TUM School of Computation, Information and Technology
Technische Universität München

Topic-Driven Characterization of Social Relationships for the Analysis of Social Influence

Jan Lukas Hauffa

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: **Prof. Dr.-Ing. Jörg Ott**

Prüfer*innen der Dissertation:

1. apl Prof. Dr. Georg Groh
2. Prof. Dr. Jens Großklags

Die Dissertation wurde am 22.08.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 16.02.2023 angenommen.

Abstract

The digital transformation of our everyday lives has changed our communication habits, but is it a change for the better? The proliferation of politically motivated attempts at manipulating the public opinion is just one example for the challenges that online social media pose to today's society. We believe that two fundamental questions must be answered to enable us to approach these challenges in a meaningful way: What can we learn about the nature of the social relationships people form online by observing how they interact on online social platforms? Specifically, is it possible to detect if one person exerts influence on another? Reviewing the pertinent literature from computer science and the social sciences shows that there is no consensus with respect to either question. In a user study, we are able to show that a characterization of a relationship that is derived from the content of communication is perceived as useful by human judges.

We argue that topic models, unsupervised Bayesian models for textual data, can produce content-based representations that are interpretable by humans and at the same time are useful as intermediate representations for further computational analysis. Using a topic model, the content of communication within an interval of time can be represented by a probability distribution over topics, which usually correspond to discrete concepts. We collect communication data from three online systems: Twitter, Facebook, and e-mail. Despite the outward differences of these platforms, we find that the collected data can be described well by a common functional model of online communication. Investigating the application of topic modeling to online communication, we find that performance can be substantially improved by taking the temporal development of the conversation into account.

Having thus established a conceptual framework and practical tools for learning about the nature of online social relationships, we approach the problem of social influence detection at two levels of granularity. At the micro level, we attempt to test if a change in the behavior of one person can be explained by the earlier behavior of another, which would be evidence for a causal relationship. However, theoretical issues with causal identifiability and an insufficient correspondence between the results and our intuitive understanding of social influence call into question the validity of this approach. At the meso level, the detection of social influence can be reframed as a prediction problem: which parts of a person's social environment are most useful for predicting that person's future behavior? In this way, we are able to identify several patterns of influence that are stable across all platforms.

Kurzfassung

Die digitale Transformation unseres täglichen Lebens hat unser Kommunikationsverhalten verändert. Eine Veränderung zum Guten? Die Zunahme an politisch motivierten Versuchen, die öffentliche Meinung zu manipulieren, ist nur ein einzelnes Beispiel für die Herausforderungen, vor die soziale Medien unsere heutige Gesellschaft stellen. Wir gehen davon aus, dass zwei grundlegende Fragen einer Antwort bedürfen, damit wir diesen Herausforderungen auf angemessene Weise begegnen können: Was kann man über das Wesen sozialer Beziehungen lernen, die online geschlossen werden, indem man das Kommunikationsverhalten auf sozialen Online-Diensten beobachtet? Insbesondere: Ist es möglich, zu erkennen, ob eine Person Einfluß auf andere ausübt? Eine Sichtung der relevanten Literatur aus der Informatik und den Sozialwissenschaften zeigt, dass es auf keine dieser beiden Fragen eine konsensfähige Antwort gibt. In einer Nutzerstudie können wir aufzeigen, dass eine Charakterisierung einer sozialen Beziehung ausgehend vom Inhalt der Kommunikation von den Teilnehmern als aussagekräftig bewertet wird.

Wir stellen zur Debatte, dass Topic-Modelle, also unüberwachte bayesianische Modelle für Textdaten, inhaltsbasierte Repräsentationen sozialer Beziehungen hervorbringen können, die sowohl vom Menschen interpretierbar als auch als Zwischenprodukt für die weitere rechnergestützte Auswertung tauglich sind. Mittels eines Topic-Modells kann der Inhalt von Kommunikation innerhalb eines Zeitabschnitts als Wahrscheinlichkeitsverteilung über Topics dargestellt werden, wobei jedes Topic üblicherweise mit einem bestimmten Konzept assoziiert ist. Wir gewinnen Kommunikationsdaten von drei Online-Diensten: Twitter, Facebook und E-Mail. Trotz ihrer äußerlichen Unterschiede stellen wir fest, dass sich die gesammelten Daten gut von einem gemeinsamen Funktionsmodell der Online-Kommunikation beschreiben lassen. In einer Studie zur Anwendbarkeit von Topic-Modellen auf Online-Kommunikationsdaten zeigt sich, dass ein Modell an Qualität gewinnt, wenn es die zeitliche Entwicklung der Konversation berücksichtigt.

