various devices affect tweeting behaviour.

The second perspective involves the interactions between the data collection software and the API. Following Kaptelinin and Nardi (2012), it can be argued that the API affords actions and access data in certain ways with a given set of restrictions (see Chapter 4 of this thesis). The actions are mediated, but not as human-computer interaction but rather as a series of programmed instructions in a form of computer-computer interaction.

With these two perspectives of affordances in mind, we now move to a central aspect of mediating for this thesis: filtering. With a focus on data, a view of Twitter as a filtering mediator is suitable. Filtering, as a theoretical notion, is grounded in gatekeeping theory, a model originally used for analysing food distribution chains and then adopted for the study of the selection of news items, explaining why some items reach the end-users and some do not (Shoemaker & Vos, 2009). Gatekeeping filters reduce a stream of information to smaller numbers of messages to a given audience (Shoemaker & Vos, 2009). In the case of Twitter, a user is exposed to the messages filtered and posted by the user accounts followed, and a consumer of the API can access the available data in the predefined ways programmed into the API.

There are major discrepancies in the way that gatekeeping theory can be applied to Internet-based and traditional media. Shoemaker and Vos (2009, p. 132) argue that the theory is applicable to new media if the researcher returns to the original application of studying social change. The traditional usage and view of the theory is challenged by Twitter, as its users receive information from a diverse range of sources and not only traditional news sources (Bastos, Raimundo & Travitzki, 2013). Gatekeeping theory has recently been extended to encompass networked gatekeeping, which makes it more suitable for investigations of social networking sites (e.g. DeIuliis, 2015).

The perspective of a filtering mediator is fruitful for two relevant purposes. First, it helps to facilitate an approach that focuses on the affordances of Twitter that shape the actions of its users and their interactions. Users themselves are filters too, filtering forward information to their followers as opposed to filtering out
information (Weinberger, 2011, p. 10-11). Information is filtered, edited and forwarded not only by professionals, but also by ordinary users (Klinger & Svensson, 2015). While users can be seen as gatekeepers, platforms can also act as gatekeepers (e.g. Bozdag, 2013; Helberger et al., 2015). Following the non-neutrality of Twitter as argued by van Dijck (2013), the platform itself filters information and, by doing so, it presents information personalised for individuals, resulting not only in an algorithmic but also in a human bias (Bozdag, 2013).

Second, from a data collection perspective one could conceive of the Twitter API as a filter as it determines what data are available and how. Weinberger’s notion of forward filtering is applicable here too. While in theory all data produced on Twitter are available through its API, some data are filtered forward. A reservation must be made here though, as some content might be filtered out at an early stage. Due to the real-time orientation discussed in Chapter 1, complete conversations are not filtered forward but are, at least to some extent, accessible through combining different parameters of the streaming API. By viewing Twitter from this perspective, I build on Twitter as a mediator while focusing on the constraints that exist between data and user, and between data and researcher. The Twitter API then, seen as a mediator, affords a researcher to interact with a set of endpoints through a predefined set of actions. The data returned are structured in a predefined way.

Two relevant concepts that are useful for understanding Twitter as a filter are **programmability** and **popularity**. Programmability can in turn be divided into technological and human programmability and is defined as:

> [T]he ability of a social media platform to trigger and steer users’ creative or communicative contributions, while users, through their interaction with these coded environments, may in turn influence the flow of communication and information activated by such a platform. (van Dijck & Poell, 2013, p. 5)

An example of technological programmability is the switch from star to heart; from favourite marking/bookmarking to liking. Programmability is affected by the users in two ways; by the content they contribute, amplify through retweets and likes, and by disobeying coded instructions or protocols (van Dijck & Poell, 2013). Examples of such disobedience are the circumvention of the 140 characters limit by replying to one’s own tweets, and by introducing the hash and at symbols and
retweets. Technical and human programmability shape the information flow mutually (van Dijck & Poell, 2013).

*Popularity* is based on both algorithmic and socioeconomic aspects (van Dijck & Poell, 2013). What might seem to be a neutral platform is instead a platform where popular users are more visible. By emphasising certain users and content, this shifts attention away from Twitter as decentralised and democratic as users and their contributions are treated non-symmetrically (van Dijck & Poell, 2013). If certain users are emphasised, or filtered forward, these users are more visible than others and are more likely to gain followers, and in the process be more visible. This Matthew effect, those who have much will receive more (Merton, 1968), seems to apply to Twitter as well. Other examples of popularity-related features and aspects are trending topics, endorsing by retweeting, and promoted tweets. Furthermore, when searching Twitter using keywords or hashtags, the default resulting list shows top tweeters. However, Twitter provides the option to see tweets from all members, but it should be mentioned that Twitter search is also algorithmically adapted to provide the content most relevant to the searcher (Twitter, 2016h). A researcher utilising the search API is likely to collect a data set which is not neutral, but rather built by results more relevant to the account used for collecting data. All these considerations highlight the need of a perspective or framework for studying Twitter usage, in which Twitter itself plays a role as a filtering mediator.

### 3.3 Affordances of Twitter

#### 3.3.1 Technology

The technology of Twitter can be understood through five significant concepts (van Dijck, 2013). The *Data* concept includes information that users produce themselves, for example tweets and profile descriptions. *Metadata* is typically produced by the platform, sometimes with the consent of the user (e.g. allowing location data). *Algorithms* are used for processing data. It is in essence a series of programming instructions to execute a given task, for example recommending users to follow. *Protocols* determine the formats of messages and how messages can be exchanged. *Defaults* are the standard settings for the users of the platform.

On Twitter, metadata includes timestamps, for example, the recipient of a tweet, URLs and hashtags. According to Mayer-Schönberger and Cukier (2013, p. 93)
there are 33 metadata fields accompanying a tweet, but the accurate figure at the time of writing is 34 (Twitter, 2016d). Changes are frequently made as to what metadata are tied to which content. Regardless of such changes, an important point is that metadata accompanying a tweet are datafied. Datafication is the process of transforming any information into quantifiable formats (Mayer-Schönberger & Cukier, 2013, p. 15). With metadata such as timestamp, geo-location and language, we can derive more than just the content of the tweet, for instance, the mood patterns of people in relation to geography and day of the week (Mayer-Schönberger & Cukier, 2013, p. 93).

Data and metadata are central to the studies performed in this thesis, but they cannot be viewed separately, as both of them are affected by algorithms, protocols and, to some extent, defaults. Dormehl (2015, p. 249) has suggested that algorithms are constructed with a hidden bias. Actually, we do not know much about the algorithms as they are typically black-boxed, i.e. the nuts and bolts of what is being done lack transparency. However, it is possible to reverse engineer the algorithms through testing of different aspects of the platform from various perspectives, thereby figuring out how they work. Twitter might at first sight seem neutral, but research has found clear indications that the streaming and search APIs yield different results (e.g. González-Bailón et al., 2014). An important reason for this is that the search API is optimised to power the search engine (e.g. Gayo-Avello, 2013; Gaffney & Puschmann, 2013). We should probably assume that there is some kind of bias in the streaming API as well.

Interestingly enough, protocol rules can sometimes be circumvented by users. In Twitter’s case this has been done by adding # and @ to the messages, a functionality which is now embedded into the platform (e.g. Bruns, 2012). The API is an interface steered by protocols, ruling what data are accessible and how to access it. It is described by van Dijck (2013, p. 29) as “a set of codes that specifies protocolized relations between data, software, and hardware”. This protocol specifies, among other things, how data can be accessed with regards to query and data structures. Finally, defaults need to be considered as well. Defaults might on the one hand make actions such as registration easier, but one size does not fit all and decisions could very well be made by the platform owner for self-interest purposes (e.g. Sunstein, 2013). An example relevant to this thesis is searching using a hashtag. The default result list is at the time of writing the “Top” listing which seems to order tweets by some aspects of (algorithmic) popularity. This ordering of tweets has an impact on what is more likely to spark conversations.
3.3.2 User affordances

Apart from the technological aspects of the platform there are also usage affordances to consider. In the input field where users create tweets, Twitter has asked different questions over the years since its start-up. At one time, the question concerned what the user was doing, similar do the Facebook status, but now the question is “What’s happening?” This indicates a shift towards a focus on Twitter as a news outlet on the one hand but also a shift towards the more general, as this question could also include what the user is doing. Another possible implication is that the new question of what is happening invites less reflection and more directness. This type of change impacts on how users of the platform communicate. The 140 character limit forces short and distinct messages or the user can circumvent the limitation by responding to his/her own tweets. One of the most important characteristics of Twitter is its information model for conversational tweets. Any reply from user B to user C is only visible to A if A follows both B and C (Twitter, 2016g), illustrated by Figure 1a. This is circumvented by some Twitter users by initiating a reply with a dot or other character (Quora, 2016), which makes the tweet visible to the followers of the author of the tweet (Figure 1b). Also, a retweeted reply will show up in A’s timeline if A follows the retweeter. At the time of data collection for Studies IV and V it seemed as if when searching by hashtag, a tweet from B to C was visible to A if A followed either B or C, directly or transitively (Figure 1a-f). During spring 2016, it appeared as this restriction had been relaxed. Seemingly, any reply including the search term(s) is visible to the searcher. However, the problem of follow-on conversation applies here too. Only the tweets in the thread that match the search criteria are visible. Twitter continuously tinkers with its interface and functions. As of writing June 2016, the most recent updates are outlined in a blog entry (Twitter, 2016a). The 140 character limit will change slightly during 2016, with replied-to users at the start of the tweet and media attachments (URLs, images, videos) at the end of the tweet not counting. Moreover, the dot notation will not be needed in a non-reply starting with a username; such a tweet will be visible to all users anyway. However, if a reply is to be seen by all followers it needs to be retweeted by the author. Retweeting an own tweet is also a new function.

Even though Twitter seems to shift towards the conversational, the way conversations are presented does limit the set of participants to tightly knit groups of users, followership-wise, implying that followership matters for the study of
conversations. An analysis of threaded conversations combined with a followership graph of the users involved should shed some more light on how threads evolve. With conversations not likely to spread outside the tightly knit group of participants, retweeted messages have every chance of spreading. This is another good reason to separate these two interaction types from each other in the analysis.

Many researchers seem to accept the hashtagged tweets as a conversation, but arguably there is more to the conversation than that. If we view a Twitter conversation in a similar way to a discussion forum, a tree structure of a tweet and its follow-on conversation emerges. This tree structure includes both hashtagged tweets and tweets without hashtags. Granted, for some hashtags there is little or no follow-on conversation, but this has not, as yet, been explored by researchers. There has been a discussion of the extent to which conversation happens on Twitter and if tweets can be seen as conversational or informational (van Dijck, 2013). There are also contrasting views on how conversational mentions really are.
While Honeycutt and Herring (2009) found that 91% of the mentions were directed, Bruns and Highfield (2013) found that many mentions were not conversational per se but rather used to talk about people.

3.4 Theoretical aspects on sampling from Twitter

3.4.1 Data complexity issues in relation to sampling

Given the development of big data, Mayer-Schönberger and Cukier (2013) reached a number of dramatic conclusions regarding the future of quantitative research, for instance that with access to very large data sets, sampling strategies is less important, if needed at all. The argument is that we are currently shifting our grasp from “some to all”, implying the future standard of $N = \text{all}$. It is argued that the methodology of sampling is becoming outdated:

> Reaching for a random sample in the age of big data is like clutching at the horse whip in the era of the motor car. We can still use sampling in certain contexts, but it need not – and will not – be the predominant way we analyse large datasets. Increasingly, we will aim to go for it all. (Mayer-Schönberger & Cukier, 2013, p. 31)

As early as 1951, in one of the most important works in epistemology, Willard Quine argued against such naïve assumptions. Notions about powerful empirical mechanisms appear problematic. Theory is under-determined by data or evidence. Thus choice, and the processing and testing of data are all underpinned by theoretical assumptions. In the context of Twitter research, similar arguments have been put forward by González-Bailón et al. (2014) and Gaffney and Puschmann (2013), and in the context of big data by boyd and Crawford (2012) and González-Bailón, the latter stressing that “the computational challenges created by Big Data means that sampling is often a necessity, and systematic attention needs to be paid to its impact on analyses and findings” (2013, p. 155). Importantly, without firehose access we do not have complete Twitter data. With hashtag, keyword, user or geo-location based methods we do not even have access to complete conversations.

Indeed, Mayer-Schönberger and Cukier (2013) themselves emphasise that big data builds on theory, which is also echoed by Dormehl (2015, p. 239) who points out
that data mining is founded on theory and adds that algorithms “often reflect the biases of their creators”. Large data sets do not necessarily outperform traditional sampling. Big data entails a series of complex problems requiring us to be inventive in approaching (not reaching) completeness. Biases are influential all the way through the process from data collection to analysis and interpretation. For the study of conversations, it is necessary to filter the stream of big data into well defined units of analysis.

3.4.2 Software as black boxes
An important aspect of Twitter as a filter is that interactions with its data are only made possible through its API, or through software interacting with the API. Understanding the API and, at times, its rather technical documentation is a challenge for any non-technical researcher. Without the understanding of the API or the software these are merely black boxes to the researcher. In science and technology studies, black boxes are facts or artefacts that are taken for granted (e.g. Sismondo, 2004). Arguably, a piece of software is an artefact that researchers often take for granted, seen as a sophisticated and objective representative of pure mathematics and algorithms. Given the domination of disciplines with a non-technical character in research on the political Twitter, there is a heightened risk that the output is routinely accepted as correct or complete. One example showing how this is problematic can be found in the study by Choi, Sang and Park (2014), who collected data using NodeXL and combined mentions, retweets and followings in the same network, resulting in a huge bias to followership relations (mentions: 3,093; retweets: 62,495; followership relations: 826,446). The authors did concede in the conclusions that “it should be noted that three types of relationships [...] were combined to address the proposed research questions. Given that some studies have reported differences between mention and retweet networks, the results of this study need to be verified by focusing on each relationship separately” (Choi, Sang & Park, 2014, p. 597).

Another issue relates to the data collected, and more importantly, to the data not collected. In Table II, the authors list nine identified opinion leaders, and if one adds the followership counts together, the total is 891,152 (Choi, Sang & Park 2014, p. 591), more than all followership relations in the combined network. How many followership relations per user were captured by NodeXL, considering that a vast majority of the user accounts in the study were not represented in the table? Barash and Golder (2011) point out that NodeXL is not an optimal tool for data
collection from Twitter. The study by Choi, Sang and Park acts as an example of how a black-boxed application makes results difficult to interpret. Yet, this is not the only example of studies that have made use of such software without critically reflecting on data collection (e.g. Himelboim, Hansen & Bowser, 2013; Himelboim, Smith & Shneiderman, 2013)\(^6\).

Social scientists have been sampling for a long time, but traditional sampling and collecting data from the Twitter API are two very different things. First, if there is such a thing as a random sample (e.g. boyd & Crawford, 2012; González-Bailón, 2013), the insights from such a sample are limited to usage statistics and the like. Second, if we use data collection techniques from one endpoint of the REST API we can, for example, collect lists of friends and followers and profile data if we know which users we want to collect such data from. By specifying one parameter when utilising the streaming API, we can also collect tweets matching a set of keywords or hashtags, tweets that are sent from or to a set of users, or tweets that are posted from a location within a given radius from a given geographical point. What we cannot do with such a set is to analyse complete conversations or map the relationships between all participants within the conversations, due to replies not matching the search criteria being omitted. A sample collected by specifying one parameter is what I refer to as a convenience sample. Although it is straightforward to collect tweets matching a keyword or hashtag, by following a set of users or by specifying a set of coordinates, such a set is not complete.

### 3.4.3 Convenience sampling

One of the most important arguments in this thesis is that research should not be tailored after the easily available data provided through the Twitter API. Research questions should guide data collection, and it is especially important when the aim is to analyse a conversation. Given the research interest and the relatively easy access to data it is surprising that there has been a (next to) complete lack of study of conversation so far. Bruns was among the first to reflect on the follow-on conversation (e.g. Bruns, 2012; Bruns & Burgess, 2012; Bruns & Stieglitz, 2013b). Since then, much research has been done on hashtag-based samples while

---

\(^6\) Note that Hansen, Smith and Shneiderman are authors of the book *Analyzing Social Media Networks with NodeXL*. 

52
acknowledging and accepting its limitations (e.g. D’heer & Verdegem, 2014; Highfield, Harrington & Bruns, 2013; Moe & Larsson, 2012), but far from all researchers have reflected on such issues. In some cases, the authors have acknowledged the drawback but argued that it was not significant given the research focus (e.g. Croeser & Highfield, 2014). There are examples of researchers pointing out the benefits of a hashtag-based approach as it “remains valid and important [...] especially where research focuses on identifying broad patterns in Twitter activity from a large data set” although it is also conceded that “no data set captured by using the Twitter API is guaranteed to be entirely comprehensive” (Highfield, Harrington & Bruns, 2013, p. 322). Another example is from Overbey et al. (2013, p. 1354), who claimed that the resulting network following a hashtag-based approach well approximates the size and has “few false inclusions”. Moe and Larsson (2012, p. 744) concluded that “even though the use of a specific hashtag will entail biases, the resulting data set does, compared to other instances of off- and online mediated communication, provide a basis for a comprehensive analysis”. While many researchers accept and reflect on the lack of follow-on conversation, little is known about what data are missing when there is no follow-on conversation.

3.5 Dimensions of Twitter activity and relationships

3.5.1 Social network theory

An understanding of social networks is crucial for any analysis of a networked space, such as Twitter. Hence, the next building block in the theoretical framework is network theory, which is utilised as a means for mapping and analysing the Twitter conversations. This section outlines core concepts and issues concerning social networks and what can be learned from them. Social network analysis investigates how social units are interrelated (e.g. Wasserman & Faust, 1994). Similarly, communication network analysis explores how individuals are linked together by their communication interchanges (Rogers, 2003). Theories behind network analyses imply that units are influenced by each other (Wasserman & Faust, 1994). In social networks, nodes are connected with arcs if the graph is directed or edges if the graph is undirected. Otte and Rousseau (2002), who proposed social network analysis as a strategy that can be applied within information science, give us a few applicable definitions:
• Density: An indicator of connectedness of a network, given as the number of lines in a graph divided by the maximum number of lines.

• Degree centrality: An indicator of how central a node is, given as the number of connections the node has with other nodes.

• Betweenness (centrality): The number of shortest paths passing through a node. Nodes with a high betweenness are the bridging nodes of the network.

Granovetter (1973) introduced the concepts of strong and weak ties and suggested that the latter type is important for diffusion across a network. Such ties can be seen as bridge links between highly connected parts of the network (Rogers, 2003, p. 306) and are encouraged by social networking sites (van Dijck, 2013, p. 7). This is particularly true for Twitter according to Tufekci (2014), as it is based on a unidirectional followership model. Therefore, its users are more likely to create bridges between sub-networks. Diffusion also requires that there are links within the network that are at least partly heterophilous (Rogers, 2003, p. 306). In the setting of a company, Burt (2004) found indications that people with bridging network positions are more likely to be innovative and to gain support for ideas. In the Twitter setting studied here, this would entail that people in bridging positions (i.e. high betweenness) are those who are able to spread their ideas to different clusters in the network.

Centola (2015) showed that social diffusion requires a combination of moderate homophily and moderate consolidation. High consolidation and high homophily creates islands which might be connected with small bridges. Low consolidation and low homophily create a more coherent network, in which the parts are connected with wide bridges. Centola concluded that “the social groups that emerge when structural consolidation actively constrains the process of network formation are in fact necessary in order to produce network topologies that can support the dynamics of social integration” (2015, p. 1315). The network structure is important to consider in diffusion studies. Centola’s findings indicate that both a coherent Twitter network and a network with distinct and separated clusters would constrain the information flow.