Nachdem wir auf diese Weise ein konzeptuelles Gebäude und die notwendigen Werkzeuge zusammengestellt haben, um die Charakteristika von Online-Sozialbeziehungen zu bestimmen, nähern wir uns dem Problem der Erkennung von sozialem Einfluß auf zwei verschiedenen Aggregationsebenen. Auf der Mikro-Ebene versuchen wir zu testen, ob eine Verhaltensänderung durch das zeitlich vorangehende Verhalten einer anderen Person erklärt werden kann, was Evidenz für eine kausale Beziehung darstellen würde. Theoretische Bedenken in Bezug auf kausale Identifizierbarkeit und eine unzureichende Übereinstimmung zwischen den Testergebnissen und unserer intuitiven Vorstellung von sozialem Einfluß stellen jedoch die Gültigkeit dieser Herangehensweise in Frage. Auf der Meso-Ebene kann die Erkennung von sozialem Einfluß als ein Vorhersage-Problem ausgedrückt werden: Welche Bereiche des sozialen Umfelds einer Person sind für die Vorhersage des zukünftigen Verhaltens dieser Person am informativsten? Mit diesem Ansatz gelingt es, mehrere typische Muster sozialen Einflusses in allen Datensätzen nachzuweisen.

Contents

1 Introduction

Throughout its history, the Internet has been a medium for individual and group communication. Electronic mail (e-mail) was among the Internet's first applications, and continues to be in widespread use, while early public discussion platforms such as Usenet have given way to the *Social Web*, and its most prominent figureheads, Facebook and Twitter. According to the Digital 2020 Global Overview Report (Kemp, 2020), 3.8 billion people, approximately half of the world's population, use social media. While this figure is derived from public statements of social media companies, and therefore should be considered an optimistic estimate, it highlights a strong, ongoing trend in Internet usage.

Social media evolved from systems for computer-mediated communication (CMC), so they are best understood in the context of their ancestors. Herring (2007) classifies CMC systems in terms of ten attributes. We focus on two of them that arguably correspond to the most fundamental design decisions: *Synchronicity* refers to the question whether communication happens in real time or is time-delayed (i.e., messages are retrieved by the recipient at an arbitrary later time). *Privacy* of discourse, as defined by Herring, requires that messages are only visible to the intended recipients, while in public discourse any other user may read a message and respond to it. This results in four basic forms of communication. Services that cover these four classes have existed since the early days of the Internet. E-mail provides private, asynchronous communication, while Usenet, an online discussion board service, used to be a popular form of asynchronous, public communication. Internet Relay Chat (IRC) is a platform for synchronous communication, both public and private. Each of the four classes has seen the introduction of new services over time; for example, IRC lost in popularity to Web-based chat services and instant messengers, which in turn were largely replaced by mobile solutions, first plain text messages, then messenger apps like WhatsApp. Still, all of these services can be succinctly characterized by the class of communication they offer.

Social networking services (SNS) and social media are different in that respect. While there are no generally accepted, formal definitions of the two terms, a frequently stated defining characteristic is that they build upon a CMC system and add two particular features (boyd and Ellison, 2008): they provide a way for users to present themselves, usually in the form of a profile page with pictures and semi-structured text, and incentivize users to publicly declare their social relationships. The sets of users and their relationships make up the social network graph. Social networking services often serve as platforms for the dissemination of user-generated or external content. The terms "SNS" and "social medium" are used interchangeably, although the term "SNS" is usually applied to services that focus on the networking aspect, and "social medium" to services that focus on content distribution. Traditionally, self-presentation and networking have been separate from communication. Consider, for example, personal homepages from the early days of the web, which often

contain some biographical data and link to other, related pages, but offer no in-medium way of contacting the author. The first popular SNS, SixDegrees, commenced operation in 1997 (boyd and Ellison, 2008). The current mainstream offerings, Twitter and Facebook, are maintaining their popularity, but new services that are tailored to a more specific audience are rapidly catching up (Kemp, 2020). Examples are Instagram, an SNS built around photo sharing, and mobile-first services like TikTok, a video sharing app with limited social networking features.

The popularity of online communication and social media is rooted in technological developments that took place over the past three decades and are still unfolding as of today. The advent of affordable dial-up Internet access in the 1990s started the trend of increasing availability of Internet access to the general population and its growing adoption for personal, rather than work-related use. The subsequent technological shift to always-on broadband made an impact on an Internet user's ability to communicate online: Dial-up access implies that the user is available for communication only when connected to the Internet, while a constant connection means the user is available whenever he or she is in front of the PC. The introduction of smartphones and mobile broadband services had an even stronger effect. Internet access is now close to ubiquitous, and availability almost constant. As of 2019, 53% of all web traffic originates from mobile phones (Kemp, 2020). This development is aptly summarized by Twitter user kappa_kappa (2016):

> *"Remember when we used to say 'brb'* ['be right back'] *all the time when we were online? We don't say it anymore. We no longer leave. We live here now."*

Technological development is transforming our communication habits, but is it a change for the better? One one hand, social media enable the exchange of information and ideas across geographical, political, and societal boundaries. They are credited with a supporting role in grassroots political movements such as the Arab Spring (Christensen, 2011). On the other hand, they are associated with a number of negative social phenomena: the deliberate spread of misinformation, e.g., by social bots in the 2016 US presidential election (Bessi and Ferrara, 2016), and the creation of *echo chambers*, social environments in which hateful ideologies and conspiracy theories are circulated and amplified, while opposing voices are being suppressed (Ferrara, 2015). Social media are also more generally held responsible for creating a more polarized and emotionally charged discourse (Riebe et al., 2018). To understand these complex phenomena and to be able to counteract their effects on society, we must first understand the more fundamental mechanisms that drive social interaction in online spaces.

To a researcher, social media and their dual nature as platforms for self-presentation and communication present a unique opportunity to learn about interpersonal relationships. The public accessibility of social media makes it possible to collect and evaluate large amounts of data about people and their communication behavior in a natural environment, i.e., an environment that is not artificially constructed for observational or experimental purposes. In addition, the social network graph makes the relationships between users visible, so that individual behavior can be viewed in the context provided by the social environment. In this thesis, we use social media data to approach the problem of identifying the driving forces behind people's online behavior from two directions: bottom-up by characterizing the

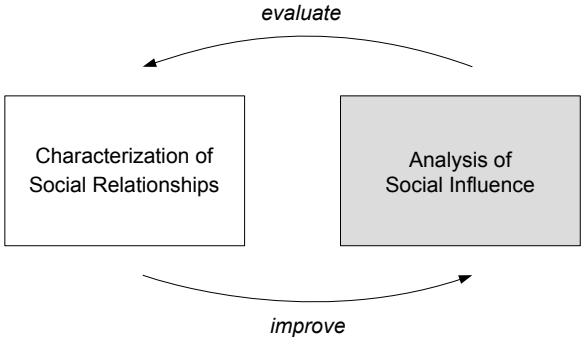


Figure 1.1: Symbiotic relationship of the two main research goals of this thesis

relationship between two people according to the content of their communication, and top-down by inferring the network structure of social influence from changes in communication behavior.

The study of individual users of an online system and their characteristics and preferences has received much research attention; see, for example, the body of work on recommender systems (Yang et al., 2014). When studying social interactions, the smallest unit of analysis is the *dyad*: two people in a social relationship. Dyadic relationships are well understood from a sociological and psychological perspective, but we intend to show that the field of social computing currently lacks a coherent and comprehensive language for representing dyadic relationships and reasoning about their characteristics. Currently, social relationships are mainly characterized in terms of interaction meta-data: Who communicates with whom, when, and how often? We argue that there is untapped potential in the analysis of the communicative content of interactions, and propose a low-dimensional representation of a relationship that is solely derived from content. By means of probabilistic topic modeling (Blei, 2012), textual content can be transformed into a vectorial representation. The resulting topic vectors are human-interpretable (Chang et al., 2009b) and can therefore, for example, serve as the basis of a qualitative study of social behavior, but are also suitable as an intermediate representation for further computational analysis.