As we already have seen, social network analysis has been a popular tool for analysing aspects of Twitter communication. In Twitter networks, the graphs are directed if the purpose is to analyse who follows whom, who is mentioning whom,
and who is retweeting whom. The flow of information can thus be studied if the graph is directed. An undirected graph would be relevant if the purpose is to show mutual connections, or mutual messaging. Social network analysis can give information about the density of the graph and its sub-graphs. It can also reveal which actors are in bridging positions, and can thus be regarded as gatekeepers, as well as which actors are more likely to get a message spread throughout the network. This leads us to the concept of spreadability, which refers to the potential to share content with other people, both from a cultural and a technical perspective (Jenkins, Ford & Green, 2013). The users with a higher spreadability are those who have received the most retweets and by the largest number of users.

In the study by Dubois and Gaffney (2014) it was found that certain network metrics were more defining for some groups of users than for others. Centrality metrics were indicators of elite users such as politicians and mass media actors, but knowledge and expertise, detected through content analysis, and interaction within the network identified other users such as political commentators and bloggers. Xu et al. (2014) also used network analysis to identify opinion leaders. They found correlations between high betweenness values in a follower network and retweet counts. Local users (in proximity to a given event) were also more likely to be retweeted. Critique towards influence metrics based on retweets was put forward by Tufekci (2014), who noted that a retweet can take any meaning between negative and positive.

The concept of opinion leaders is relevant to studies of social media platforms, where such an actor might influence the opinions of other participants in the conversations. The users on Twitter act as filters by filtering forward and recommending information to its followers. As anyone can produce, this implies that anyone can be an opinion leader, but identifying opinion leaders is not possible without questioning their followers and others exposed to their messages about how opinions have been influenced. As such studies are outside of the scope of this thesis the concept used here is *elite users* as by Wu et al. (2011). The criteria for elite user in this thesis differ from Wu et al. (2011), who included media, celebrities, organisations and bloggers in their definition. Here, the label is used for users that are prominent in different aspects of Twitter usage and relationships. As anyone can be a gatekeeper, as claimed by Klinger and Svensson (2015), anyone can be an elite user on this platform.
The point of departure is that elite users are more active in their communication, both regarding singletons and mentions (activity), and more often mentioned (visibility) and retweeted (spreadability). Their prominence might be a result of early adaption to the platform and frequent posting about the topic of the conversations – they have a certain self-possession, which may well be inherited from the offline world. This thesis seeks to identify elite users from the three metrics activity, visibility and spreadability.

As celebrities tend to gain high actor scores with very little activity, a combination of metrics from different network types is preferable. Communication network analysis shares some similarities with social network analysis. The difference is that communication network analysis is focused on communication flow and the units of analysis are the interpersonal communication relationships (Rogers, 2003). In the case of Twitter, social network analysis is used as the overarching concept, also including followership. Communication network analysis is used here to investigate communication as it emerges in retweeting and mentioning or replying.

3.5.2 Implications of the followership graph and visibility of conversations

In order to deepen our understanding of Twitter as a filter, we need to consider the implications of the followership graphs and how Twitter makes replies visible. It is frequently stated that a prerequisite for a well-functioning democracy is that citizens are exposed to multiple, not single, political viewpoints (e.g. Pariser, 2012; Sunstein, 2009). With platforms making use of recommender systems, their users are more likely to be divided into homophilous groups. Homophily and heterophily indicate the degree to which individuals in communication networks are similar or dissimilar to each other according to a given attribute (e.g. Rogers, 2003).

While recommender systems facilitate filtering of information they are also unlikely to lead to “serendipitous discovery” by any user (Dormehl, 2015, p. 145). Pariser (2012) introduced the concept filter bubbles which are created for users by algorithms beyond their control. The filter bubble is created when algorithms filter data for a specific user based on previous actions and preferences. When people do the same more or less consciously, the result is an echo chamber (Sunstein, 2009). Echo chambers are discussion forums in which mainly like-minded people communicate, or one perspective dominates the contents or the opinions expressed within the forum. Filter bubbles and echo chambers are thus likely to inhibit exposure to a variety of perspectives of the world, and to reinforce opinions and
viewpoints. Such environments are likely to create polarisation, making citizens less likely to agree or compromise (Weinberger, 2011, p. 82). Polarisation is a consequence of echo chambers, pushing people towards extremes by reinforcing and corroborating their opinions and viewpoints (Sunstein, 2009).

However, it can be argued that polarisation may have some benefits. According to Sunstein (2009, p. 76), society could benefit from polarisation as “when many different groups are deliberating with one another, society will hear a far wider range of views”. Weinberger (2011, p. 85) highlights that homogeneity could be a requirement for knowledge construction:

> Knowledge has always needed communities to flourish. Communities need walls so that they can let in the right amount of diversity, even if too frequently they err on the side of homogeneity.

Weinberger (2011) also argues that even if information is exchanged within echo chambers, the transparency of the web might expose its users to various different ideas before they get to the echo chamber. Transparency can also function in a different way. If a conversational thread on Twitter is developed within a tightly knit group of users, comprising only like-minded people, the conversation is findable through the search function, or by a participant retweeting some of the content to followers with different viewpoints. The conversation is open, but not filtered forward to anyone.

Let us now return to the conclusions drawn by Centola (2015). It seems that an effective information Twitter network needs its echo chambers, but also fairly wide bridges between them. This is supported by Rogers (2003) who claims that communication is more effective when homophily is high. When networks are heterophilous on the other hand, “followers generally seek opinion leaders of higher socioeconomic status, with more formal education, greater mass media exposure, more cosmopolitanism, greater contact with change agents, and more innovativeness” (Rogers, 2003, p. 362). If Rogers is correct, and this applies to heterophilous Twitter networks, elite users such as politicians and journalists should have a more diverse follower base than non-elite users. The non-elite users would be more inclined to follow like-minded peers but when it comes to elite
users, they would be likely to follow like-minded as well as non like-minded. The data from Study II gives some indications of such tendencies, but more research on this is needed.

3.5.3 Boundaries of conversations

An important starting point regarding boundaries is to define what belongs to the conversation and what does not. Hashtagged tweets only constitute a subset of a conversation, and data sets including only such tweets can be seen as biased. Given that problems associated with sampling unresolved, it is necessary to deal with different ways of collecting data related to Twitter activity. By combining two different approaches in collecting tweets, the issue of incompleteness can be solved to some degree. In this thesis, the pursuit of collecting complete conversations started with mapping the usage of one hashtag from the aspect of polarisation. The second step was to follow the activities of an elite group over three four week periods during one year. The third step, and the one that applies most here, was to deal with the specific task of approaching completeness in the context of Twitter conversations, by combining the approaches in steps one and two.

Notions of total data sets imply that firm boundaries exist between repositories of data. One example, utilised by Mayer-Schönberger and Cukier (2013) is the strategy taken by Steve Jobs in his fight against cancer. Rather than allowing his doctors access to a small sample of his DNA, Jobs paid a six-figure sum to sequence his entire DNA. However, the doctors did not even attempt to map all of the human entity that was Steve Jobs. The example seems to build on an assumption that humans can be reduced to our firmly inscribed DNA sequence. However, while our DNA supplies a core for our biological manifestation, there is so much more to being human. In other words, although the total DNA sequence of Steve Jobs was mapped, this was still merely a sample of a larger system.

---

This was tested on the data from Study II where the users were grouped into one of the categories Left, Centre-right and Right-wing, if their profiles clearly stated political stance. Among the users, 189 politicians and 67 mass media actors were identified. The remaining 660 users were labelled as others. Politicians were followed by on average 2.1 different political categories (SD = 0.74), mass media actors by 2.5 (SD = 0.66) and other users by 1.7 (SD = 0.96). All differences in means (t-test, independent groups and Mann Whitney U) are significant at p = 0.01.
At issue here are really disciplinary projections on what constitutes complete data. From the point of view of the DNA-specialist, Steve Jobs was completely mapped. However, another specialist can approach the same subject with different tools and focus. There are essentially endless possible perspectives on the human body, each leading to separate instruments of collecting data. Specialists are often in competition with each other, arguing that their sets of data are particularly important to solving specific problems. In such cases, it might be possible to claim that one approach is more effective than others. If so, we are once again stuck with sampling. The DNA of Steve Jobs was a sample and perhaps not the best one as, obviously, it could not be instrumental in saving his life. Arguably, the answer to the riddle of his specific tumour lay elsewhere.

The example of DNA is otherwise an illustrative case of a data repository relatively fixed in time, i.e. Steve Jobs had the same DNA sequence his whole life. However, in most cases of real-time streams of heterogeneous data, this kind of stability is absent. Indeed, synchronous mapping of real-time data with high velocity is often sought to create day by day predictions. This is another fundamental problem for the notion of N = all. Data sets have histories and while we can aspire to track and capture as much of this as possible, we cannot access N = all.

The argument here is similar to that developed by boyd and Crawford (2012) when they suggested that a major challenge for researchers approaching data sets is to understand not only the limits of the material but also the limits of the questions that can be asked and the interpretations that can be made. The argument is that it is the analyst who constructs these boundaries by introducing a form of disciplinary bias, thereby, at times, creating the illusion of completeness. In Studies IV and V, this has been pursued in connection with the analytical entity of conversation, in the context of a set of Twitter hashtags. In the case of the conversation regarding the December Agreement, which was the case chosen for these two articles, the topic was defined by a series of hashtags. The approach taken required a stable use of hashtags, and that the participants did not substitute hashtags for other ones as the conversation evolved. This is a form of selection bias (He & Rothschild, 2015). Finally, there is another aspect of boundaries that is relevant here. Following van Dijck (2013) it is fruitful to see conversations on platforms such as Twitter taking place within a larger ecosystem of social networking sites. Positioning itself between other social media platforms has historically been important to Twitter:
Twitter steered its technological design to favour ubiquitous integrated use of its basic architectural elements; in order to interlink as many social networks as possible, the microblogging service gradually adapted its hardware and tweaked its software to fit other services’ standards. (van Dijck, 2013, p. 70)

The conversation might not just happen on Twitter, but traverse over different social media, websites, and offline. How to track a conversation that cuts across different platforms is a challenge that is not dealt with in this thesis.

3.6 Summary

The framework binds together five main theoretical aspects designed to create a sophisticated vantage point for the purpose of this study. The framework is not designed as an all-encompassing perspective on the data streams investigated. By defining the setting as interest-based, elite-centred, and as a social network on a commercial platform, the most relevant aspects are those which are related to the platform itself (i.e. filtering and affordances), social networks and elite users in relation to these. Viewing the platform as a mediator is also relevant, but more so if the social aspects are focused on. The choice to view the API and the users of the platform as filters are fruitful for the methodological focus in this thesis. In the next chapter, the methodological constraints that follow from the affordances of Twitter, both to the user and the client interacting with the API, are discussed.
4 Method

This chapter focuses on issues related to collection and analysis of data. It is not a conventional method chapter; rather, it discusses meta-methodological aspects as well as specific methodological considerations for the studies. The project makes use of three main types of data. The first one is relationship data, which is manifested in the three sub-types followership, mentions (including replies) and retweets. The second one is content data, which consist of the tweets that the actors produce. The third one is profile data, including profile descriptions written by the actors themselves, alongside number of followers and friends.

From the literature review it is clear that there is a lack of studies of conversations on Twitter, and in the research framework it is argued that for many researchers both the Twitter API and software designed to work with the API function as black boxes. This lack of transparency regarding the specific filtering properties of API is fundamentally problematic for Twitter researchers. In order to produce valid research results, social scientists need to make informed decisions regarding population and sample. Therefore, the chapter starts with a discussion on filtering in relation to data collection methods, which is then followed by a discussion on digital methods (Rogers, 2013b), in relation to this research project. Following this, data collection issues are discussed and methods for collecting data using software are described. As previously stated, this project has moved from collecting data using a hashtag to aim for collecting complete conversations around hashtags. At the time of data collection, there was no software available that had functionality for collecting conversations. Sections 4.3 and 4.4 describe and discuss the move from a hashtag-based and a user-based data collection method to the composite method and issues related to software for data collection. This is followed by a section about preprocessing, which here includes the steps taken to convert the raw data into analysable units.

4.1 Filters as methodological problems

Twitter is not a neutral actor as a platform. An important aspect of filtering is the tailoring of information for users by recommender systems. Filtering works differently to users of the platform compared to users of the API, for example researchers. Users and researchers interact with different technologies, i.e. the user
interacts with a user interface that is not identical to the API the researcher makes use of. From the user perspective one might view Twitter as a more social version of a filtering gatekeeper as its members decide who to follow, but as previously stated the platform has moved from being a utility to emphasising the popular. From the API user perspective, filtering made by the platform on data accessible through the endpoints entails that not all data are available, and it is unclear which tweets are visible in the timeline of any given user at any given time. In relation to ownership, three overarching methodological problems are identified here:

- Restrictions on the data flow or filtering of data flow, as Twitter determines what data are available.
- Filtering decisions depend on a variety of different needs which are not necessarily compatible to the needs of the researcher, partly because of Twitter being a commercial platform with commercial interests.
- Data availability and rules and restrictions considering data availability may change at any time.

Within the cluster around the elite users of the political Twittersphere there are different types of filtering activities which become visible at the moment they are researched. The filtering activities are different depending on type of information flow, but in its most simple terms the information flow is filtered forward as the member sees information shared or disseminated by the actors that the member follows. There are fundamental differences between the two activities retweeting and replying. Retweeting passes content on to a wider audience, which means that the number of people that will see the tweet increases. Contrary to this, replying restricts the content to smaller audiences, since only a specific subset of the followers of both the replier and the user replied-to will see the reply. The fewer the joint followers of the two interacting users, the smaller the audience exposed to the conversation. This is an effect of the way Twitter presents conversations to its users, which was illustrated in Chapter 3. This asymmetrical adjustment of channel width is not apparent to all users, and most certainly not visible in the data collected through the API.

For the researcher, the black box of Twitter API and its regulations constitute filtering mechanisms that require investigation. An important issue to deal with is how software and API steer possibilities to analyse the data and, crucially, researchers using an API must be aware of both built-in limitations and biases
Considering the available API endpoints it is perhaps not surprising that studies of tweeting make use of merely one such endpoint and specifying just one parameter for collecting data. Basically, a filter is created when using the streaming API. This filter can contain a combination of up to 400 keywords/hashtags (track), 5,000 user IDs (follow) and 25 geographic coordinates (locations), and can be updated every other minute. A bias problem is encountered when using only one of these three parameters, or when the search API is used. Only those tweets matching the search query are returned, and, in all likelihood, not all relevant tweets that are publicly made available by their authors. Additionally, the context of the tweets is lost, meaning that the tweets that are part of a conversation are analysed without knowledge of the content replied to or the reaction to the tweets. For a conventional hashtag oriented study, this leads to a partial data set in which reactions to the captured tweets are only possible to analyse if they too include the hashtag. A methodological imperative coming out of this discussion is that Twitter researchers need to understand filtering aspects of the API in order to pursue and develop workarounds.

A further methodological imperative concerns programming knowledge in order to avoid data sets that in a simple way reproduce the bias of the Twitter API. Similar thoughts have been expressed by Larsson (2015), who identified a need for technical skills to be able to deal with the data in the social sciences and the humanities. Knowledge of programming, alongside visualisation and analysis of such data, are not common within these areas (e.g. Steen-Johnsen & Enjolras, 2015). At the start of this research project, there was no software that could gather all the data needed. yTK could however be used to track search words, and Topsy (now a paid service) to gather archived tweets, but only a small sample. Later, at the same time as Study II was published, the system The Digital Methods Initiative Twitter Capture and Analysis Toolset (DMI-TCAT) was released and described (see Borra & Rieder, 2014). Some free tools do exist (see Gaffney & Puschmann, 2013), but whenever the API and the terms of the usage of a social web application change, they need to be updated by the developers.

Other options were considered within the framework of the current project, such as NodeXL and Webometric Analyst. However, to get access to the data needed whilst also avoiding the vulnerability of relying on specific software at a time
when Twitter API underwent significant changes\textsuperscript{8}, the choice fell on the software yTK which constitutes an open source solution with an architecture that could be adapted and extended for the needs of this project. The software was extended with functions for collecting user profile data and lists of friends. From Study II it also became clear that more tweet metadata, such as \textit{in reply to status id}, were needed. This, alongside an adaptation of the methods connected to the streaming API to follow users instead of tracking words, made it possible to collect replies to tweets, thus enabling the mining of threads.

In Study III, the most prominent participants in the conversations around #svpol were selected based on activity, visibility and spreadability during eight consecutive weeks. However, although the data collection was suitable for mining threads, those data were never analysed deeply. A number of problems were identified (see Data collection below), and refinements were needed. The refined method was used for Study IV. The method developed captured most of the data needed, but the data set was still likely to be biased. As noted above, the streaming API allows for following up to 5,000 users, and the set of users can be updated every other minute. More than 5,000 users posted tweets containing the hashtag during the data collection period, which entailed an enforced focus on the most active participants. The criteria for selecting the 5,000 users might seem somewhat arbitrary. The method for Studies IV and V selected the most active users during a one week long sliding window. Other relevant options are to use a shorter window if velocity is high and selecting the most visible participants. Regardless of the length of the data collection window, the enforced focus on the most active users make it less likely that tweets sent by a less active user to another user of a similar activity rate are captured (see 6.7).

4.2 Digital methods

In his book \textit{Information Politics on the Web}, Richard Rogers introduced the concept \textit{web epistemology} (Rogers, 2004). He would much later describe this as...
the study of how natively digital objects such as links, tags and other web-related objects are handled by devices (Rogers, 2013b). This involved a shift in the discussion of epistemology when utilising digital methods, and a move to consider the Internet as a source of data, method and technique (Rogers, 2009; 2013b). Methods that import standard methods from the social sciences and the humanities in this context, labelled as virtual methods, have shifted the focus away from the data of the medium according to Rogers (2009). Both Rogers (2013b) and Thelwall, Vaughan and Björneborn (2005) have proposed correlating web findings with non-web data. Such an approach would also be useful in dealing with potential selection biases, for example, when comparing the identified political affiliations of the categorised users with opinion polls.

Digital methods follow the methods of the Internet as a medium in that they apply techniques that are employed on Internet for “authoring and ordering information, knowledge, and sociality” (Rogers, 2013b, p. 38). How do we follow the medium on Twitter? Some different ideas of this have been presented in the literature. Bruns described his approach as “exploratory”, “open-ended”, involving “experimental trial and error” examination of data to investigate “what useful and reliable data may be gathered” (2012, p. 1329). He also referred to the question posed by Rogers of how to “rethink user studies with data that are (routinely) collected by software” (Rogers, 2013b, p. 20). Another interpretation of the notion of follow the medium is given by Arvidsson et al. (2016), whose data collection approach was based on popularity. A third interpretation is given by the creators of DMI-TCAT mentioned above. Their software work closely to the Twitter data and metadata, and by doing so it leaves “the ‘primary’ material untouched” (Borra & Rieder, 2014, p. 267).

The methodological recommendation of following the medium relates to the earlier discussion on the importance of having knowledge about the API as well as about programming. Obviously, such knowledge makes it possible to work closer to the source data and its metadata. Furthermore, these are skills that are needed to collect relevant profile and follower or friend data from a reasonable set of users. While DMI-TCAT fulfils the following the medium criteria, it is not able to collect conversations. Important questions regarding what data one tool collects compared to another need to be investigated, but this is outside the scope of this thesis.