Coming from the opposite direction, we investigate how users of a social medium influence each other in their communication behavior. Social influence is closely linked to information dissemination in social networks, considering that receiving a piece of information from someone and passing it on to others is visible evidence for a particular form of influence. More generally, one can test each edge (a, b) of the social network graph for time-delayed similarity of the behavior of a and b to reconstruct the underlying influence network. Due to its central role in information dissemination, social influence is directly involved in any attempt to control the flow of information, such as viral marketing or the manipulation of public opinion in political discourse. By detecting and visualizing influence relationships, we aim to create transparency, which may ultimately contribute to improving the quality of discourse.

Each of these two lines of research — the characterization of relationships and the detection of influence — stands on its own, but the two are connected by the motivation to

improve our understanding of human social behavior in online social networks. A systematic effort at detecting social influence requires a numeric representation of dyadic human interaction as described earlier. If our proposed representation leads to new insights about social influence, this is proof of its utility. This symbiotic relationship between the top-down and bottom-up approach is depicted in figure 1.1.

1.1 Outline

This thesis is structured as follows: In chapter 2, we review literature from the fields of sociology, psychology, and computer science on the topic of social relationships. We argue for a content-based representation of social relationships and present the results of a user study, which tests the perceived utility of content-based representations in the context of online social networking services. Further systematic experimentation on the properties of social relationships and social influence requires observational data from social media. In chapter 3, we obtain data from three platforms, including SNS and traditional CMC systems, provide basic descriptive statistics of these datasets, and discuss the ethical issues associated with using social media data for research. Using the information gathered up to this point, we are able to design a content-based representation of social relationships that takes the particularities of each platform into account. In chapter 4 we discuss how the framework of probabilistic topic modeling provides a principled way to distill communication data into the desired representation.

This concludes our work on characterizing social relationships, and we turn to the problem of detecting social influence in online communication. Chapter 5 lays a common foundation for the following experiments. We review existing work on social influence, derive a basic cognitive model of the influence process, and discuss how the analysis of social influence can benefit from the topical representation of social relationships. In chapter 6, we study social influence at the micro level and propose and evaluate several methods of detecting dyadic influence relationships. In chapter 7 we move on to the meso level of analysis and investigate to what extent nodes in a social network are influenced by their social environment. Finally, chapter 8 summarizes our findings and discusses them in the context of the original research questions.

Source code of the data processing pipeline and the experiments is available on GitHub (https://github.com/jhauffa/crawler and https://github.com/jhauffa/influence) or from the author on request.

2 Characterizing Social Relationships

Parts of the content of this chapter have been previously published at the IEEE Third International Conference on Social Computing (Hauffa et al., 2011). The user study described in section 2.2.2 was conducted in the context of the diploma thesis of Tobias Lichtenberg (2011), which was jointly supervised by Jan Hauffa and Georg Groh. The results have been previously published at the 2012 International Conference on Social Informatics (Hauffa et al., 2012) and as an arXiv preprint (Hauffa et al., 2014). This chapter contains verbatim and near verbatim quotations from prior publications (Hauffa et al., 2011, 2012, 2014), which are highlighted in gray.

From the moment of our birth, the formation and maintenance of social relationships is part of our everyday lives. Yet, the concept of a relationship is surprisingly hard to put in words. We start by restricting the scope of our analysis to relationships that are social, dyadic and interpersonal; in other words, the presence or absence of a relationship is a property of a social group of size two, and the members of this group are individual human beings. There are two well-known definitions for relationships of this kind: Kelley (1983), as cited by Clark and Reis (1988), nominates *interdependency* as the defining characteristic of a relationship. Two people are interdependent if their "behaviors, emotions, and thoughts are mutually and causally interconnected" (Clark and Reis, 1988). This results in what Sillars and Vangelisti (2006) call *non-summativity*: A relationship is characterized by the nature of the interdependency of the related people rather than by the sum of their attributes and actions, and can therefore be said to exist as a distinct entity.