The data-driven and exploratory approach has also been applied here, with a purposeful aim of working as close to the raw data and metadata as possible. When
decisions of what data to collect within the specified sphere have been enforced, data collection has been steered by activity and popularity (e.g. following the most active 5,000 participants, focusing on followings among the most prominent users of a hashtag). To be able to follow the medium, the strategy taken here was to modify yTK. Its original functionality is restricted to tracking keywords or hashtags by combined calls to the search and streaming APIs. In the first rebuilding phase, functionality for collecting friend and follower data, as well as profile data, was added. For Study III, the basic tweet streaming functionality was changed from using a list of keywords or hashtags (track) to use a list of user IDs (follow). When collecting Data set 3, track and follow were combined.\(^9\) Some metadata fields omitted by yTK were also added, for example the metadata field denoting to which a tweet is a reply to.

### 4.3 Data collection

Tweeting constitutes a new form of human communication and development of data collection method for social science research goals has been lacking. Research has been dominated by two major approaches; user-based and hashtag/keyword-based. There is a further possibility in collecting data given geographical coordinates, which so far have not been utilised to a large extent. As have been argued earlier, researchers need to be aware that these approaches involve sampling procedures at least to some extent outside the control of the researcher. Building on the literature review in Chapter 2 it becomes possible to state that politics related Twitter research in most cases has relied on one approach only, and in those cases two or more approaches have been used, these have not been combined to collect conversations.

In this project the hashtag and user-based approaches have been combined. By simultaneously tracking a hashtag and following the most active participants, replies to the captured tweets are collected. Due to the lack of an archive to search from, it is currently the only way to capture a conversation without firehose access.

---

\(^9\) The API documentation shows an example of how to combine two parameters. See https://dev.twitter.com/streaming/reference/post/statuses/filter
I have argued that software and/or programming expertise is needed to collect, to aggregate and to analyse the data. Technological aspects also involve problems such as having to rely on a persistent Internet connection, software compatibility, and processing power to deal with the data. These issues alongside boundaries of conversations are central to the methodological problems dealt with in this research project.

In Study II, the hashtag tracking method was used in order to map the activities and relationships within the political Twittersphere. When doing so, I became curious of the existence and amount of replies not including the tracked hashtag. This triggered experimentation with a combined approach in which one hashtag and the most prominent participants in the conversations were used as search criteria towards the API. As the hashtagged tweets were collected, the most frequent posters were added to the search criteria, and by doing so tweets sent by and to these were returned by the API. Those tweets that either included the hashtag or were replies to a tweet in the database were stored. The set of followed users within the conversations was updated every other minute. After eight weeks of collecting data with this method, a user-based approach was applied for a longitudinal study of the elite actors of the Swedish political Twittersphere (Study III).

The experimental setup made use of the search API for collecting tweets matching a hashtag, and the streaming API was used to follow the 2,000 most active users of the hashtag, and the 2,000 most active repliers to tweets in the database over the whole period. This design made it less possible for new actors to be included in the user ID list and thus the data set has a possible bias towards those who were active during the first few weeks. Moreover, even if there are examples of Twitter research finding that a few thousand dominate the conversations in different settings (see Study IV), the method has a bias towards the most active users of the hashtags.

However, when Twitter researchers Gaffney and Puschmann (2013) claimed that Twitter was going to withdraw the search API, the method had to be refined once again. The refined method is described in detail in Study IV. In short it relies on two streaming API lists, one with a set of hashtags, and one with a set of the 5,000 most active users of the hashtags during a one week sliding window. To fill gaps created by software, hardware or Twitter failures, and collect threads tagged into the conversations, the software was programmed to query the REST API endpoint.
statuses/lookup for tweets replied-to but not in the database until the whole thread was captured.

4.4 Special considerations of data collection

A methodological insight coming out of the current project is that data collection on Twitter needs to be considered in the context of ownership issues. This aspect appears to be relevant for data collection on most other privately owned social media platforms as well. Three problematic aspects of ownership have been identified in the course of this project. The first concerns restricted and filtered availability as Twitter determines what data are available through its API and how it can be accessed. The second problem is that platform owners such as Twitter design their filtering mechanisms based on numerous corporate and user needs. The preferences of researchers do not seem to be highly prioritised in this design process. The third methodological problem connected to ownership is that the availability of data can change at any time, at the behest of the platform owners. At the start of this project there was always a risk that any chosen application to study would cease to exist, or change existing data access options.

The literature review (Study I) revealed that webometrics already had had its fair share of data ownership issues, most notably the link search facility of search engines being withdrawn. Twitter is no different with its API being updated since the start of the project, with the major update from 1.0 to 1.1 being made between Studies II and III, in March 2013. The original version of yTK would not work following the API updates. Therefore, it had to be reprogrammed (a couple of months after the update, a new version was released). This was not the only time updates of yTK were required. During the first period of Study III the application suddenly stopped working. It turned out that the part of the software used for authorising the application towards Twitter had to be substituted with an updated version. Approximately one year later, between data collection of sets 2 and 3, the application stopped once again. This time it had to do with access to the streaming API. yTK had treated an HTTP message from Twitter case insensitively, which worked fine until Twitter changed the case of some letters in the message. The change made the message a slightly different one than what yTK expected, with the result of yTK interpreting it as an error. This required the location of one line of code which needed to be updated.
As the API was updated, restrictions regarding data access were changed. Overall, a consumer of Twitter data can now access more data in general, but less data regarding for example followers and friends. With version 1.0, 350 calls to the REST API could be made each hour, but that was counted over the whole API. With API 1.1, 60 calls can be made for followers and friends, whereas 180 calls can be made for user data and searching for tweets. It is thus possible to make more requests per hour than before (60+60+180+180=480 for this example), so the challenge is to utilise this effectively and efficiently. All the data for Study II were collected using API 1.0, and for Studies III, IV, and V, API 1.1 was used. Note that these restrictions do not apply for the streaming API, which instead has its 1% limit (see Chapter 2).

4.5 Preprocessing of collected data

In the previous section, various problems relating to platform ownership were discussed. However, there are other methodological issues connected to the unstructured character of user generated content. This is a common problem within web mining research and dealt with through preprocessing. Specifically, projects involving large amounts of user generated data also involve substantial cleaning work (e.g. Batrinka & Treleaven, 2015). Although there are aspects of Twitter data that can be considered as structured and therefore very suitable for data mining as is, the free text fields filled by humans could contain anything. For example, Twitter users could on the one hand specify a very accurate location or on the other hand a very fuzzy location (Crampton et al., 2013). Removal of spam is also necessary. This could be performed using automatic methods but here a semi-automatic method was utilised. A semi-automatic method is more reliable in the context of smaller languages, such as, in this case, Swedish. Simple algorithms were used to create lists of content and users frequently occurring in the database (i.e. frequent hashtags, identical tweets, frequent posters) and these were then manually checked. One example of how to identify potential spammers is a time series analysis of different activity groups. Figure 2 shows how the least active 90% group suddenly became dominant for a few days during the initial data collection for Study III.
This type of spam was produced by a few accounts making use of the hashtag without addressing other users or retweet other users’ tweets. Another type of spam was found in Study IV where hashtagged tweets mentioned celebrities with no apparent interest of the topic reflected by the hashtag, and account names of services such as ShareThis and YouTube. Many of the non-relevant mentioned users found in Study IV were tagged into the conversation just because the hashtags were frequently used. Such automatically generated content can skew the statistics. The researcher needs to be aware of the various types of spam. There is a lot of spam detection methods developed within computer science, but development of such methods is outside the scope of this thesis.

In the preprocessing stage it was also necessary to decide how to define the retweet. There are different styles of retweeting but “RT @username” is a prototype (boyd, Golder & Lotan, 2010). This pattern has been used in several studies (e.g. Bruns & Burgess, 2012; Highfield, Harrington & Bruns, 2013) and also in this project. Other versions include “MT” or “via” among others (e.g. Bruns & Moe, 2013; Pearce et al., 2014), but there were very few examples of these in my data sets. Bruns and Moe (2013) labelled the modified versions (e.g. “RT”) as manual retweets and retweets made via the retweet button as button retweets, a labelling used in this text. The retweet button does however not add the “RT @username” (Bruns, 2012) so following the prototype implies that other types of retweets are excluded from the analysis. The argument for using the prototype to
identify retweets is that tweets including this pattern are definitely retweets. In the software used for collecting data there was no way of finding out whether a button retweet was a retweet and if so, of which tweet. The Twitter API returns a field denoting which tweet is retweeted in a given retweet, and its metadata (Twitter, 2015b), but when the software was adapted to capture this information, it was found that all retweets started with “RT”. It is not transparent whether Twitter adds “RT” to all retweets, button retweets included, or only denotes manual retweets as retweets. This is another example of ownership issues that affect the research.

Preprocessing procedures to ensure anonymity of the tweeters during analysis and in all research products were also necessary. These are discussed in Chapter 5 and outlined in Appendix. Other steps outlined in Appendix are conversion of tweet data and metadata to network graphs and descriptive statistics.

4.6 Summary

The Twitter platform and its API is a filter through which researchers can access data. By tailoring research questions after data availability, much research of the platform has been performed on incomplete data sets. Both the API and software designed to work with the API are likely to be treated as black boxes. By understanding the API and combining different parameters of the endpoints, it is possible to approach completeness of conversations. Open source software such as yTK might be black-boxed to anyone who does not possess programming knowledge; however, with programming knowledge the black box problem can be avoided.

An important issue regards research ethics. In the next chapter, a discussion takes off in the debate of whether informed consent is needed or not when collecting publicly available data, as in the case of Twitter. The discussion covers both ethical and legal aspects of the publicly available data on the platform.

10 Note that this can be tested on one’s own tweets. Such tests were not made here, however.
5 Ethical consideration with regards to user provided information

Researching user generated content on social media platforms gives rise to new forms of ethical problems and considerations. Twitter research involves unique challenges in this area and due to the relatively brief time in which the object of study has existed, there is currently a lack of solid guidelines. As the current project involves a focus on methodological issues, it appears important to approach ethical issues quite extensively, in the form of a separate chapter. Having said that, some issues remain that cannot be fully clarified within the scope of the present chapter. Crucially, it appears insufficient to merely work with traditional ethical norms and guidelines.

While Chapter 4 dealt with the issue of platform ownership as a methodological issue, the current chapter addresses how research ethics become embedded within information policy. The current project must fulfil requirements to various forms of code of conduct: Twitter policy, European policy and national (Swedish) policy. Such information policies can involve conflicting recommendations and rules. In particular, requirements regarding open data and transparency tend to clash with regulations regarding handling of personally identifiable information.

It is notable that national data protection regulation can vary from country to country. In particular, the EU has a tradition of a strong data protection regulation (directive 95/46 EC; COM 2012(11)) which the US totally lacks. In principle this introduces asymmetrical boundaries into the practices of researchers, i.e. a project that is legal in the US could become illegal in the EU.

This chapter covers ethical and legal aspects of research on the publicly available data through the Twitter API. Sources consulted are the Twitter documentation (Twitter, 2016b; 2016e; 2016f), legal documents such as the (Swedish) Personal Data Act (SFS 1998:204) and various researchers’ views on these aspects. Based on the ethical conclusions, the steps taken to ensure anonymity when analysing data are outlined for each data set.
5.1 “You are what you tweet”

Ethical issues in Twitter research are rarely reflected on, partly because Twitter is a public platform (Zimmer & Proferes, 2014). Twitter itself emphasises this in the Privacy Policy (Twitter, 2016e) by stating “What you say on the Twitter Services may be viewed all around the world instantly. You are what you Tweet!” Social media research is dependent on platform owners, their instructions and terms of usage (e.g. Weller & Kinder-Kurlanda, 2016). Privacy and ethics are fundamental issues within social science research, and it has been suggested that it could be problematic to collect freely available data (e.g. Staksrud, 2015).

A fruitful view is to look at tweets and profiles as documents. This view is taken by Wilkinson and Thelwall (2011), who indicate that documents can be analysed without triggering ethical issues, but they also emphasise that researchers should normally ensure that data subjects are anonymous. If we take the document view it is relevant to ask whether the document has reached a level of originality. If so, by maintaining confidentiality the researcher violates many national laws, for instance the Swedish copyright regulation (Ågren, 2000). The analysed tweets would need to be cited in a similar way as a research article. However, when large numbers of research articles are used in a summary such as in Chapter 2 in this thesis, only the ones the researcher exemplifies with are cited. Obviously, when a tweet is quoted it should also be cited, but should tweets be cited if the dynamics of a conversation thread are described, rather than individual messages or parts of messages? One alternative is to reveal the tweet ID of the first message in a thread, which could be done when it is certain that no participant can be harmed. But how can such a decision be made? Even though the tweeter is responsible for his or her own postings, and has the option to delete a tweet after it is posted, it would be problematic if a researcher pointed to a tweet (or a thread the tweet is part of) that is very controversial, illegal perhaps, or reveals sensitive information.

A complicated issue is what constitutes personal data. Narayanan and Shmatikov (2010) have argued that it is difficult to define “personally identifiable information” for two reasons. First, the concept is used in different ways in various types of legislative texts. Second, a multitude of different forms of information can be used for re-identifying anonymous data. Given such critical discussions, it becomes difficult to identify any form of user generated content that in no sense can be regarded to be personal data. According to The Swedish Data Protection Authority, personal data is any piece of information that by itself or combined with
other pieces of information can be related to an individual (Datainspektionen, 2013). Strictly interpreted, a tweet could be seen as personal data, but this depends on the information given by the author of the tweet. We also have the veracity problem – if the tweeter is using a real or false identity, and if it is real, whether the personal name is used or not – which is different between platforms. Whereas Facebook wants one account per person or organisation, such issues are not important to Twitter. Political opinions are considered as sensitive data and, strictly interpreted, these cannot be used in research without explicit consent and clearance by an ethical board (Datainspektionen, 2013). But the Swedish Personal Data Act states that the paragraph that prohibits treatment of sensitive data does not apply when such data are made public, which it arguably is in the case of publicly available tweets and profile descriptions, especially when considering that these data are open to all users of the Internet.

Moreover, when the data are collected from unstructured data, for example running text, most of the Personal Data Act is not applied, but instead there is a legislated feature which states that the personal integrity must not be violated (SFS, 1998:204). There is also an intriguing clash between the Personal Data Act, the Archiving Act, research ethics, reproducibility and ownership. The purpose of the Personal Data Act is to protect personal data (SFS, 1998:204), which research ethics also strive for. The Archiving Act states that a Swedish university should archive what is commonly accessible (SFS, 1990:782). Reproducibility is important to ensure the validity of research, and finally, Twitter grants access to its data under certain conditions.

As noted initially in this chapter, data protection regulation has a strong tradition in the EU but not in the US. With this in mind, it is perhaps not surprising that European researchers have been much more concerned with ethical issues compared to their American colleagues.

5.2 Consent or not?

The traditional approach of seeking consent from participants when performing research on human subjects has been contested in relation to Twitter research. Elgesem (2015, p. 15), for example, argues for “weaker grounds for obtaining consent to use non-private information that individuals themselves have made available in a public forum”, exemplifying with the study of political discussions
on Twitter in which the participants seek public attention. Wilkinson and Thelwall (2011) agree with those who argue that consent in settings where data from the public web are analysed is not needed, but users may have privacy expectations and might not be aware of being studied (Weller & Kinder-Kurlanda, 2016). According to Elgesem (2015), political postings may be collected without consent. As previously stated, using a hashtag signals a wish to be seen publicly and to make a tweet searchable. It could be argued that not using hashtags is the opposite; a wish to not make a tweet searchable (a similar argument could be made regarding geographic location, as allowing Twitter to store such data is a choice the user makes, and one could also question whether retweeting a tweet with a hashtag signals the wish to be searchable). If so, collecting tweets by username or user ID is problematic. The question is then if any reply to a political posting would also count as a political posting.

Moe and Larsson (2012) discussed ethical issues regarding hashtagged studies of political tweeting during an election campaign. Their experiences with the ethical board suggested that this type of data should be treated as survey or questionnaire data, entailing some kind of consent from the respondents. Hashtagging a tweet could be viewed as giving consent as any user of a hashtag should understand that the tweet is findable, but the authors concede that there are tweets (those without hashtags) that might not be viewed as public (Moe & Larsson, 2012). A problem that arises from the view that hashtagged tweets can be collected and analysed without consent is that showballed users, i.e. those who only receive tweets, must be excluded from the analysis. This would require that after collecting data, all mentions of passive users in the data set must be removed from the tweet content and metadata. In fact, just storing a mention of a user that at the time of the tweet has not used the tracked hashtag (so far during data collection) could be equally problematic.

Larsson (2015, p. 145) identified a need to “take the open or closed nature of the data into account and shape our research approaches accordingly” when studying Twitter activity. There seem to be no standards for handling the lack of informed consent in social media research (Weller & Kinder-Kurlanda, 2016). However, when looking at the Twitter Privacy Policy (Twitter, 2016), it seems as by just using the service, this type of consent is given from all users with public profiles. The following passage makes this particularly clear:
For instance, your public user profile information and public Tweets are immediately delivered via SMS and our APIs to our partners and other third parties, including search engines, developers, and publishers that integrate Twitter content into their services, and institutions such as universities and public health agencies that analyze the information for trends and insights. When you share information or content like photos, videos, and links via the Services, you should think carefully about what you are making public. (Passage: Tweets, Following, Lists and other Public Information)

In fact, at the end of the second paragraph of the Privacy Policy, this is rendered even clearer: “When using any of our Services you consent to the collection, transfer, storage, disclosure, and use of your information as described in this Privacy Policy”. The argument that these types of texts are too lengthy and written in “obscure legal language” (e.g. Prabhu, 2015, p. 163) hardly holds here. The Twitter Terms of Service (Twitter, 2016f) are equally emphatic in these respects:

Most Content you submit, post, or display through the Twitter Services is public by default and will be able to be viewed by other users and through third party services and websites. [...] You should only provide Content that you are comfortable sharing with others under these Terms. (Passage: 1. Basic Terms)

You are responsible for your use of the Services, for any Content you provide, and for any consequences thereof, including the use of your Content by other users and our third party partners. (Passage: 5. Your Rights)

---

11 Note that this text has been changed during this study. The text from December 2014 (see https://goo.gl/XkNMaH) is similar to the current one while the texts from March 2013 and August 2012 (https://goo.gl/QP7xwo and https://goo.gl/1whPhh) have been less specific: “For instance, your public user profile information and public Tweets may be searchable by search engines and are immediately delivered via SMS and our APIs to a wide range of users and services, with one example being the United States Library of Congress, which archives Tweets for historical purposes”.
The third policy document of relevance here is the Developer Policy (Twitter, 2016b). The document describes how data should and should not be used. Downloadable data sets can be made available but these can only include tweet IDs and/or user IDs, and not content or metadata. Twitter allows non-automated export of up to 50,000 public tweets and/or user objects per day, in the form of spreadsheets or PDF files, and the like. It is also stated that expressed consent is needed when protected or private content is shared or published. Any user who does not allow this should make his or her account private. The terms clearly state that data posted on Twitter can be used by third party, and as long as Twitter allows its contents to be analysed, by researchers, research companies, opinion analysts, and the like, it will be analysed. However, just because the content is public does not mean the Twitter user wants or expects the content to be archived and analysed (Zimmer & Proferes, 2014). This can of course be seen from another perspective. Elgesem (2015, p. 26), for example, suggests that it is not a “reasonable expectation [...] that the information will not be used in research”. Integrity is not absolute, but rather dependent on context (contextual integrity, see Nissenbaum, 2004).

5.3 Public or private?

The blurring of boundaries between the public and private is undoubtedly a problem (Elgesem, 2015), and due to this we cannot assume that participants in settings like the ones studied here always seek public attention (Eysenbach & Till, 2001). Previous studies have indicated that far from all users read these terms and conditions on social media but those who do find the parts concerning consent understandable (e.g. Custers, van der Hof & Schermer, 2014). But this type of consent does not specify how one’s content and actions can be used. Would a user expect that his/her political viewpoint, which to Moe and Larsson (2012) belongs to the private sphere, would be used as part of an analysis? However, if we view the public profile as a document representing the viewpoint, would the content still be private? The document itself would be considered public and so the political viewpoint has been made public. Twitter considers content in tweets and profiles as public (Twitter, 2016e), so in practice any user of the service considering his/her political viewpoint as private should not state such opinions on a platform where they are made public.
Obviously, it is not only the data collection that might be problematic. Research within computer science has shown that it is possible to de-anonymise users and reveal sensitive information given a social network (Narayanan & Shmatikov, 2009) or even weaker indicators such as texts, geographical features and temporal distributions of activity (Korayem & Crandall, 2013). A network graph is like a fingerprint. Each user has a unique set of followings and interactions, although such patterns may change over time. However, given the findings in these two papers, the researcher must be very careful before sharing a data set, which could lead to unauthorised secondary use and loss of control from the data subjects perspective (Zimmer, 2010). Would it for example be possible to derive a participant’s political position from the tweet IDs representing the tweets analysed in Study V? If all those tweets were collected with the metadata, reconstructing the threads and matching number of tweets and participants with the analysed threads is a trivial exercise. For those threads where a smaller number of participants take part, at least some of their political positions can be inferred.

The qualitative analysis of the threads did not focus on who said what but rather what content was discussed, the existence of echo chambers and consensus formation. As these were the questions, the descriptions of the threads are highly unlikely to lead to a person, except perhaps for a leading politician. The political positions identified were only used for statistical purposes. In the thread analysis, a branch of a discussion was coded as echo chamber or polarising opinions meeting but without any connection to the tweeting user (whose identifiers were encrypted before the content analysis).

### 5.4 An ethical position on Twitter data

Tweets are semi-ephemeral. For most of the users of the platform, tweets are likely to appear as ephemeral. The tweets are displayed in reverse chronology and seem to disappear as new tweets are posted. As described above, data collection is performed in real-time and/or by searching in an archive which includes tweets posted during the last couple of days. But any public tweet and its follow-on conversation are findable through a tweet ID within the conversation. Collecting non-tagged replies to hashtagged tweets falls within a grey area; the user does participate (at least indirectly) in the hashtagged conversation but might not be aware of the tweet being findable using the methods utilised here. While the
(public) reply is by default not filtered forward to those who do not follow both users, it is always visible in the expanded view (View conversation) and findable through the search function. It is public which should not come as a surprise to the user. After all, the point of posting in a public forum is quite likely to make opinions available for anyone.

There are contexts where data collection based only on hashtags is not enough, simply because the follow-on conversation is too large to not consider, which has been proven in this thesis. The ethical standpoint taken here is that data collection based on hashtags and prominent participants is within the ethical space, but that complete anonymisation needs to be done in all of the research products. Categorisation based on publicly available content such as profile descriptions and tweets is also considered being within the ethical space. However, revealing such information in a research product is considered outside of this ethical space, as it could be possible to relate such information to a physical person or organisation.

Based on the research cited in this section and the Recommendations from the AoIR (Association of Internet Researchers) Ethics Working Committee (Markham & Buchanan, 2012), the following considerations have been taken into account and executed in these studies:

- The context is adjudged being the same for researcher and user (no ambiguous hashtags have been used).
- The context is defined by the hashtags; all content is related to the hashtags directly or indirectly.
- When indirect content is accessed, it is accessed as replies to hashtagged tweets (and replies to replies) (Data set 3), and as tweets posted by or to prominent users of the hashtags (Data set 2, in which only hashtags were analysed).
- Hashtag users probably perceive the context as public, users replying to elite users probably too. A grey area is a non-tagged reply to a non-elite user.
- Ethical expectations should be that user IDs, usernames, opinions expressed and tweets are not made public.
- Information collected should not be traceable to an individual or organisation. Who said what will not be revealed.
Access to the data was only granted to me as researcher, and data were purged from the yTK database regularly. The most important issue regards privacy. With political viewpoints potentially being sensitive data, steps have been taken to ensure that the users were anonymised before the analysis took place (see Appendix). No user IDs, names, usernames, profile descriptions or tweets have been revealed in any research product of this project (including this thesis, the articles the thesis is based on, posters and visualisations).

In both cases where users were coded, Data sets 1 and 3, the data fields name and username were left intact for the sole purpose of identifying politicians, mass media actors and celebrities. However, these were anonymised after the coding and before the analysis. In Data set 1, all participants were assigned a numerical ID from 1 to 916 and the usernames, names and user IDs were subsequently removed. In Data set 3, the usernames, including usernames in tweets, and user IDs were encrypted using an irreversible algorithm\(^\text{12}\). The user IDs of the most prominent participants in Study II, the followed users in Study III, and the thread participants in Study V were kept as separate data files disconnected from all other data, including the tweets, for calculating overlaps between the sets.

5.5 Summary

By combining a set of hashtags with a set made up by the most prominent participants of the conversations, ethical issues are raised. It has previously been argued that hashtagged tweets are consciously being made searchable, and thus can be considered public. Capturing the follow-on conversation might be problematic as the content is not explicitly made searchable. The ethical stance taken in this research project is that such content can be collected, but that the identities of the participants must be protected. Hence, all participants were anonymised before analysis of relationships, content and conversations, and by doing so it is also

\(^{12}\) Such encryption is not completely secure as it is possible to encrypt all possible combinations of alphanumerical strings up to a certain length (here: the limit of the field) until the target encrypted string is matched, which is possible if the encryption algorithm is known.
ensured that they are anonymised in all research products. No quotes are used in any research product. For practical reasons there are moments when any type of identification field accompanied by a code representing political stance is visible to the researcher. The data were anonymised as soon as the categorisation was made. Any data not used in the analyses have been removed and all raw data except for tweet IDs have been removed after the acceptance of the articles. In cases where complete anonymisation is necessary, the recommendation would be to use irreversible encryption of usernames and user IDs as soon as data are collected, and before the researcher analyses the data.

With the possible ethical pitfalls and issues in mind, following users and tracking hashtags is necessary to provide some context for those studies that are either hashtag or user-based, otherwise we cannot know what we can infer from such studies. The potential use of data collected might not be apparent when seeking consent (Prabhu, 2015). Similarly, it is not apparent who will become prominent as the conversations unfold. Given that we on beforehand do not know who will participate in the conversations we cannot ask for consent, and with 10,000-20,000 participants it is also impractical to do this. However, as pointed out by Weller and Kinder-Kurlanda (2016), there is a need for more research on user expectancy regarding privacy. Finally, it is important to keep in mind the larger ethical issues involved given the non-transparent activities of platform owners. The platform owners are happy to emphasise how they operate on behalf of the common good but also make use of its members’ data in a non-transparent way (van Dijck, 2013, p. 10-11). Even so, Twitter appears to be more transparent than other platform owners, such as Google and Facebook.

The next part of the thesis summarises the five studies. For each study, the context and position is outlined with a summary and how the study contributes to the thesis project.
6 Summaries of the studies

This chapter summarises the five studies. The first summary concerns the literature comparison of webometrics and web mining. The next four studies, all focused on political Twitter activity, are ordered by the time of the data collection. Study II had its focus on polarisation tendencies in the usage of #svpol (Swedish politics). Study III focused on the hashtag usage within the cluster around the elite users of the same hashtag. Study IV aimed for a complete data set from which conversational threads can be extracted and Study V is an analysis of these threads. Short descriptions of the articles, including publication data, data collection approach and corresponding data set can be found in table 2.

In all of these three data sets a set of around prominent 1,000 users are identified. Section 6.6 includes a comparison of the prominent users across the sets. Finally, as one of the most central contributions of this thesis is the critical discussion on selection bias in Twitter research, this chapter also includes a comparison of data collection methods with different parameters. More specifically, the focus is on collection of conversation-based data. While the composite method has several advantages, questions still remain to solve regarding its ability to collect complete data sets. Moreover, a comparison of the search API and streaming API made by González-Bailón et al. (2014) suggested there is a bias involved when utilising the APIs. Section 6.7 discusses API bias in relation to data collection and evaluates the methods used to collect conversations with regards to completeness. This evaluation is based on a series of tests made with different data collection parameters. These tests are separate from the studies of the thesis.
Table 2. Overview of the articles.

<table>
<thead>
<tr>
<th>Study</th>
<th>Publication data</th>
<th>Approach</th>
<th>Data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>III</td>
<td>Lorentzen, D. G. (manuscript). Is it all about politics? A hashtag analysis of the activities of the Swedish political Twitter elite.</td>
<td>User</td>
<td>2a, 2b, 2c</td>
</tr>
</tbody>
</table>

6.1 Webometrics and web mining

6.1.1 Context and position of article

The starting point of this article was to do a thorough literature review of the main research area for this thesis. I was trying to deal with my identity as information science researcher, and webometrics was a natural field to start working with, but webometrics is not the only field devoted to analyses of the web. From previous
experiences of log analysis, the starting point of this project was the idea of combining information science with computer science. Hence a literature comparison of webometrics from information science and web mining from computer science was a relevant start. As sub-fields of information science and computer science respectively, they have inherited traditions and methods from their parent fields, informetrics and data mining. As previously mentioned in this summary essay, web mining is typically divided into the three web aspects structure, content and usage, and some papers on webometrics have made use of similar sub-areas.

Computer science and information science were found to be the two fields with the most number of Twitter studies before 2013 in the review by Zimmer and Proferes (2014). But the sub-fields web mining and webometrics are of course not the only fields studying web structure, content or usage, especially when it comes to a specific topic such as political use or conversations. Nevertheless, the two fields studied in my review have dealt with and solved problems relevant for studies of web content, usage and structure.

Twitter is not an ordinary web site. It is accessible through the web, SMS, various mobile phone apps or other web sites connected to Twitter through the API. In one way, researching Twitter by utilising the API is similar to traditional data mining, which is performed on structured data in databases. But not all data accessible through the API are structured. There are content, such as the tweets, which are written text. The text in turn sometimes contains references to web pages outside of Twitter through the use of URLs. Web mining, webometric and traditional data mining techniques are all relevant to apply to Twitter data.

6.1.2 Summary of article

Using broad searches in Scopus, a body of literature was found from which articles were selected. For roots and definitions, reviews, early papers and highly cited papers were selected. Since the area for the thesis is social media studies, articles dealing with social web platforms and user generated content were selected. For identifying differences between the fields, articles seemingly being unique for the fields respectively were selected, as were articles dealing with disciplinary identity. A classification scheme including the aspects Topic, Approach, Method, Type of data, and Data type category was utilised. The literature comparison was made with the aid of a taxonomy for categorising disciplines by Becher and Trowler
This taxonomy includes two dimensions. The first dimension stretches from pure (more theoretical and oriented towards basic knowledge) to applied (oriented towards solving specific problems). The other dimension regards to what extent the fields are hard (more quantitative and value-free) or soft (more qualitative and focused on values). Keyword and citation analyses accompanied the analysis of the articles.

The definitions used in the article for each of the fields are “quantitative study of web-related phenomena” for webometrics (Thelwall, Vaughan & Björneborn, 2005, p. 81) and “the discovery and analysis of useful information from the World Wide Web” for web mining (Cooley, Mobasher & Srivastava, 1997, p. 558). In the article, the evolution of the fields, common grounds and differences, and signs of and opportunities for collaboration are outlined and discussed. Focus was on methods and applications, research problems and citations. Both fields were born in the late 1990’s. Web mining was found as being the larger field, if considering the number of publications, with close to nine times as many as webometrics. Both fields have been devoted to all of the three mentioned sub-areas, but webometrics has had its focus on structure (mainly link analysis) and web mining less on structure and more on the other two. Both fields have embraced social media and its user generated data.

A basic difference, which arguable comes from the traditions these fields inherit, is that webometrics is very much focused on exploratory and empirical studies, but with a couple of examples of a combination with method development, whereas web mining is to the same extent focused on the development of algorithms. Both fields do seem to quite large extent borrow their modes of practice from their mother fields. The fields were considered as both being hard and applied, although webometrics was found to be softer and purer in some areas. A key conclusion of the article is that webometrics could benefit from collaboration with web mining, or at least make use of computer science algorithms.

In the literature review it was also clear that the field most vulnerable towards platform owners’ filtering and data access restrictions following API changes is webometrics. A method broker (Thelwall & Wouters, 2005) needs to be aware of changes applied to the APIs and data access. However, the method broker is less vulnerable than the researcher relying on third party software, which will not function following major changes. Algorithms created can always be applied as
long as data are accessed. Webometricians have previously been affected by API withdrawals and are likely to be so in the future.

6.1.3 Contribution to thesis project

This study enabled me to draw sound conclusions regarding the wide array of methods available. The richness of data did not seem to make procedural decisions straightforward and unproblematic. The many examples of developing and redeveloping algorithms within web mining point to difficulties in managing data. The heterogeneous character of big data forces the researcher into many more complex decision-making situations compared to traditional research. When such difficulties are involved outside of computer science, there is a risk that new methods are not developed and current methodology is not reflected upon. Researchers might use a method simply because it is well established, even if it might not be able to perform well enough.

The study allowed me to develop an identity as a researcher, not as a webometrician, but rather as an informetrician with a focus on social media. I am troubled by the way ownership, technology and boundaries restrict the possibilities of doing research on Twitter data and activity. The distinction between being a webometrician and an informetrician is that the former seems to be more devoted to exploratory studies while the latter is more inclined towards method development. But these labels are not static and two researchers with similar interests might label themselves differently from each other. I also see my work built upon web mining. Many small data preprocessing and analysis steps have been performed through the use of software, such as R, built in functions in Python and Gephi. The exploratory part of investigating the data sets involves application of such technology. By utilising programming skills, the boundaries between computer science and information science (and other disciplines) become blurred. It all comes together naturally as part of an umbrella area that has rapidly emerged in recent years: data science.

Regardless of methods used, where they belong, and whether they are built in within software or not, the use of programming to customise data collection tools has been necessary in order to collect data to answer the research questions. This is crucial, because otherwise the researcher would tailor aims and research questions depending on software, which in turn is tailored for the API. The access points of the API sometimes need to be combined to make it possible to collect data that can
answer certain research questions. Programming skills are often needed to do this, either by modifying open source software or by directly accessing the API. These conclusions are supported by Golder and Macy (2015), who stated that programming skills are needed both to access and process the big (semi-structured) data that are produced through Twitter. Among skills required they mention API programming, handling of XML or JSON structures and using regular expressions. Other applicable skills include manipulation and storage of large data sets, which may require distributed programming, machine learning, sentiment analysis and topic modelling.

6.2 Polarised conversations

6.2.1 Context and position of article

The aim of the study was to analyse followership and communication patterns among the most active, visible, and spreadable users of the hashtag #svpol, from the perspective of polarisation. In a pioneer article on political Twitter conversations, Larsson and Moe (2012) outlined the communication patterns among the high-end users and general usage patterns of the hashtag #val2010 (Election 2010). My article builds and extends on this as it includes the top 916 users of the hashtag #svpol and analyses the three aspects followership, mentions and retweets from the perspective of polarisation. This study broke new grounds in combining these aspects in one study. The approach was exploratory and gave birth to a number of new questions, both empirical and methodological.

6.2.2 Data

Based on a pilot study made in the summer of 2012 where tweets matching #almedalen (a political event) were collected, #svpol was chosen as the hashtag to track during four weeks of September the same year. After each week, the collected tweets were run through an algorithm for calculating a value representing the activity surrounding each actor, based on its activity (tweets posted), its visibility (mentions received), and its spreadability (retweets received). The 500 actors with highest values were kept and at the end of the four weeks, the lists of actors were aggregated into a final list of 916 actors who then became the focus of the study. The Twitter API was queried for these actors’ lists of friends and their profile data.
6.2.3 Methods

The analysis was made in two parts; one comprised of descriptive statistics and the other a social network analysis. Social network analysis was applied to the followership, retweet and mention networks. The analysis was accompanied by descriptive statistics considering both the whole data set and the tweets posted by the prominent users. Content analysis was applied to the profiles for categorising into the political camps left, centre-right and right-wing.

6.2.4 Results

The results revealed differences between the mention network and the two other networks (followership and retweets). While the mention network did not reveal any clustering tendencies, the other two did. The users of the hashtag thus seem to prefer to follow and retweet like-minded, but did engage in conversation with other groups. However, the followership graph was only internal (i.e. between the 916 prominent users). A complete followership graph of these followers could perhaps reveal different structures.

When comparing user groups, different behaviour can be seen. The least active group was the least conversational and retweeted more often than more active users. However, members of the category right-wing were the most vocal, and retweeted as often as the least active group. These results highlighted the importance of looking at the defined conversation both as a whole, and its different user groups.

6.2.5 Contribution to thesis project

While the analysis of relationships and interaction focused on the 916 most prominent users, there were in total 10,294 unique users of the hashtag during the four weeks. It is a reasonable conclusion that there are many more silent users following the potential opinion formers, or the prominent participants in the conversations. The prominent participants averaged 3,065 followers. Add to this that by following another user, Twitter users reveal significant information about who they relate to. The study looked at activity from different angles; the interaction and relationship types, the political groupings and the activity groups.

The results clearly showed how different behaviour can be across the ideological groups and user groups, and how polarisation appears in different ways when comparing followership, retweet and mention networks. This highlights the
importance of being aware of what is being studied. Polarisation is visible in the first two, but not in the mention network. The mention network was incomplete however, as it did not take follow-on conversation into account. Also, the patterns found in the different analyses triggered new questions. What are the defining traits of a mention? How is it compared to a reply? What happens to a retweeted message when it crosses a boundary into another political territory? These questions need to be addressed within the framework of qualitative studies.

6.3 All about politics?

6.3.1 Context and position of article

Study II raised a number of new questions, of which one concerned the activities of and around what could be labelled as the tweeting elite or potential opinion leaders. The purpose with “Is it all about politics?” was to investigate what topics are discussed by prominent participants of the hashtag #svpol and how the topics are connected. The study made use of hashtag and redistribution analyses of the activities of and around 985 users. The data for the analysis were collected during three separate time periods, each lasting four weeks. There are few examples in earlier research focused on what the most prominent Twitter participants within political topics tweet about apart from politics. One example is the study by Ausserhofer and Maireder (2013), in which it was discovered that less than a fifth of the tweets were about politics.

6.3.2 Data

985 prominent users within the #svpol cluster were identified through an eight week long pilot study. During these eight weeks, the tweets including the #svpol hashtag and the follow-on conversation were collected. The user IDs of these users were put in a list sent to the streaming API which continuously returned tweets posted by and to these users, as well as retweets of their tweets. Data were collected during three periods over one year (2013-06-02—2013-06-30, 2013-11-24—2013-12-22, and 2014-05-11—2014-06-08). 2,303,403 tweets from 154,993 Twitter users were collected, of which 1,288,746 were posted by the followed users. 54% of the tweets had an English language code and 44% had a Swedish code.
6.3.3 Methods

To identify the most prominent hashtags over the three periods, the number of tweets and users per each hashtag and day were counted using computer assisted methods. The top 100 hashtags overall for the complete set (including retweets) and a subset (excluding retweets) were used as basis for the analysis. Trend analysis was made to map the activity of the different user groups (the elite set and the full set which includes those interacting with the elite) with focus on mentions, retweets and original tweets, as well as for outlining trends regarding usage of hashtags.