Hinde proposes a definition that is compatible, but more narrow in scope. He states that a relationship "involves a series of interactions in time" (Hinde, 1976), thus focusing on interaction as the main mechanism that creates interdependency. Hinde does not place constraints on the nature of these interactions, but, according to an axiom of Watzlawick, it is impossible to interact without any verbal or non-verbal communication taking place (Watzlawick et al., 1967, cited by Sillars and Vangelisti, 2006). We may therefore assume that relationships are to a large part enacted by — and can be understood through — communication:

> *"Relationships exist in the structure and content of communication over time. [...] Interaction both stimulates changes in cognition and affect and is the medium through which those changes become real for the self and others."* (Parks, 1997)

Kelley's and Hinde's definitions both acknowledge the role of continuity over time. The length of the time span for which people are interdependent, or sustain regular interaction, is what distinguishes a relationship from a series of inconsequential encounters (Rogers, 1998). Using their definitions as a starting point, we can identify three fundamental elements of a social relationship: interaction, reciprocity, and continuity. Miller summarizes their respective functions as follows: "in order to maintain the [dyadic] group both participants

must construct reciprocal interaction with a high level of involvement with each other." (Miller, 2007)

The aforementioned definitions originate from the social sciences, while this work is rooted in the field of social computing, which is concerned with the intersection of human sociality and information processing. A central premise of social computing is that computer systems can support people in their everyday communication and, conversely, the resulting communication data can be analyzed to learn about social behavior. By its very nature, social computing has strong ties to the social sciences. In this work, we use the methodological framework of social network analysis (SNA), which is common to both fields. Instead of people and their relationships, in SNA the units of analysis are *actors* and their *relational ties*. Actors can be arbitrary entities, but are usually expected to be homogeneous in some respect. Actors are connected by ties, which are induced by a binary relation (in the mathematical sense) on the set of actors. Depending on the point of view, the resulting structure can be seen as a social network, a model of social organization that takes inventory of entities and their relationships, or as a graph, to which all methods and algorithms of graph theory can be applied. Wellman (2007) lists a wide range of examples for actors and relational ties, including cities and the means of transportation that connect them. The implication is that actors do not necessarily have to be individuals or social collectives. As long as the actors are manifestations of social behavior, the results of the analysis of the network graph can be interpreted in terms of social behavior as well. In this way, SNA facilitates the analysis of human sociality by means of computation.

Since we operate within the framework of social network analysis, we use the associated terminology. In the social sciences, the terms "relationship" and "relation" are used almost interchangeably (cf. the etymological remarks of Conville and Rogers, 1998). To avoid confusion with the mathematical concept of a relation, we always refer to dyadic social ties as relationships, and use the term "relation" either strictly in the mathematical sense, or figuratively for the class of relationships that is represented by a relation on the set of actors.

The definition of a relational tie in social network analysis is much less prescriptive than any of the definitions of social relationship discussed previously. This discrepancy between SNA and other methodological frameworks of the social sciences can be in part explained by a paradigm shift, starting with Granovetter's research on weak ties (Granovetter, 1973). Granovetter defines tie strength as an abstract, unobservable property of a dyadic relational tie, but hypothesizes that it is linearly associated with a number of observable characteristics: continuity over time, emotional intensity, intimacy, and reciprocity. By comparing these characteristics to the fundamental elements of a social relationship we identified earlier, one must conclude that a relational tie, as its strength decreases, at some point ceases to be a relationship in the sense of Kelley and Hinde. Surprisingly though, Granovetter finds that the value an individual derives from a relational tie is not a linear function of tie strength. Weak ties fulfill a distinct structural role as bridges between groups of strongly tied actors, and therefore facilitate the diffusion of information across group boundaries. By means of an experiment and multiple meta-analyses, Granovetter demonstrates that weak ties are valuable in information transmission, job search, and community formation. Consequently, he argues that an understanding of social behavior on the macro level cannot be achieved by focusing solely on the small, densely connected groups formed by strong ties, but requires

knowledge of the weak ties that interconnect such groups.