Co-word analysis, a technique for outlining the main topics and relationships between the topics (e.g. Borra & Rieder, 2014; Callon, Law & Rip, 1986; He, 1999), was utilised to outline how the hashtags were connected to each other. Here, only non-retweets with at least two hashtags were used as data. When retweeting, the additions of “RT” and a username might result in hashtags at the end of a tweet being excluded or cut-off, thus distorting the results. A hashtag metadata field is attached to the tweet, but these data were not collected by the software in this study.

6.3.4 Results

A large share of the hashtags was related to politics in a broader sense, but with a few exceptions such as the #nowplaying hashtag. Political hashtags at international, national and local levels were identified. Some hashtags were stable over the three periods, but event-related hashtags with occasional spikes in the usage were also found. Different patterns of usage were found when comparing the most prominent hashtags in the full set (including retweets) with the subset (excluding retweets). The activities of the elite users were stable over time and focused on national politics. The redistributed material fluctuated more and was clearly event-related. Sudden events such as demonstrations and scheduled events such as elections were followed by an increase in activity by the followers of the elite users. A notable sudden event during data collection was the death of Nelson Mandela. Among the more stable hashtags amplified through retweets were those related to Wikileaks and whistleblowers. Such hashtags were among the top 20 most used by the elite users, but not as prominent as when the full set of tweets was considered.
The network of co-occurring hashtags indicated that hashtags were closely related to each other. No distinct clusters could be identified although the network could be divided into a national domestic part and an international part. The Swedish political Twittersphere is seemingly related to a global political Twittersphere. This finding could be of interest to follow up on. Who are the bridging users between the domestic and the global spheres?

6.3.5 Development of followership graphs

Study II provided a snapshot of a followership graph showing indications of polarisation. In this study, two similar graphs were made, both representing internal followings. These were made outside of the scope of the article and were based on lists of friends for each user, collected after the first (2013) and third (2014) data collection periods. Hence, the graphs indicate how followership changes over one year. Table 3 shows a summary of the networks. Density measures the connections between the nodes divided by the maximum number of connections in the graph. In the 2013 data, 982 users followed at least one other user within the group. In the 2014 data, the corresponding figure was 925. Even though there were fewer connected users, more connections were created among the users. Thus the network became denser during the year.

Table 3. Summary of followership graphs. Note the differences in density.

<table>
<thead>
<tr>
<th></th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Density</strong></td>
<td>0.069</td>
<td>0.088</td>
</tr>
<tr>
<td><strong>Avg. degree</strong></td>
<td>67.636</td>
<td>80.997</td>
</tr>
<tr>
<td><strong>Nodes</strong></td>
<td>982</td>
<td>925</td>
</tr>
<tr>
<td><strong>Edges</strong></td>
<td>66,419</td>
<td>74,922</td>
</tr>
</tbody>
</table>

Both graphs were very similar to the followership graph in Study II (Figure 3). In all three cases one cluster was more distant to the other two. This was the right-wing cluster in the polarisation study. In these two graphs, the cluster detection algorithm modularity was utilised to identify and colour the main clusters. It is very likely that the red clusters at the left hand are dominated by right-wingers.
Figure 3. Followership graphs. Left: 2013, right: 2014. The graphs reveal three major clusters, two of them being closer to each other.

Similar to Study II, the clusters to the left are denser than the other two (Table 4). This means that there are more connections between the nodes in this cluster compared to the other clusters, i.e. the members of the red cluster follow each other to a larger extent than the members in the other clusters. Here, the community detection algorithm Modularity (Blondel et al., 2008; Lambiotte, Delvenne & Barahona, 2009) in the network software Gephi (Bastian, Heymann & Jacomy, 2009) was used, as opposed to Study II in which the users were categorised manually.

Table 4. Density comparison, followership networks.

<table>
<thead>
<tr>
<th>Year</th>
<th>Cluster position</th>
<th>Density</th>
<th>Nodes</th>
<th>Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>Top left</td>
<td>0.175</td>
<td>313</td>
<td>17,113</td>
</tr>
<tr>
<td></td>
<td>Top right</td>
<td>0.122</td>
<td>265</td>
<td>8,561</td>
</tr>
<tr>
<td></td>
<td>Bottom right</td>
<td>0.146</td>
<td>394</td>
<td>22,566</td>
</tr>
<tr>
<td>2014</td>
<td>Left</td>
<td>0.218</td>
<td>285</td>
<td>17,653</td>
</tr>
<tr>
<td></td>
<td>Top right</td>
<td>0.156</td>
<td>308</td>
<td>14,723</td>
</tr>
<tr>
<td></td>
<td>Bottom right</td>
<td>0.193</td>
<td>323</td>
<td>20,049</td>
</tr>
</tbody>
</table>
6.3.6 Contribution to thesis project

This case study revealed some intriguing methodological issues. During the first round of data analysis, some spikes in activity and sudden increases in share of retweets were found. This is a typical reaction to sudden events. A date-time delimited database search revealed that Nelson Mandela was referred to very frequently during one day, both with the name but also as a hashtag. However, as the trend chart for the non-retweet set revealed, Mandela was not referred to frequently enough as a hashtag for it to be visible among the most used hashtags. This initial finding was the main reason for doing two analyses of the hashtag trends, one excluding retweets and one including retweets. This is yet another example of how different results can be gained with small methodological tweaks.

Another reflection is about co-hashtag analysis. While laying the indexing task on the tweeters (or more accurate: benefitting from their behaviour as a part of data analysis) makes them (unintentionally) co-researchers in the way that they themselves describe what they are tweeting about, it is not necessarily the case that they are the ones who best perform this action. But is there other ways to create topical maps of two million tweets? One could argue that co-word analysis could be performed using the full texts, and not only hashtags. The problem with such an analysis is that it is time and computer power consuming and that it would require part-of-speech analysis to provide more accurate results. Here it is assumed that a co-hashtag analysis is a good estimator of the most popular topics and of how they are related.

6.4 Approaching completeness

This article was co-authored with Jan Nolin who contributed with contextual and conceptual parts to the text. I was responsible for most of the text, data collection, analysis and evaluation.

6.4.1 Context and position of article

There are many examples of Twitter studies using data collection methods based on either keyword or hashtag matching, or a set of users. In contrast to this, investigations of conversations are rare. The hashtag is in this thesis considered to be a problematic unit of analysis for two reasons. First, from a digital methods perspective significant parts of the set of tweets collected through hashtag,
keyword or user-based approaches are intertwined in threads and focusing on a fragmented data set tends to misconstrue discussions. Second, the API has a filtering presence which tends to skew the empirical material for the researcher. This article presented and evaluated a new method for collecting conversational threads.

To prove that there is conversation beyond the hashtags, a case thought to be more conversational than most was used. The data collection method was first tested during spring 2013, but after some flaws and possible stumbling blocks were revealed the method was refined during autumn 2014. The first data collection test was performed in December 2014. The case chosen was the extra-election scheduled in March 2015, but as this was cancelled in late December, the focus turned to the December Agreement. The main aim with this study was to collect a more complete conversation around a set of hashtags than what a hashtag-based data collection approach would provide. Moreover, the study attempted to assess and discuss the possible bias introduced by the researcher through data filtering during the various phases of data analysis. The article compared the hashtagged set with the full set which included follow-on conversation. It also compared different network types with different conversational levels.

Is the restriction of Twitter data stream problematic for researchers? Given common restrictions regarding research budgets the implications for university-based research are substantial. Gaffney and Puschmann (2013, p. 57) have suggested that restrictions of APIs leads to “tailoring research questions to data availability”, thereby avoiding multifaceted and complex issues. This also creates an asymmetrical relationship toward corporations that acquire full access.

6.4.2 Data

A set of hashtags representing political conversations related to the December Agreement was identified through a pilot study of #svpol in December 2014. Data were collected by simultaneously collecting tweets matching either of a set of 13 hashtags and following the activities around the 5,000 most active participants in the conversation, using the streaming API for both. Replies to these 5,000 actors were recorded if they were replying to a tweet in the database. By doing this, follow-on conversation related to the collected tweets was captured. All hashtags were related to the election and the December Agreement. The hashtag for general Swedish political conversations (#svpol) was also included. The data collection
started a couple of hours after the announcement of the cancellation and was ended two weeks later. In total 177,847 tweets from 16,349 users were captured.

6.4.3 Methods
I performed the tests on networks emanating from the data set. A subset including all tweets with any of the hashtags (Set A), and the set including all tweets collected (Set B) were compared. I then created four different communication networks: Set A mention, Set B mention, Set B retweet, and Set B reply. These refer to different levels and types of conversation. In Set A mention, only hashtagged mentions are included (talk about mentions, and initiatives to conversations), in Set B mention, all mentions are included (all conversational tweets not being retweets), in Set B retweet, all retweets are included (relaying and diffusing information), and in Set B reply, only replies are included (having conversation). Overlaps between conversational sets were studied in two steps. I used Jaccard coefficient to calculate overlaps of top 1,000 and 100 actors between Set A mention, Set B mention, Set B retweet, and Set B reply with regards to the network metrics PageRank (see Brin & Page, 1998), weighted degree and betweenness centrality.

6.4.4 Results
The approach taken showed that it is possible to mine conversational threads from Twitter through use of API without firehose access. The evaluation demonstrated that a large portion of topically related tweets did not include the chosen hashtags and that communication beyond the hashtag can be extensive. This was proven by comparison of mention networks where the hashtagged network was much less dense than the network including all mentions. The full network involved 9,036 users compared to the hashtagged network with 5,497 users. The number of interchanges through mentions was in the full network 59,178 and in the hashtagged network 16,429. This indicates that Twitter users heavily engaged in threaded conversations could very well be excluded from analysis of more traditional kind, where the focus is on top 1,000 actors based on number of messages sent and received.

The comparison over prominent users in four different networks confirmed this. There were overall small overlaps between the various network types and metrics with the one exception Reply versus Set B mentions (as the former is a subset of the latter). The differences in terms of prominent users in the various networks
suggest that the ones who are prominent in a network mirroring one type of communication might not be prominent in a network based on another type of communication. Specifically, differences between the retweet network and the thread-based network were large. One reason for these findings could be the effort level required for passing along a message compared to engaging into a conversation. Another reason is the asymmetries in sending messages to others and receiving messages. As concluded by González-Bailón et al. (2014), the users often mentioned are not necessarily the ones that are most active.

A possible drawback with using this approach is the enforced focus on the top 5,000 participants in the conversations, but if these are addressed by the tail end participants those tweets can be captured as well. The case in itself might also be more conversational as other topics and/or in other contexts. The method might not be feasible for conversations with many more users or a high frequency of messages. The method should be tested on such conversations and, ideally, such tests should be compared with firehose data. Having acknowledged this drawback, the method should be applicable for several contexts where conversations develop.

6.4.5 Contribution to thesis project

The conventional hashtag or keyword-based topical Twitter research misses out on the follow-on tweets or communication (e.g. Bruns & Moe, 2013; Bruns, 2012), which is comprised by replies to the tweets matching the search criteria for data collection (i.e. a hashtag) but not themselves matching the same criteria. This type of research approach has been utilised in many studies, some of them in a similar setting as the present one (e.g. Larsson, 2014; Larsson & Moe, 2012). Another approach is to follow the activity of a set of users to collect tweets posted by or to these (e.g. Ausserhofer & Maireder, 2013). This approach captures replies as long as not a user outside the set of followed users is replied to by another user outside this set.

The paper extends the toolbox for a Twitter researcher by showing how it is possible to collect (almost) complete conversations around a given topic. It is postulated that in thread-based networks, all tweets are conversational as they either are replies or replied-to (obviously it is possible to talk about person x in a reply to person y). This claim is not controversial. After all, replying to someone has a conversational purpose. The study of the conversation as unit is arguably the most underdeveloped sub-area of Twitter research so far. It is remarkable that there
are so few attempts at collecting conversational threads by researchers without access to the firehose. Likewise, it is surprising that among the studies utilising firehose access, there are no examples of analyses of conversation threads (it is unclear how the data set used by Kim and Yoo, 2012, has been collected).

6.5 Twitter conversation dynamics of political controversies

6.5.1 Context and position of article

In December 2014 it was decided that a re-election would be held in March 2015 after no party or constellation of parties had won majority in the Swedish general election. For the first time in contemporary Swedish politics, a budget proposition failed to receive enough support in the Parliament. But on the 27th of December it was announced that the election was cancelled following an agreement between seven parties, named the December Agreement. Given the sudden eruption of emotional political discussions on Twitter at this time, this seemed to be a unique opportunity to study conversations on this platform.

The purpose of this paper was to investigate the lifecycle of and participation in controversy conversations on Twitter following a political event by analysing conversational threads. Here, some of the questions not dealt with in Study II were pursued. While the (incomplete) mention network might indicate that opposite points of view meet in the conversations, at least among the most active, visible and spreadable participants, it remains unclear whether the participants reach consensus or if they keep hold on to their viewpoints.

6.5.2 Data

Data set 3 was used (see 6.4). A subset of the threads based on volume and number of users was selected for content analysis and descriptive statistics.

6.5.3 Methods

Trend analysis of shares of mentions, replies, retweets and original tweets was utilised to understand how tweeting behaviour is affected by sudden events. Content analysis (e.g. White & Marsh, 2006), useful for interpreting statistics gathered by quantitative methods (e.g. Thelwall, 2009), was utilised to investigate themes and conversation dynamics in the conversation threads. The analysis was based on the concepts of controversy conversations (e.g. Mazur, 1987; McMullin,
A combination of social network analysis, statistical analysis and content analysis was utilised to analyse the threads. In the first step, the overall life cycle of the conversations and characteristics of the threads discussing the December Agreement were outlined. In the second step, threads were sampled for qualitative content analysis focusing on the identified phases *sparking of conversation*, *development* and *dynamics* of conversations, and *closure*. Threads were sampled until saturation was reached.

### 6.5.4 Results

Looking at the overall lifecycle, the activity immediately after the announcement of the agreement was more focused on information diffusion than having conversations. During the first 24 hours, the least active group was far more active than during the rest of the studied period. After the first day, the activity shifted from information diffusion to having conversations. At the tail end of the data collection period another event, the attack of Charlie Hebdo, influenced the activities in the conversations. Following that event, the share and number of tweets of the least active group increased once more, and the share of conversational tweets decreased. Generally, the most active 1% group had the largest share of tweets apart from the first day. When overall activity was low, the dominance by the most active group was substantial. The group comprised by the 9% highly active users was stable with regards to share of tweets per day.

The December Agreement was discussed intensively during the first four days and then substituted by other topics. Citizens dominated the analysed threads but there were a few examples of participating political actors, who are probably expected to have a more restrictive approach to their participation in social media. Echo chambers as well as discussions where non like-minded argued with each other were identified. In most of the analysed threads different opinions did meet, but consensus formation was rare. There were many interactions but also many examples of replying tweets not inviting further conversation. Politicians were often addressed, mainly by citizens declaring an opinion, but in some tweets also demanding answers. Very few examples of replies by politicians to such tweets were found. Many threads developed quickly, had a short lifecycle and short time spans between the replies. Hence the conversations could be difficult to participate in. The main conclusion of the study was that Twitter might not best be described as a forum for democratic discussions. Rather, it may be depicted as a source of opinions and reactions to political statements. The conversations involved strong
emotional currents including sense of betrayal by political leaders of the
democratic system, the voters and ideological positions.

The study confirmed that conversations expand beyond the hashtag although few
examples of democratic debate were found. Twitter conversations seem more
likely to develop within tightly knit groups followership-wise as compared to
ideals of “small worlds” and “6 degrees of separation” (Milgram, 1967). One
possible explanation for this can be that the way Twitter presents conversations to
its users (see Chapter 3) makes a reply visible in the timelines of only the followers
of both the sender and receiver. In a sense, then, the technology of Twitter seems
to favour development within something akin to echo chambers, but in this case
the conversations did not involve only like-minded. The study concluded that the
community within the political Twittersphere is fluid. It is loosely knit at hashtag
level and ad-hoc communities are created when a reply is made.

From a methodological point of view it is interesting to note that only 442 tweets
of the 3,472 included in the 61 threads about the agreement included any of the
tracked hashtags. Hence, hashtag usage is not common in replies. This entails that
if, for example, sentiment analysis is to be made on tweets in contexts where
threads are likely to emanate, the hashtagged tweets would not be a sufficient or
representative sample of the opinions expressed.

6.5.5 Contribution to thesis project

The solution of the completeness problem resulted in new and more complete
conversation data to analyse compared to data sets collected with hashtag or user-
based approaches. In Twitter research, this type of data is unique. The involved
methodological development allows attention to be shifted from the tweet, the
profile and the relationship as unit of analysis to the whole conversation as a unit.
Tweets can be analysed in context.

This exploratory study is an example of how the conversation unit of analysis can
be analysed. Further research is much needed here in order to develop a rationale
for studying threaded conversations, but also to gain an understanding of how
conversations start, evolve and close. Furthermore, as previous research of political
conversations has dismissed deliberative democracy and the like on the grounds of
incomplete conversations, analysing conversational threads could be relevant to
political scientists.
Although the study found examples of conversations on Twitter, the claim in 6.4.5 that all replies have a conversational purpose is somewhat rebutted by the findings. On a closer inspection, it is clear that there are many replies that do not invite further discussion. A similar finding was made by Mascaro and Goggins (2015). In their analysis of tweets sent to the moderator of a televised debate, only every fourth was an attempt to conversation.

6.6 Users in the studies

The three studies focused on the same topic and zoomed in on the activity of a smaller group; the 916 and 985 of the most active, visible and spreadable users of the #svpol hashtag and the 1,011 participants in conversation threads about the December Agreement. The overlaps between these sets of users were fairly small. The Jaccard Coefficients were 0.27 (Data sets 1 and 2), 0.08 (Data sets 2 and 3) and 0.05 (Data sets 1 and 3). Because of the different approaches to select users (i.e. those who are active in conversations might not be the same who have a high general activity or spreadability) and the timing of data collection, differences between the sets of users in Data set 3 and the other two are expected. This is also due to Data set 3 being more complete, and the usage of a larger number of hashtags in the data collection process. Not the same users receiving tweets are the most active in mentioning others (e.g. González-Bailón et al., 2014; Wang, Wang & Zhu, 2013). It is more surprising that there is such small overlap between Data sets 1 and 2 given that data were collected for Study II in September 2012 and for Study III in April and May 2013 (the first round). The inclusion of follow-on conversation in Study III could help explain some of the differences. In each study, I have also found that accounts had been removed or shut down, and maybe some of the lack of overlap is an indication of the persons or organisations behind these accounts creating new accounts.

6.7 Further testing of the composite method

When González-Bailón and her associates (2014) compared the streaming API with the search API, significant differences were found. These results contradicted my own one week test, in which differences were found to be fairly small. All 45,449 tweets retrieved from the search API were also retrieved from the streaming
API, but 2.4% of the tweets in the streamed set were not included in the search set. It is possible that the geographic locations from which González-Bailón et al. (2014) collected data matter. As the researchers made use of a Spanish case and collected tweets using the search API from Britain, and using the streaming API from Spain, the search API might have delivered results that were (algorithmically) adjudged as being more relevant for a British searcher.