Another reason why SNA did not adopt one of the more stringent definitions of social relationship is inherent to the definitions themselves. Relationships and strong ties are defined in terms of characteristics that are either not directly observable, e.g., emotional intensity, or require substantial interpretation, such as continuity: How long does interaction have to be sustained to constitute a relationship? SNA can accommodate an interpretive workflow as described by Charania and Ickes (2006), where the researcher selects suitable candidates, elucidates the relevant characteristics of their relationship via structured interviews or questionnaires, and finally assigns the relationship to one of a set of predefined classes. However, social networks are just as useful as a representation of unfiltered observational data. The interpretive work of a domain expert yields low amounts of high quality data. In social computing research, the situation is reversed: observational data is cheap, but noisy, and often requires further processing to be useful. A "friendship" on Facebook or "following" someone on Twitter, for instance, does not necessarily indicate a strong social tie. Pending further analysis, one cannot say much about the nature of the relationship that goes beyond mutual awareness.

We find that, in the social sciences, there is a *conceptual gap* between observable reality and the predictions made by theories, which is usually left to be filled by the researchers conducting a study or experiment, drawing on their expertise and life experience. We provide more examples for this gap in the following sections. Manual processing of observations invariably introduces subjectivity and, by treating the life experience of the researcher as normative (Duck et al., 1997), might also impart the researchers' biases onto the data. By using algorithmic, data-driven methods from the field of social computing, we aim to reduce bias from this particular source, while keeping in mind that bias may also arise from the method and scope of data collection, in addition to general societal biases reflected by the data. Neither theoretical work in the social sciences nor the algorithmic contributions from social computing have managed to close this conceptual gap as of yet. The often heard call for "big data" research to be grounded in theory and in the results of qualitative, "small data" studies (boyd and Crawford, 2012; Kitchin and Lauriault, 2018) requires researchers from different disciplines to find a common language. Our work on the characterization of social relationships should be understood as a contribution to that ongoing research program.

The *characterization* of a social relationship is the process of distilling the observed behavior of the involved actors into attributes that reside on a higher level of abstraction than the observations. Theories from the social sciences motivate different ways of characterizing social relationships. For example, according to the theory of social capital, relationships are characterized by the amount of resources that are, potentially or actually, exchanged between two actors, while in the context of social networks, tie strength is the primary attribute of a relationship. These are quantitative characterizations; conversely, a social relationship can be characterized in a qualitative way, which means finding a way to represent its type, role, or function. The value of a characterization ultimately depends on how much it contributes to answering the questions of the researcher. However, we argue that a characterization that is generally useful across a wide range of tasks has two properties: human interpretability and utility as an intermediate representation, which could be informally described as "computational interpretability". Human interpretability directly contributes to

the explainability of inferences made from the characterization, and therefore improves the transparency of a computer system built upon it. A characterization that is interpretable can also be judged in terms of its quality by domain experts. Complementary to interpretability, the utility of a characterization as a building block for more complex systems can be determined by extrinsic evaluation (as defined by Galliers and Sparck Jones, 1996).

In this work, we aim to leverage the large amount of communication trace data (Howison et al., 2011; Golder and Macy, 2014), both meta-data and actual communication artifacts, that is produced in the context of social networking services. In the following section, the problem of using observational data for characterizing social relationships and representing the gathered knowledge is analyzed from different angles. As an outsider to the discipline, one cannot hope to perform an exhaustive survey of all related research in the social sciences. Nevertheless, we make an attempt to identify the main areas of conceptual overlap between the social sciences and social computing, so that we can contrast theoretical approaches to characterization with tangible, data-driven methods, and identify common problems. Finally, we argue that content-based representations best embody the complementary qualities of interpretability by human and machine. The former quality is demonstrated in a user study at the end of this chapter, while the latter is the subject of extrinsic evaluation in chapters 5 to 7.