The intent of this section is to present the results from a number of tests which were made in order to allow for better understanding of this discrepancy, as well as testing the ability of the composite method to capture complete data sets. The tests made use of two cases, Swedish politics (#svpol) and Australian politics (#auspol). Data were collected from the same geographic location for two weeks per hashtag. The composite method was compared to the hashtag method by utilising the streaming API for the first and the search API for the second. The composite method was used with different parameters for each of two weeks. During the first week, the 5,000 most active participants in the conversations during a 72 hour long sliding window were used as streaming parameters, alongside the hashtag. When using the streaming API to track keywords and user activities, there is a risk of streaming more content than the 1% limit introduced by Twitter in its API. Hence, during the second week the parameters were set to 1,000 users and a 24 hour window.

The size of the follow-on conversation differed between the sets. #auspol seems to be a larger Twittersphere with 127,025 hashtagged tweets posted during the first week and 184,070 during the second week. Collecting follow-on conversation increased the first data set by 22% and the second by 17%. Corresponding figures for #svpol were 45,199 tweets (29% increase by including follow-on conversation) and 44,522 (26% increase). Looking at the comparison between the APIs, the search API outperformed the streaming API in terms of completeness of tweets including the hashtag when the most active 5,000 users over 72 hours were followed, but more tweets were collected with the streaming API when 1,000 users were followed (Table 5).

The tests show that there are no major differences between both APIs, but recall that the streaming API returned all tweets that the search API returned when not collecting follow-on conversation in the first one week test. When using the composite method, not all tweets returned by the search API are collected with the streaming API. Interestingly, the sets collected with streaming API did miss out on
tweets collected with the search API already from the start, regardless of how many users were followed.

Table 5. Number of hashtagged tweets and overlaps per method and parameter.

<table>
<thead>
<tr>
<th>Hashtag</th>
<th>Number of users</th>
<th>Method</th>
<th>Tweets</th>
<th>Share of tweets in corresponding set</th>
</tr>
</thead>
<tbody>
<tr>
<td>#auspol</td>
<td>5,000</td>
<td>Stream</td>
<td>127,025</td>
<td>95%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Search</td>
<td>130,316</td>
<td>97%</td>
</tr>
<tr>
<td></td>
<td>1,000</td>
<td>Stream</td>
<td>184,070</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Search</td>
<td>183,604</td>
<td>98%</td>
</tr>
<tr>
<td>#svpol</td>
<td>5,000</td>
<td>Stream</td>
<td>45,199</td>
<td>93%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Search</td>
<td>46,809</td>
<td>97%</td>
</tr>
<tr>
<td></td>
<td>1,000</td>
<td>Stream</td>
<td>44,522</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Search</td>
<td>43,483</td>
<td>96%</td>
</tr>
</tbody>
</table>

In an attempt to understand if tweets were missed out on was due to the 1% streaming cap being reached, all streamed tweets were stored in a separate table just to keep track on how many tweets were streamed during a given hour. However, the number of tweets was not anywhere near the 1% limit so the problem must lie elsewhere. The architecture of yTK makes use of a MySQL database which is not the most efficient database solution for storing large data sets (e.g. Bruns & Liang, 2012). This could entail that the problem has to do with software. Regardless of what causes the problem, it seems as when streaming by hashtags and user IDs simultaneously there is a data loss which increases the more users are followed. These results indicate that there are biases related to data collection. It seems as it is not possible to collect a complete set when attempting to capture the follow-on conversation without firehose access. The problem seems to apply for the hashtagged tweets that comprise the basis for the data collection, but missing tweets in the follow-on conversation cannot be ruled out.
6.8 Summary

The five studies have contributed to the thesis project in different ways. The first study highlighted differences in approaches to web data from two research fields, webometrics from information science and web mining from computer science. By positioning myself as an informetrician with a focus on the social web and influenced by web mining, I found a methodological space for the project. Study II mapped the relationships and communication among the most prominent users of a hashtag during four weeks. The empirical findings brought new knowledge to the research on this topic by showing that followership and retweet networks are polarised, but that the mention network was not. The user-based Study III, in which the most prominent participants in political conversations were followed over one year, contextualised the findings in Study II. It was concluded that most of the hashtags used by these actors were related to politics, and that the actors are related both to a Swedish and an international political Twittersphere. Studies IV and V combined the two approaches in the aim to collect complete conversations given a set of hashtags. While Study IV identified significant differences in sets of prominent participants depending on data sampling, Study V focused on both quantitative and qualitative aspects of the conversations.

Throughout the whole project, the method development has been aimed at reaching completeness, and in the process the focus has shifted from data-driven research (Studies II and III) to question driven research (Study V). With access to the complete conversations, research can be focused on the various aspects of the actual conversations on the platform. However, as the tests of the composite method in 6.7 indicated, completeness is never reached without firehose access. It is yet unknown what causes this incompleteness when streaming with hashtags and user IDs simultaneously. Nevertheless, next to complete conversation threads can be mined by using the API within what is possible according to technological aspects and whilst still not moving outside the rules and regulations of the API.

In the next chapter, the research questions will be discussed from the methodological and empirical results in Studies II-V.
7 Discussion

Social media researchers have had a focus on the platform of Twitter due to straightforward access to its data, either through firehose, whitelisting or common access to the API. This has led to a clustering of research around Twitter as well as some development of platform specific methodological and theoretical discussions, although this thesis has shown that there is more to do within this area, especially regarding reflections around data collection issues. Even though data have been easily available through the API, research on Twitter has been problematic. The Twitter-research community has been continuously affected by new policies and regulations. In many cases, this has made it impossible to replicate earlier studies (data cannot be shared in a straightforward way) or even reuse previously developed methods (methods adapted for API v 1.0 are obsolete when applied to v 1.1). The API issues are familiar within the webometrics research field. The method broker, developer and evaluator spoken of by Thelwall and Wouters (2005) is challenged by such issues and need to be aware of changes in API, access, rules and regulations and ethical considerations. Another problem is that sampling is initially performed by non-transparent filters introduced by the platform owner. With data sets collected in partially unknown ways, the researcher has a problem with unknown unknowns, not knowing what data were not collected. Concerns about how large or representative the sample is compared to a complete set (e.g. boyd & Crawford, 2012), and the risk of giving too much focus on the most active participants (e.g. Gaffney & Puschmann, 2013; González-Bailón et al., 2014) further complicate the data collection issue.

One way of working around this problem is to use different and adapted methods over a longer time period to collect a more cohesive set of data, making case studies more information rich. The four studies of Twitter activity and relationships included in this thesis paint a more comprehensive picture together.
than what they would do individually. The incomplete mention network in Study II is complemented by a conversation analysis in Study V. The retweet patterns among the most prominent users of a hashtag in Study II is complemented by an analysis of what topics are filtered forward and amplified by the followers of the elite users in Study III. Similarities in the structures of the followership graph in Study II and those in Study III are striking despite the small overlap in users in the two data sets.

The purpose of this thesis was to critically engage in discussions about both methodology and specific methods for collecting and analysing Twitter interactions and content. The process started with a data-driven approach but then moved towards investigating methods for collecting and analysing conversations. This resulted in the development of the composite method. By capturing follow-on conversations, i.e. tweets not matching the search criteria but tied to the collected tweets through the reply function, it is possible to analyse (next to) complete conversations. With a complete data set, a relevant issue is how to analyse the conversations. Even if the data sets are large and the number of users involved is usually more than can be handled manually, we also need to make use of qualitative methods of analysis. The conversation-based data set opens up for new questions and new kinds of investigations, for example, how conversations evolve, what the character of the conversations is, and why people interact in the ways they do. Tweets and conversations can also be analysed in context. Such investigations are helpful for explaining for example the communication patterns, trends, and development of conversations, which can be identified with quantitative methods. In sum, this project has evolved from the mapping of relationships and communication within a hashtagged space towards aiming for completeness and collecting conversational threads, and finally to an analysis of what the characteristics of the conversations are and a discussion of how this space relates to a larger context.

The thesis has been guided by three questions:

- What kind of problems for collecting and analysing data can be identified within contemporary research on Twitter-based political communication?
- In light of existing difficulties, which kinds of approaches can be developed in order to improve on current research practices?
What could be the relevance of such methodological contributions for other Twitter research investigating different contexts?

As previously noted, there are a number of ways to utilise the Twitter API to collect tweets. The ones that are most often used by researchers build on hashtags, keywords or user IDs. It is also possible to stream a 1% sample of all tweets and to stream tweets by specifying locations. The latter two can immediately be discarded if the purpose is to collect conversations around a topic.

Study IV clearly showed that substantial parts of the discussions are missing with a hashtag-based data collection method. Of course, not all topics have these levels of interaction, but nevertheless a researcher cannot analyse a part of a conversation without knowing what is not captured. Crucially, without the follow-on conversation, one cannot study conversation. It is easy to be seduced by the big data flowing through the filters of Twitter, but researchers need to look beyond this and embrace the opportunities of the conversations extending beyond the hashtags.

The research questions will be discussed from three main themes identified through the process of the thesis work. These are:

- The conversation and its boundaries.
- The filtering platform and its affordances.
- Sampling, bias and completeness.

The chapter will conclude with a discussion about what insights can be gained by studying Twitter data based on the results of Studies II-V, and reflections about when the three data collection methods used in this thesis are best suitable. What are the characteristics of this kind of empirical material? As researchers are exploring the boundaries and the potential of this new form of social data, there is an increasing need to ask questions on what can and what cannot be done with it and to think about when Twitter data could replace other (social science) data.

### 7.1 The conversation and its boundaries

Given research made on the political Twittersphere, there is clearly a need to develop methods to capture and analyse conversations before any attempt to determine to what extent Twitter enables deliberative democracy. Having solved
the completeness issue (at least partially), this thesis moved on to analyse threaded conversations. The definition of a Twitter conversation as entity used for this thesis is *all tweets connected through the reply functionality*. This is a technological definition that has been used in order to create boundaries for collection of data. However, there is also a need for researchers to further investigate the character of Twitter conversations and to more clearly define what it means to tweet. There are crucial differences between face-to-face conversations and their counterparts on Twitter. The “analogue” conversation according to Pask (1976) is a means to reach an agreement of understanding of a topic. The conversations are interactions between cognitive systems, leading to the construction of knowledge. Obviously, conversations on Twitter are much messier. Participants in the studied Twitter conversations in Study V might agree on the main topic, but perhaps not reach an agreement of understanding of sub-topics. The motivation does not seem to come from learning something different or new.

Although Twitter and numerous researchers have created a convention of talking about certain activities on this platform with the word “conversations”, we should be aware of the problems involved. Maybe conversation is not the right word and perhaps the use of it creates problematic connotations. Twitter conversations seem to have some aspects in common with interactions on online forums. Reflecting on such discussions, Herring (1999; 2010) argued that computer mediated communication can be seen as conversations, despite identifying a number of differences compared to the face-to-face counterpart. Except for the obvious lack of visual cues from the participants, online conversations are shorter and have a different turn-taking. The examples referred to by Herring are different from Twitter in some aspects. Due to how conversations are presented the Twitter thread does not truly resemble a discussion forum thread, and with the openness of the platform, any tweet has the ability of getting many replies. But as the thread becomes deeper the audience shrinks, although it is possible that anyone can find a tweet in the conversation by using the search function. In addition, Twitter conversations stand out from other forms of interactions due to the rigid 140 character limit imposed on turn-taking. This encourages brief and to the point arguments. If the user disobeys the programmed rules and circumvents the character limit by replying to its own tweets, it affects the conversation. Other participants might reply to a part of the content and before the entire argument is posted.
In the introduction I argued that aspects such as deliberative democracy can be better studied with the conversation-based sample compared to the hashtag-oriented sample, but the analysis of conversational threads in Study V showed that conversations according to the ideals of deliberation (e.g. Fishkin, 2011; Graham, 2008) are rare. The study indicated that participants opened up for discussions but all too often a substantial discussion failed to materialise. This confirms results of previous studies that have questioned the democratic function of the platform (e.g. LaMarre & Suzuki-Lambrecht, 2013; Larsson & Moe, 2013; Larsson & Moe, 2012; Sæbø 2011; Yardi & boyd, 2010). The problematic feature of the popularity principle on social media platforms as identified by van Dijck (2013) counteracts one of the core conditions of a deliberative discussion by Graham (2008) on Twitter: equality. In Study V, it was clear that elite users were more often addressed, and in turn should be more visible. The more visible users attract more followers, and are then further emphasised by the platform. The idea that deliberative democracy is not reachable on Twitter seem to be well supported, both by previous research and Study V. Shifting the focus to how people communicate on a platform such as Twitter, as proposed by Eveland, Morey and Hutchens (2011), is a reasonable option which was taken for Study V.

As there were examples of threads with many tweets and many participants, it can be said that Twitter indeed enables discussions. However, problems regarding such benefits persist. One is related to the presentation of conversations which implies that any reader can see a replying tweet only if he or she follows both the sender and receiver. It is possible to find a tweet by using the search function but as few threaded tweets are hashtagged, the conversation is hard to find. Another problem is related to the velocity of conversations. The conversations in Study V were played out during a few hours or days, and were intense and probably difficult to keep up with considering the high frequency of tweets. It is unlikely that the participant has the whole conversation in mind when replying to a tweet. The situation is complicated by different rendering of the thread in the various devices and apps the participants are using. In my own usage of the Twitter app for Android phones and tablets and the iPad I have experienced that fewer tweets show up in hashtag searches and expanded conversations in these devices compared to performing the same searches and viewing the same conversations on the Twitter web site. To fully understand how conversations evolve on Twitter, research on how participants perceive the conversation with regards to the various devices is needed. Another reason for taking devices into account in the analyses is that
different tweeting behaviour across devices has been found in the literature (e.g. Murthy et al., 2015; Veenstra et al., 2014; Xu et al., 2014). Moreover, as argued by Tufekci (2014), the focus needs to be widened to include other platforms. Twitter is only a part of a social media ecology (van Dijck, 2013), and is hence only one piece of the puzzle of the daily lives of its users.

Let us now return to the DNA of Steve Jobs as discussed by Mayer-Schönberger and Cukier (2013) (see 3.5.3). Instead of taking a sample of the DNA, the entire DNA of Jobs was sequenced. This is an example of a complete, clean and stable data set, but still it is only a part of a human being. Most big data practices are much messier. If we connect this to the Twitter data set, the hashtagged tweets comprise a (biased) sample of a conversation. The conversational thread seems a very clean unit of analysis which is clearly demarcated. However, Study V showed examples of branches deviating from the initial topic of the thread. The idea that such deviating conversations can exist within a larger conversation on Twitter has been proposed by Magnani, Montesi and Rossi (2012). The December Agreement is a part of Swedish politics. The conversations around the issue of Swedish politics are parts of other conversations, most of them domestic, but some of them international. Some of the conversations contain other conversations about other topics. A conclusion relevant to boundaries is that it is more complex to define a conversation than using the technical definition of a thread of tweets connected through replies. The affordances of Twitter, the devices and the apps further complicate the issue of defining what a conversation is. With qualitative analysis, which takes both the context of the tweet in relation to the tweet it is a reply to, if any, and the metadata representing the device and app used by the participant, it is possible to zero in on these boundary issues somewhat.

Seemingly, the Twitter researcher needs to find new ways to talk about and research conversation, as it is a new type of communication. There are some general traits for computed mediated communication as well as some specific traits for the Twitter conversation to consider. These are that the involved participants do not meet in time or space, the conversations have the potential of involving huge amounts of participants, contributions are asymmetrical as some speak a lot while others speak very little, there is a breakdown of sequential reply and the turn-taking structure and finally there are problems of overview and ordering.
7.2 The filtering platform and its affordances

Following van Dijck (2013), this thesis takes the view that Twitter is not neutral. Twitter is steered by popularity as not all users are treated equally by the platform. Those who are already popular are emphasised and thus more visible than other users. The emphasis on the popular users and topics, the visibility of replies and the data that might be filtered out by the API all contribute to the non-neutrality. The idea that the API is not neutral is supported by the data, mainly through the method triangulation in section 6.7. Not all relevant data are captured and we do not know what data we do not get. Previously in this thesis I have outlined a number of relevant changes in the Twitter interface, its functionality and its API. Other changes are in the semantics, such as the shift from a star to a heart, which means that tweets are not bookmarked as before, but rather liked. This semantic shift makes it problematic to compare research on favourites with research on likes. The changes can be related to technical programmability, which is about how the platform affords its users to interact, but also to popularity. One can see that programmability and popularity are connected to each other. The platform affords the construction of hashtags. Eventually the community around the hashtags settles for one or a few that are preferred. The frequently used hashtag becomes a trending topic and attracts more users through increased visibility. Another aspect of popularity is that tweets from more popular users and tweets that become more popular through replies, likes and retweets are likely to be more visible. For a researcher, it is important to be aware of the affordances at the time of the study. Research results from when Twitter was a utility differ from research results from a popularity steered Twitter.

But it is not only about how the Twitter web site affords its users to interact. How conversations are visible in various devices affect how they evolve. The question of who is exposed to what can thus be directed to the device used when accessing Twitter. Regardless of if some filtering occurs in handheld devices compared to the web site, overviewing a conversation on a smaller screen is more difficult. Considering the discussion in the section above, this leads to the question of how the interface of the combination of the device and the app or web page affords the user to participate in a conversation. We would then need to view affordances in relation to the mediated action perspective (e.g. Kaptelinin & Nardi, 2012), devices (e.g. Murthy et al., 2015) and the interactional features of Twitter (e.g. Haustein et al., 2014; Holmberg et al., 2014; Lawrence et al., 2014; Zappavigna, 2011), and
also considering the aspects representation, technology and socio-culture (Fragoso, Rebs & Barth, 2012). We would also need to investigate if and when some filtering occurs when using the various devices and if this can be connected to the popularity principle.

Aspects such as social networking, homophily, consolidation and echo chambers are all highly dependent on the way Twitter presents conversations. Social information flows are dependent on explicit social network structures, and in the case of Twitter, the relevant network type is the followership network. Returning to Centola’s (2015) conclusions, a network that facilitates good information flows has moderate consolidation and moderate homophily, characteristics that the followership networks in Studies II and III seem to have. These followership networks all have distinct clusters that are connected with fairly wide bridges. In Study II these clusters were found to be fairly homophilous. This could be an explanation for the large extent of follow-on conversation, which requires a tightly knit followership network as well as some diversity regarding opinions. However, followership is not all that counts here, as some users might find tweets by hashtag searches, or make use of lists rather than followings. A relevant aspect here is the information diffusion process, which could be influenced by both the usage of lists and following other users. Again, who is exposed to what and how?

We can talk about an architecture of ad-hoc communities. This relates to the stream of tweets including the same hashtag as an ad-hoc public (Bruns & Burgess, 2012). At the broadest level, the ad-hoc community around a hashtag includes all the Twitter users making a hashtag search or tweeting with the hashtag. The hashtagged tweets are also visible in the timelines of all followers of the tweeter. When a tweet is replied to, something akin to an echo chamber is created by Twitter, as an effect of the way the platform presents conversations. Replies are visible in the timelines of those following both the sender and receiver, which makes the audience of such tweets narrower. Underlying is the followership graph, which seems to be polarised, and so the ad-hoc community created should then be more like echo chambers of like-minded, but the ad-hoc communities in Study V were diverse considering political positions. Retweets in contrast redistributes content to all of the sender’s followers, which results in larger audiences. But retweeting seems to be mainly performed within clusters of like-minded. For conversations to thrive, the Twittersphere needs to be tightly connected and include a diverse range of opinions. Retweets on the other hand can redistribute information between the denser networks through weak ties.
The much used social network analysis method is applicable in settings like these. In this thesis, the differences between networks with regards to prominence have been outlined with different parameters, metrics and communication type (Study IV). For example, those who are prominent in a retweet network are not likely to be prominent in reply-based networks. The weak ties that are important for information diffusion are less important for the development of conversations on the platform. Only analysing a retweet network in one setting does not give the answer of who is influential in a given Twittersphere. What is filtered forward and how is this further emphasised by the platform through popularity? This question is also related to sampling issues discussed in the next section.

7.3 Sampling, bias and completeness

For a number of reasons, researchers need to be aware of how they sample data from Twitter. I have previously introduced the concept of a convenient sample, which refers to collecting data from Twitter by either specifying a set of keywords or hashtags, a set of users or a set of geographical coordinates. Such a sample comprises a biased and fragmented representation of a conversation. Not only is the sample incomplete, but it also has a bias towards the parameters specified by the researcher. With the lack of reflections around what such a sample represents or not, it seems as many researchers accept that the collected data set is sufficient. The many examples of disagreements and the small share of tweets including a hashtag in the threads in Study V indicate that the opinions expressed in hashtagged tweets are unlikely to be representative of the entire conversations. Even if the convenient sample includes a large number of tweets, N is not “all”.

Moreover, sophisticated sampling techniques are needed because of sparsity and noise. While big data evangelists argue that increased amounts of data allows clear and certain signals, González-Bailón (2013) has suggested a need to remove noise through use of theory. It is not necessary so that more data points increase certainty. One argument for such a position is given by Schneier (2015) who wrote that terrorism in relation to big data is “a needle-in-a-haystack problem, and throwing more hay on the pile doesn’t make that problem any easier”. In relation to this, the search for completeness has had a major purpose in finding those units of analysis that are meaningful to the researcher. This process should be guided by the research questions.
Twitter research has either failed in its attempt to analyse conversations or lacked the ambition of collecting data with such a unit of study in mind. It is clarifying to compare with a similar online phenomenon. If the data source was a discussion forum, would a researcher collect and analyse only those entries made by a set of users or those including one of a given set of keywords? This question might seem absurd, but this is more or less what Twitter researchers have done so far. Interestingly enough, the hashtag-based Twitter study has emerged as a strong tradition justified implicitly or explicitly by unclear notions of this being the most natural practice. I have repeatedly throughout this thesis argued against this justification.

Even in attempts to systematically capture something broader than the hashtag, the researchers of the political Twittersphere have fairly often accepted a definition of conversation which includes only parts of the conversations (section 2.3). This can, of course, be attributed as an effect of the API, its restrictions and regulations. In this thesis I have made pioneering work relating to more complete collections of conversations. Starting with a more appropriate definition of a conversation as communication that stretches beyond the hashtags, the methodological approach is based on the collection of hashtagged tweets and expanded by following the most active participants. Study IV showed the large differences between a set based on hashtags only and a set with follow-on conversation. Such a set cannot be collected by using just one parameter of the API endpoint. Streaming tweets about a topic and the tweets sent to or by the participants in the conversations is the solution to the data collection problem. Looking up each collected tweet for replies over and over again is not very practical. Filtering the user-based stream is better, although not optimal. The combination of streaming users and keywords proved to be an adequate solution for the chosen context.

One of the most troubling problems regarding data collection is that the material collected tends to be biased towards the opinions of elite users. This bias is created because of the critical decision of who to focus the data collection and/or analysis on. Many studies have decided on 1,000 or fewer users (e.g. Zimmer & Proferes, 2014) and several studies have found domination in activity by a similar number of users. The studies of this thesis has also focused on around 1,000 users and found similar results, although with small overlaps between the three data sets. One would need to ask whether the dominance by around 1,000 users is an effect of bias somewhere in the process. If Twitter emphasises the more popular users and tweets, these are more visible. This initiates a vicious cycle. More visible users and
tweets gain even more visibility. Such patterns ripple through the API to the researcher, who has to make decisions of what to sample as it is not possible to stream all users within a topic. It has been argued that the traditional tool of quantitative social science, the random sample, may not be possible through the Twitter API (e.g. boyd & Crawford, 2012; González-Bailón, 2013). Therefore, it makes little sense to randomly sample participants in the not random sample to follow. The alternatives, then, are to follow the popularity model in some way, either focusing on the most active or the most visible users, or the most recent users regardless of how active or visible they are.

The reasonable aim is not completeness but rather to decrease the effects of incompleteness. Each topic has different usage statistics and different number of participants. In 6.7 a test of the composite method with different parameters was made. The chosen hashtags #svpol (Swedish politics) and #auspol (Australian politics) were used alongside the 1,000 and 5,000 most active participants in the conversations. The clustering around #auspol was much larger than the one around #svpol, but it was still possible to collect follow-on conversation (although the completeness of the threads themselves was not assessed). A good point of departure is to identify a typical number of users per time unit. If 5,000 participants are active within 24 hours then our sliding window could be engineered so that it picks the 5,000 participants during the past 24 hours. The incompleteness issue is more critical for high velocity conversations with far more than 5,000 dominating participants. In Data set 3, 5,000 users had participated within 24 hours, but this was a special event attracting more participants and the number quickly faded. Naturally, larger conversations attract 5,000 users sooner. This implies that it is not likely that we can get a good grasp on the size of the follow-on conversation in every possible case. The higher the velocity of the conversations is, the more likely the 1% streaming cap is reached. However, while applying the composite method on such conversations will likely result in incomplete threads, these fragments can still be used as candidates to more substantive discussions. If the method of analysis is qualitative, it is always possible to navigate to the Twitter page of any tweet in a conversation and view the entire thread.

The enforced limitation of following 5,000 users implies that there are unknown unknowns regarding tweets related to the conversation. These are the interactions between users outside the set of 5,000 users and not including any of the tracked hashtags. Although the less active users are not utilising mentions as much as other users (e.g. Bruns & Highfield, 2013; Bruns & Stieglitz, 2013b; Study II), it is not
unlikely for an interaction between two of these users to occur in a thread, and thus the thread as the researcher views it is broken. The more people involved and the higher the velocity, the higher the likelihood that threads are broken due to interactions among the users outside of the specified user filter. The bias towards the most active (or visible) users grows with the number of participants and the velocity.

Researchers must try to reduce their own bias in data collection, otherwise, with a convenient sample from a totally unknown full set, what value has the research? Just because many researchers accept the hashtag-based sample, albeit with conceding its limitations, does not mean one should always settle for such a sample. The thesis project has had a long standing aim for completeness. In Study II, it was concluded that the incomplete mention network might skew the results. In Study IV, it was confirmed that different filtering options gave different networks. I have here shown how different a hashtag-based sample can be compared to a sample with follow-on conversation, which questions findings in hashtag or keyword-based studies without follow-on conversations. Obviously, researchers who find it relevant to study conversations or discussions cannot do so with such an incomplete set.

7.4 What insights can be gained?

The studies in this thesis indicate that Twitter is an accessible source for attaining insight into public opinion. The seemingly straightforward access to data is a huge advantage. However, it is easy for researchers to be seduced into believing that the quality of the empirical material is better than it really is. Earlier research has mostly discussed the drawbacks such as shortness of messages and the low representativeness of the population. There have also been discussions about ethical issues, particularly concerning politically oriented discussions. The current thesis has posed questions regarding what insights can be gained in collection of data that becomes situated in the gray area between the systematically collected sample and complete data set. It has also been argued that a fundamental problem with the data collected within such a gray area is that the context of the various tweets is lost. This makes sophisticated scrutiny of Twitter activities very difficult. Furthermore, the thesis presents a strategy for navigating out of this gray area. This is partly made through the usage of conversation as the analytical unit and partly
through the introduction of the composite data collection method. This, crucially, involves collection of follow-on conversations and so having access to the context of each tweet. With such a method, it is possible to attain insights into reactions to events and tweets. The low representativeness implies that Twitter cannot, and should not, replace other sources of public opinion. It can, however, be used as complement. For example, what do some people think about the December Agreement? Why do parties lose or gain voters?

The results of the studies have focused on characteristics of the political Twittersphere. Study II showed that there are polarising tendencies in regards to the followings and retweets, but that non like-minded people communicate with each other in the form of mentions. Polarisation is more visible in retweet networks (e.g. Conover et al., 2011; Study II) and followership networks (Kim & Park, 2012; Study II). Hidden behind a mention network created from a hashtagged data set are complex and simple interactions of which few are consensus like, which were found in the thread analysis in Study V. Echo chambers not visible in these are visible in threads. Barberá et al. (2015) found that echo chambers are more prominent in political topics, but also that while Twitter users seem to prefer to retweet like-minded they are exposed to ideologically-wise diverse information. Exposure to diverse information is evidently true for the most engaged participants, those who participate through replies and mentions.

Engagement is a crucial aspect in which the three network types in Study II differ. The more engagement required the less dense is the network. The followership network was the densest, followed by the retweet network. As engaging in conversation requires more effort, it comes as little surprise that fewer participants are included in the mention network. A mention network in a setting like this may involve only the most engaged interested people. In Larsson and Moe (2012) it was found that communication was more reciprocal at the “high end”. In Study II, focusing on a larger high end, there was no tendency of mention polarisation but such tendencies might be visible if looking at less active/visible users. What if moderate homophily and consolidation are characteristics of the group of 1,000 most prominent users and not among the less prominent users? This question can only be addressed by looking beyond this group of 1,000 participants. It might also be so that the combination of moderate homophily and moderate consolidation is a unique characteristic for this setting. Do other settings, contexts or topics look different? More research on conversations in other settings, contexts and on other topics is needed.
Study III showed that much of the activity of and around these specific elite users is centred on political issues. One interesting finding here is that apart from the national Twittersphere, an international sphere was also involved. International topics were tweeted about and these were then filtered forward and amplified through retweets. On a methodological note, the study highlighted a need to be aware of how different communication types impact on the result. For example, when retweets were included, some hashtags were very prominent whereas they were not when retweets were excluded.

There is clearly a need for separating the relationships and communication types from each other in an analysis. If we take the network comparisons in Study IV as example, it is more relevant to focus on prominence with regards to a given aspect, than to aim for a summary across all communication types. Those who are engaging in conversations are perhaps not the ones who are effective spreaders of information. This might have to do with network topology. A tightly knit followership network seems to be a prerequisite for the development of conversational threads whereas weak ties are important for information diffusion. As Twitter displays replies in the timelines of followers of both the sender and receiver, those who have many followers are more likely to be involved in conversations. Contrary to this, a user does not need to have many followers to be able to spread a message. If a follower with many followers retweet the message, it is exposed to many users in just two steps.

Adding to this, a followership network is denser than a retweet network, which in turn is denser than a mention or reply network. This means that if these types are combined into one network, this network will be biased towards followership. Any user of Twitter interacts only with a small subset of all its friends through retweets, mentions and replies. Retweets require less effort and is thus more likely to be the preferred way of interacting for many users. This aspect of information behaviour can be seen in Study II. The various user groups were found to behave in different ways, with the least active being less conversational but more often retweeting. A similar behaviour was found among the highly active right wing-users. Slicing data sets as in Study II is necessary to understand behaviour. The results from that analysis question findings from studies where all interactions (e.g. replies, mentions, retweets) and relationships (followership) are analysed in one network.

Another challenge regards measures of influence or prominence, illustrated by the findings in Study IV. The differences in top 100 and 1,000 users in networks
according to four different metrics suggest that both data collection method and chosen metric impact largely on who can be identified as prominent. This means that using a score based on activity, visibility and spreadability as in Study II can be misleading, especially since a mention or a retweet is not necessarily positive (e.g. Tufekci, 2014). The combined metric including visibility through activity, being mentioned or being retweeted is arguably able to select those users that Twitter visitors perceive as prominent, although it does not consider whether such prominence is positive or negative.

Earlier research has shown that activity tend to increase during events of various kinds (e.g. Bruns & Highfield, 2013; Larsson & Moe, 2012; 2013; Jungherr & Jürgens, 2014). During the protests studied by Jungherr and Jürgens (2014), activity was more centred on spreading the news rather than conversation. Similar findings were made in Studies III and V, in the former most notably during Election Day and in the latter following the December Agreement and the sudden attack on Charlie Hebdo. Beyond this confirmation of earlier research, I also attempted to investigate the character of event-related activity increase. Judging by my results, it seems as more people are tweeting during events as the 10% most active users are fairly stable as user groups (Study V). When overall volume is high, and especially the least active 90% group is at its most active, the activity seems to be focused on informing and relaying information rather than on conversation. Both sudden and scheduled events seem to have a similar effect on Twitter usage in that the least active users become more active, and the share of mentions decreases. In conjunction with such events the least active are overrepresented comparing to the general Twitter activity. But these observations are also bound to certain topics. It might be so that the most active participants within a given topic are not very active elsewhere, and that the least active participants within this topic have a generalist interest. Following the activities of the least active user of a given hashtag with similar methods as in Study III would shed light on these issues.

We currently lack knowledge on the dynamics between the most and the least active within a given hashtag. Seemingly, different types of topics allow for different kinds of conversations. For some hashtags, the most active 1% or 10% participants are not dominating. Bruns and Stieglitz (2013a) found that the hashtags #stopkony and #royalwedding had separate distribution patterns when compared with most other hashtags. In these cases, the least active 90% accounted for larger shares of the tweets. Some hashtags seem to stimulate to a rush of
activity immediately after a sudden event and, as well, seem to attract more users. Some hashtags are more often retweeted, specifically observed in Study III when Nelson Mandela passed away. Hence, there are activities on Twitter that are difficult to study with the common access to the API if follow-on conversation is desired. Not only because of difficulties in predicting when a non-scheduled event is about to happen, but also because of the sheer volume of tweets from a large number of users.

The domination of a small number of users, appearing prominently in this thesis, has been identified in other settings as well (e.g. Barberá, & Rivero, 2015; Bruns & Highfield, 2013; Tumasjan et al., 2011). Generally, it can be said that a minority of around 1,000 users dominate each data set in this thesis, but the three sets of users have small overlaps. One should not overestimate the impact of Twitter, but is it possible for any follower or participant of a conversation to appreciate whether it is dominated by 1,000, 10,000 or 100,000 participants? Maybe it is reasonable that the number of participants is perceived to be far higher than what it actually is. Analysing the conversations with suitable methods is the only way to appreciate how many are participating, how many are vocal, and so on. Even so, we cannot be sure that we have captured all the relevant data. We do, for example, not know how many Twitter visitors are inactive followers of the conversations. An indication of that many other users interact with the most prominent users through replies and retweets was given in Study III. The activity of a small group of elite users led to activity among a significantly larger audience. What this means in relation to opinion leadership in this setting, both regarding active and inactive followers, remains to be studied.

In Study II there were few examples of journalists and politicians among the most active and visible users. Study V showed that politicians are often addressed by citizens and other user groups, but citizens dominate the activity. Results so far are somewhat different between published studies. The general usage of #svpol in Study II was dominated by citizens, as was the Canadian political Twittersphere around #cdnpoli (Small, 2011). Studies with a focus on a smaller set of users have found domination by elite users (e.g. Ausserhofer & Maireder, 2013; Larsson & Moe, 2012). Some studies have found activity domination by citizens but that elite users are given more attention (e.g. D’heer & Verdegem, 2014; Larsson & Moe, 2013). The different indications shown in different studies in this setting might depend more on the specific case than data collection method. Following a controversy such as the December Agreement, politicians need to act and their
actions are bound to generate much attention. However, with a large percentage of the replying tweets not including any of the 13 tracked hashtags, all users frequently replied-to are much more prominent in a network based on mentions from this more complete set than from the same set in which only hashtagged mentions would be included.

In closing this section, we can identify a couple of asymmetrical Twitter effects. Firstly, a minority accounts for a large share of tweets. This minority of around 1,000 participants is not easily overviewed by a follower of or participant in the conversation and with the amount of content they produce, this minority might very well be perceived as including more users than it actually does. Secondly, attention is given to a small set of users, often elite users such as politicians, celebrities, journalists and early adopters. Those who are given attention are also more emphasised through popularity, and this is then reinforced through the Matthew effect. Thirdly, the vocal minority does not always comprise the same users as the most visible minority. This means that we need to study activity (tweets posted) and visibility (tweets received) separately. We also need to strive for more complete data through inclusion of follow-on conversations, as only a subset of all replies is captured with convenience sampling. A reply network of a hashtag-based is skewed because of the non-hashtagged replies not captured. Fourthly, replying narrows the audience whereas retweeting broadens the audience. This means that a reply-based network is less likely to develop unless the underlying followership graph is dense, while retweeting can spread content across different parts of the Twitter network through weak ties.

7.5 Which data collection method and when?

I have focused on three different methods for collecting relevant Twitter data, whereof one is introduced in this thesis. The first two are referred to as the hashtag-based and the user-based, and the developed method as the composite. The simplified use of only hashtag-based or user-based approach has here been labelled as the convenient sample. Studies making use of such a sample do not study full conversations but rather some kind of basic communication, where people are on the one hand talking about or sometimes to other people by mentioning them, as substitutes for other users’ names (e.g. Bruns & Highfield, 2013). Typically, such a study can reveal topics introduced, the immediate reaction
to a given event and who wants to talk to (or about) whom. Adding follow-on conversation to this, we can analyse the reaction to these tweets, if someone talked to (or about) replies, and conversational threads emerging from the hashtagged tweets. It is also possible to analyse invocation, which is the invitation to another user by a mention, and this user’s reaction to this.

Hashtag and keyword-based approaches have their merits, for example, in information diffusion studies. As retweets are often copies of the tweet retweeted, a hashtag or keyword-based approach is generally enough for studies of information diffusion. Such studies would benefit from a followership (and/or a list-based) graph to provide context for the information processes, for the purposes of investigating how large proportion of a user’s followers retweet the user’s tweets. Other settings where the hashtag and keyword-based methods are suitable are in crisis communication where people might be more inclined to use hashtags even in replies. Studies of health issues (e.g. Ghosh & Guha, 2013), pandemics (e.g. Chew & Eysenbach, 2010; Kostkova, Szomszor & St. Louis, 2014) or psychological issues (e.g. Jashinsky et al., 2014) should perhaps not aim for follow-on conversation as it is the content in relation to the keyword that is of interest. On the other hand, someone might tweet with the purpose of asking for guidance, which implies the tweeter is interested in the follow-on conversation. A similar discussion can be applied to protest-related studies (e.g. Croeser & Highfield, 2014). In this case, the authors acknowledge the incompleteness but argue that it is not important for the study.

User oriented method is suitable when the research question concerns what type of content is amplified through retweets. A study similar to Study III could be made in different context from an agenda setting perspective to analyse who the agenda setters are in a given Twittersphere and what type of content is amplified through retweets. Following a hashtag-based (or perhaps conversation-based) initial data collection, different strata of users can be identified and followed over time. Analyses of interest could be if there are differences between the most active and the least active users with regards to agenda setting and amplifying. The user-based approach is also suitable for investigating how a set of users, for example politicians or journalists, interact with other users.

The composite data collection method is suitable when conversation dynamics are of interest. On a general level insights into how people communicate with each other on the platform can be gained. It is also a method that can be employed if
deliberative democracy or similar aspects are to be studied seriously. Another important reason for collecting conversations is that a more complete data set can be collected, and all replying tweets can be analysed in the context of the tweet replied to. The composite method yields a richer account of what is studied. This method is also to be preferred for archiving purposes, if for example the aim is to use historical Twitter data for analysing how its users communicated in relation to various events or topics. But the composite method is not always the best possible and it is not always feasible. For conversations to evolve it is likely that the setting would have to involve a diverse span of opinions and being fairly tight connected through followings, as well as not involving too many users. There probably also needs to be some level of controversy. The Swedish setting and the December Agreement controversy is a good example when these criteria meet.

While this thesis has made use of examples within one political Twittersphere, it is relevant to discuss other areas. Nelhans and I tested the composite data collection method on scholarly publications identified through digital object identifiers (DOI) by searching for the string “dx.doi.org” (Nelhans & Lorentzen, 2016). Potentially, such an approach could reveal how people react to research. However, we found few examples of conversation. Health-related topics could on the one hand invite a more information sharing behaviour and on the other a more conversational behaviour. Users might tweet about symptoms as a kind of self-reporting (Kostkova, Szomszor & St. Louis, 2014) but they could also take to Twitter for guidance, and if so, follow-on conversation would be relevant to capture. Keeping track on or predicting pandemics and similar (e.g. Chew & Eysenbach, 2010; Gesualdo et al., 2013; Kostkova, Szomszor & St. Louis, 2014) could be done by tracking hashtags or keywords and apply co-word analysis, and make use of geo-location data. Such data are however difficult to make use of as only a small share of tweets contains these (e.g. Gesualdo et al., 2013).

Crisis communication is probably more inclined to information sharing rather than conversations. However, to be able to learn how one can make information systems more effective for crisis communication we need to learn how we communicate during crisis. For this reason, follow-on conversation could be of interest to collect and analyse. In a study of Twitter conversations in relation to three campus shootings in the US, Heverin and Zach (2012) found information sharing behaviour as well as information seeking, information negotiation and understanding the “why” among other types of behaviour. Their hashtag-based study is an example of a potential for extension with the composite data collection
method. A problem is that the number of users tweeting during crisis is potentially high which makes follow-on conversation difficult to track.

Protests and activism are other areas which might involve many users. Veenstra et al. (2014) identified almost 90,000 tweeting users during three weeks of tracking #wiunion (Wisconsin labour protests). The very large share of retweets (62.5\%) indicates that usage was more about informing and relaying information rather than having conversations. It is reasonable to believe that most similar types of communication activities are more focused on informing and hence capturing follow-on conversation makes little sense.

Journalism and citizen journalism could be interesting aspects to study. How do people react to news reported on Twitter by journalists compared to news by citizens? Among other important findings, Stuart Allan (2014) identified a clash between the raw, unedited citizen images and the mass media images, which were edited according to profession ethics. This is just one example of how citizen journalism differs from professional journalism. Here, the follow-on conversation would be very relevant to capture, although it might be difficult to identify the relevant tweets. A possible approach is to search for them in a collected data set on a given topic. Another solution is to use a similar approach as Nelhans and Lorentzen (2016), where DOI URLs were streamed. Here, the streaming filter could search for references to images instead.

I have previously touched upon the issue related to opinion mining and sentiment analysis. These approaches are typically applied to an entity or topic, for the purpose of analysing people’s opinions of it. In cases where follow-on conversation is present it should also be included to eliminate the hashtag or keyword bias. One positive hashtagged tweet might receive many negative replies. Here, one could test the composite approach and compare results between sentiment scores of hashtagged tweets with those of replying tweets not including hashtags.

These are just a few examples of other areas, settings and contexts. The modes of communication are still new and continuously evolving, partly because of changes in affordances of the platform, and these modes can differ between domains. The possibilities are many when it comes to Twitter interactions. It is important to not handle the material too crudely; both quantitative and qualitative methods are needed when looking at specific characteristics within the studied domain.
8 Conclusions

This thesis has made a major contribution through the development of the composite method for capturing conversations on Twitter. The discovery of a significant follow-on conversation in the studied context questions findings made from data sets collected with traditional methods. It is expected that social media researchers become more conscious about data quality, sampling issues and completeness. Reflecting about what data are not collected is important, because otherwise sound conclusions can hardly be drawn from the analyses.

Even though Twitter is far from a new platform, research of activity on the platform can still improve significantly, method-wise. The use of trivial data collection methods has been common which results in avoidance of complicated and multifaceted issues. The Twitter research to date has focused on simplifications of the interactions as a conversation all too often has been reduced to what is represented by only the hashtagged posts. Theory for collecting, analysing and explaining data is still very important, and this must be put in relation to the issue of incompleteness. What can really be said about results from a study of hashtagged tweets, especially when the follow-on conversation is large or unknown? Despite many scholarly investigations of Twitter relationships and interactions, we still do not know much about how its users communicate on the platform. In this thesis, a serious attempt to study conversations has been made, but there is still much to learn through studies in other contexts and domains.

From the perspectives of information science and informetrics, this thesis provides a method to collect more complete data in order to study information flows and behaviour in a defined topical realm, and a means to make measurements of informational aspects more accurate. Any mention or reply network created from a set of hashtagged tweets will exclude those users who do not use the hashtag, even though they might reply to hashtagged tweets. Any such network is always restricted to hashtag users. The hashtagged reply-based network is also likely to reduce the prominence of those users who receive much attention in the form of replies. If non-hashtagged replies are excluded, the reply-based network is very likely to be misleading. Measurement of informational aspects such as activity, visibility and distribution of tweet types are not accurate if follow-on conversation is not collected. Other metrics that are likely to be inaccurate are the size and density of the mention and reply networks. A followership network derived from
hashtag users might also be inaccurate. Users with large follower numbers do not need to use hashtags for visibility; their tweets will reach a large audience anyway.

Alternatives to basic statistical approaches with regards to sampling must be considered as Twitter users are not representative of a voting population. The most active or visible users comprise a too small and biased sample. We cannot put a value on a variable such as the probability of voting for a given party and then draw a conclusion based on Twitter data how an election will end, but we can learn about the characteristics of the conversations. The conversations are more complex interactions than the hashtag-based sample indicates, and few replying tweets include hashtags. The conversations provide insights into opinions and reactions to events, statements and tweets.

These studies have shown that the conversations are dominated by a smaller group of about 1,000 participants that could be considered an elite, and when looking beyond this group, another 150,000 users interact with this group. Similarly, other studies cited in the thesis have found a domination of a few thousand users, but also that this differs between different contexts and domains. The studies have also shown that the quality of the conversations probably does not correspond to our ideals of deliberation. Other studies cited here have shown similar results, although with very limited data, in that only partial conversations have been studied.

8.1 Not reaching completeness

In the end, the pursuit of collecting complete data took this project to the analysis of conversational threads as a limited entity of analysis, but completeness was not reached. The data collection method tests in section 6.7 confirm this, as there were tweets captured using the search API not captured by the streaming API, and vice versa, when the streaming API made use of the composite method. When tweets matching hashtags and tweets sent by and to a set of users are streamed simultaneously, a fraction of the hashtagged tweets are not collected. By utilising the API with the common access level, not all data can be collected, but at least in the case of Study V the data set is still complete enough for analysing the conversations on Twitter. The characteristics of the political Twittersphere studied here are in some aspects probably unique, for example the underlying followership graph which seems to be structured in a way that supports interaction. The range of opinions in the Twittersphere is seemingly diverse enough for discussions to
materialise. The size of the follow-on conversation might be larger than in most contexts, but at present, there are no studies to compare with, apart from the larger Twittersphere around #auspol, which had less follow-on conversation. Nevertheless, knowing the size of the follow-on conversation, or rather assess it, is crucial for knowing whether the hashtag-based set is a reasonable alternative for analysis.

The conversations studied here were played out in a setting which does not fulfil all the characteristics of big data. The volume is relatively low, as is the number of participants and size of the audience (5% of the population). Further research is needed to establish how conversations are evolving given different variables. It is probably so that conversations with too many participants and of too much volume cannot be captured with the composite method, and hence firehose access would be needed in such cases. Some questions arise with regards when all the four V:s of big data (Ning et al., 2015) are fulfilled to a large extent, that is, there is a large volume of tweets including the hashtag that are posted by a large number of users (variety) at a high velocity, and the then much used hashtag attracting spam (veracity). To what extent do threads evolve? Is it possible to collect a sufficient part of the conversations for analysis (i.e. too many users, too much volume exceeding the streaming cap)? Further tests are needed to get an understanding of when there is too much volume and variety, and too high velocity for conversations to evolve.

In working with big and complex data sets, we might find that it is just as important to understand what we are capturing as well as what we are not capturing. Otherwise, we will be ignorant of the kind of bias involved in constructing a sample. And we must, still, develop methods of sampling. Despite claims of “end of sampling”, methodological discussions on this issue are perhaps even more important than earlier.

The studies have shown that the various methods of data collection and data analysis yield different results. They, therefore, highlight different aspects of the conversations. A key conclusion of this thesis is that Twitter researchers need to be guided theory-wise in their sampling and be aware of which questions can be asked with the chosen data collection methods. Another conclusion is that as long as Twitter has this level and type of data access, researchers need to be creative in utilising the API to be able to answer the questions researchers find interesting rather than accepting the limitations of data-driven research of hashtags or users.
As the platform continues to evolve, researchers need to be clear on what the affordances are to the users at the time of data collection, and how these affect results. This is easier said than done, as the ways in which the user accesses the platform need to be considered. As suggested in Study V, the conversation thread might look very different from one device to another. The entire conversation might not be visible at the time of a reply when using a mobile phone.

The most important thing to learn from this thesis is that researchers must be aware of the limitations when sampling data from a commercial platform. We know that Twitter is not neutral and that our sample is biased but not how biased it is and not (completely) how or to what extent this affects our results. It is important to not uncritically let Twitter be the cheap opinion poll of tomorrow, and with the limitations in regards of access, representativeness and equality, it is important to not let Twitter be the centre of the public political discourse.

8.2 Limitations

This thesis is limited to Swedish conversations, which might be of less volume than many other interesting conversations. The methods for collecting conversations have been tested elsewhere too, for example in altmetrics (Nelhans & Lorentzen, 2016), in which the discussions were very short regarding number of tweets compared to the threads in Study V.

Any analysis of a demarcated space such as the political Twittersphere around #svpol will always suffer from a context-dependent bias. In this case, there are not many alternative hashtags, although it is difficult to know what data are not collected. Study III found that the elite users of #svpol were heavy users of that hashtag when all of their available activity was collected. With the limitations acknowledged, the focus of the thesis was put on the activities around the elite users of the Twittersphere. What happens outside of it is beyond the scope of the thesis.

In a similar way as some users are emphasised with a particular set of network filtering options, a certain definition of the context affect the results too. And the context is always embedded in a larger context. A hashtagged conversation on Twitter is not separated from other conversations on Twitter, and Twitter itself is a part of a larger eco-system of platforms. These platforms are intertwined in the
everyday activities of social media users. There are also practices that fall outside of the study of conversation, such as likes, readers talking about Twitter activity in other contexts or journalists writing about Twitter activity. The dynamics around the platform are difficult to capture, but nevertheless, these dynamics need to be researched and discussed.

8.3 The next step

This thesis has discussed methods for analysing Twitter conversations and exemplified with political discussions from an information science perspective. A natural next step would be for political or social science researchers to utilise these methods to collect and analyse the conversations from their own perspectives. With the methods developed in Study IV we can finally ask questions such as: What is the quality of the conversations? Do different opinions meet or is it all about echoing opinions of like-minded? Is it a discussion that comes about, an exchange of viewpoints that could be qualified as a part of a discursive democracy? Study V showed indications of citizens voicing their opinions, but in cases where many tweets were published during a short time window it is difficult for politicians to read and reply to a satisfying number of tweets. But by collecting tweets in the ways presented here gives an opportunity to summarise feedback from citizens. Also, with Study II showing examples of cross-boundary retweeting, it would be of interest to investigate what is made with such a retweet.

While there were several examples of conversation where citizens could voice their opinions in Study V, the study also showed examples of when substantive discussions failed to materialise. Indeed, Twitter as a mediator does affect the possibilities of conversations somewhat, as it makes use of algorithms to make some users and some content more visible than others. Also, the limited visibility of replies restricts the conversations to echo chambers; they can only grow to substantial conversations if there are enough (active) followers of the two users involved in a replying tweet. The interesting thing to note here is that the clustering around the #svpol (and the other related hashtags) actually include participants from the whole political spectrum.

The methods for collecting conversations need to be evaluated by testing them in different settings and with different parameters. In Study IV, it was concluded that for conversations with high volume and velocity, the method might not be suitable.
But how do such conversations evolve? This question might only be possible to address through access to the firehose. Other, similar investigations as in Study V in different settings, for example during non-controversies or elections, or in other topics, would help in painting a less fragmented picture of how conversations evolve. Moreover, when studying conversation it is also relevant to complement with other Twitter data, such as lists and followings. Data such as avatars, links and other information related to the profiles could be used to learn more about who is participating. The thread analysis makes it possible to find out if opinions change following a discussion, but such changes could not be discerned in Study V. However, a longitudinal study is better suited for this.

Software issues are also important to study. Each piece of software comes with its own limitations, and a comparison of different software applications and a method working closer to the data (directly towards the API) is highly relevant. Finally, the researcher is certainly constructing the conversations in a way that is not apparent to the participants. Arguably, the researcher has an overview of a large collection of connected tweets that no individual participant can see in its entirety. In this sense, the Twitter researcher is allowed a privileged position to a conversation in which he or she is not a participant. Therefore, insights into how the participants perceive the conversation are needed to understand this type of communication on Twitter.
References


Weinberger, D. (2011). Too big to know: rethinking knowledge now that the facts aren’t the facts, experts are everywhere, and the smartest person in the room is the room. New York: Basic Books.


Appendix

Preprocessing Data set 1

Coding of participants

Participants were coded primarily from their profile descriptions. If no description was provided, the name of the participant was checked to identify if the user represented a politician, party, mass media outlet, or other public instance or person. When the participant was coded, its username, name, location and other identification fields were removed and replaced with a numerical ID between 1 and 916. The user ID was kept for converting tweets to graphs and statistics (see below). When this process was finished, the user ID was removed.

Converting tweets to mention graph

All mentions by and of prominent users in the data set were used to create a network graph using these steps:

1. Find all tweets in set T including a username but not including “RT @” posted by a prominent user.
2. Extract all usernames in the tweet and look up user IDs in T.
   a. If user ID exists and is in set of prominent users P:
      i. Pick up coded IDs for tweeter and mentioned actors, write to file.
Converting tweets to retweet graph

Retweets by and of prominent users in the data sets were used to create a network graph as follows:

1. Find all tweets in set T matching pattern “RT[:]* [\[]\]{}{1}{a-zA-Z0-9_]+(.)*” posted by a prominent user.
2. Extract first username after RT and look up user ID in T.
   a. If user ID exists and is in set of prominent users P:
      i. Pick up coded IDs for tweeter and retweeter, write to file.

Converting tweets to descriptive statistics

For this end, seven categories were used: 1) the most active 1%, 2) the next active 9%, 3) the least active 90%, 4) all 916 prominent users, 5) users coded as L, 6) users coded as CR, and 7) users coded as RW. Step 1 below looked up users in different tables depending on category.

1. Find all tweets in set T posted by a categorised user.
2. Count all tweets per category matching pattern “RT[:]* [\[]\]{}{1}{a-zA-Z0-9_]+(.)*”.
3. Count all tweets per category including a username and not matching pattern in 2.
4. Count all tweets per category not matching patterns in 2 or 3.
5. Count all tweets per category including “http://” or “https://”.

Finally, the user IDs captured by querying for lists of friends were converted to the coded IDs. All relationships outside this group of 916 actors were removed. When these steps were performed, all Twitter user IDs were removed from the data set. The user data file comprised the following information: coded ID, political block (if any), note (politician, mass media, celebrity, or other).

\[13\] This search pattern is called a regular expression.
Preprocessing Data set 2

Two sets were created from the initial data set, one excluding all retweets and one including all tweets.

Converting tweets to descriptive statistics

For this end, two categories were used: 1) all users and 2) the followed users.

1. Find all tweets in the full set posted by a categorised user.
2. Count all tweets per category matching pattern “RT[:]* [@]{1}[a-zA-Z0-9_]+(.)*”.
3. Count all tweets per category including a username and not matching pattern in 2.
4. Count all tweets per category not matching patterns in 2 or 3.

Extracting hashtags

All content not matching the pattern # followed by a string of alphanumerical characters was removed from the tweets. In the set without retweets, co-occurrences of hashtags were identified using Table 2 Net (Sciences-Po Médialab, 2016) which returns a network file that can be imported into Gephi. For both sets, the usage counts per day and per hashtag were counted as the number of tweets in which a given hashtag was included and the number of users of that hashtag.
Preprocessing Data set 3

Coding of participants

Participants were coded from their profile descriptions and the content they shared through tweets. If no description was provided, the name of the participant was checked to identify if the user represented a politician, party, mass media outlet, or other public instance or person. When the participant was coded, its username and name were encrypted using an irreversible algorithm. Location and other identification fields were removed and replaced with a numerical ID. The numerical IDs then replaced the user IDs in the thread table. The final user data file comprised the following data: numerical ID, encrypted username, encrypted name, language code, user type (political, mass media, celebrity, or citizen), and political stance (if any).

Building threads

The reply metadata field was used to create threads from the collected tweets. Thread ID, tweet ID, reply ID, user IDs of tweeting and replied to user and timestamp of tweet were stored in one table. The texts of the tweets were stored with their IDs in another table. Tweet IDs were then replaced with new, numerical IDs. In order to completely cut the connection with the raw data, the tweet texts in the raw data were removed. A backup of the raw data was kept until both articles were accepted.

Finalising anonymisation

The user IDs and tweet IDs were removed as soon as this process was finished. This means that during one stage of the preprocessing it was possible for me to connect a political viewpoint to a physical person (or organisation) in those cases it was clear which person (or organisation) the account was tied to. When the user coding was complete, all identifiers were removed except for name and username, which were encrypted. This encryption was also applied to mentions in the tweets to enable analysis of invitations and to see who a reply was sent to. Some kind of
identifier is needed to be able to analyse the conversation dynamics. This encryption entailed that it was not possible for me to connect the tweet, or any user mentioned, to the person or organisation authoring or mentioned in the tweet.

**Converting tweets to descriptive statistics**

Two sets (A and B) were created from the raw data; one set including all captured tweets (B) during the two studied weeks and one set matching either of the hashtags (A) used for data collection. As for Data set 1, the participants were divided into usage categories based on activity: 1) the most active 1%, 2) the next active 9% and 3) the least active 90%. These steps were then performed:

1. Find all tweets in set T posted by a categorised user.
2. Count all tweets per category matching pattern “RT[:]*[@]{1}[a-zA-Z0-9_]+(.)*”.
3. Count all tweets per category including a username and not matching pattern in 2.
4. Count all tweets per category including a tweet ID in the reply metadata field.

**Creating networks from the data sets**

Mention and retweet networks were created as for Data set 1. In Study IV, we compared Set A mentions, Set B mentions, Set B retweets (which included slightly more tweets than Set A retweets) and Set B reply (including only mentions being replies).
Part IV

Studies I-V

Due to copyright reasons the articles included in this compilation thesis are not available in the digital version.


III. Lorentzen, D. G. (manuscript). Is it all about politics? A hashtag analysis of the activities of the Swedish political Twitter elite.


Publications in the series Skrifter från VALFRID


Jarneving, Bo (2005). *The combined application of bibliographic coupling and the complete link cluster method in bibliometric science mapping.* (ISBN 91-89416-12-0) Skriftserien; 30


Olson, Nasrine (2010). Taken for granted: the construction of order in the process of library management system decision making. (ISBN 978-91-89416-26-0) Skriftserien; 45