2.1 Quantitative and Qualitative Approaches to the Characterization of Relationships

In the social sciences, as well as in the field of social computing, diverse efforts have been made to characterize social relationships. We briefly review the main approaches and the associated conceptual and practical issues. We begin by discussing quantitative characterizations in the context of weighted social network graphs, and move towards characterizations that are successively more quantitative and holistic, in the sense of taking an increasingly complex notion of social context into account.

2.1.1 Social Network Analysis

Social network analysis is a straightforward way of making sociality accessible to computational analysis. A social network is defined by its graph, a tuple (V, E) with a finite set of vertices V, corresponding to social entities (*actors*), and a set of edges $E = \{(x, y) \mid x, y \in V \wedge x \neq y\}$, corresponding to social relationships (*ties*) among the entities. In other words, a social network graph is induced by a binary relation, which is usually constructed by observing or modeling an actual class of social relationships. The relation is either non-symmetric or symmetric, and therefore induce a directed or undirected graph, respectively. If the relationships of interest are characterized by reciprocal interaction, a symmetric relation is appropriate, while a non-symmetric relation allows modeling a mixture of social and parasocial (Stever and Lawson, 2013), or otherwise one-sided relationships.

When constructing a social network, the first task is to identify the actors, followed by defining a binary relation. A common assumption is that, within a particular online ser-

vice, one account corresponds to one natural person, but this is not necessarily true. One counter-example is posed by the social media accounts of companies and celebrities, which are usually maintained by a marketing team; another counter-example is the maintenance of multiple "sock puppet" accounts by one person in order to amplify his or her voice in discussions. However, the detection of such instances of m-to-n mapping is difficult, and often it is acceptable to treat accounts as individual entities, even if their ownership is unknown. The relation on the set of actors is usually defined implicitly, via rules that specify whether the observation data constitutes sufficient evidence for the existence of a relational tie. Allan (2006) points out that "the structure of a personal network remains highly dependent on what the analyst counts as warranting making a 'link' between two individuals within the network", so in order to avoid bias, one should choose criteria that arise directly from the observation data, are motivated by the research question, or have a theoretical justification. The separation of relationships from non-relationships can be viewed as the most basic form of characterization.

The least restrictive non-trivial criterium for the presence of a social tie is *awareness* or knowledge of a particular person, which implies a directed social network graph. Milardo (1992), as cited by Allan (2006), lists three more criteria that are commonly used in the social sciences: The first is psychological significance, either in terms of emotional closeness or specific social roles, such as friendship or kinship. The second is exchange of emotional or physical support, and the third is regular interaction above a specified frequency threshold. According to Allan (2006), these types of networks "suggest different types of questions, and are designed to address different theoretical or substantive issues." Milardo (1989), as cited by Parks (1997), refers to the networks induced by the first two criteria as "psychological networks", while the third criterium induces "interactive networks". In a study of the psychological and interactive networks of marriage partners, Milardo reports only 25% of overlap between the two networks.

Online social network services allow their users to declare social relationships with other users. These can be bidirectional, requiring confirmation from both parties, or unidirectional, e.g., "following" someone on Twitter. The set of declared relationships induces an explicit social network, similar in character to the psychological networks of Milardo. Again, the overlap with the interactive network tends to be low: in a study of Facebook "friendship" networks, Wilson et al. (2009) find that "nearly all users can attribute all of their interactions to only 60% of their friends." An implicit or interactive social network graph can be constructed by inserting edges between pairs of actors if their frequency of interaction within a period of time exceeds a threshold. Little guidance exists for choosing this threshold, yet Tsur and Lazer (2017) observe a strong effect of the threshold on the structure of the resulting graph and the inferences that can be drawn from it. In the context of tie strength, Granovetter (1973) warns about the ambiguity that results from determining the presence or absence of an edge by applying a threshold to an observed quantity:

> *"Included in 'absent' are both the lack of any relationship and ties without substantial significance [...] In some contexts, however [...], such 'negligible' ties might usefully be distinguished from the absence of one. This is an ambiguity caused by substitution [...] of discrete values for an underlying continuous variable."*

In the absence of information that would motivate a non-arbitrary choice of threshold, requiring reciprocal interaction might be a suitable replacement. A combined requirement of reciprocity and frequent interaction could even be viewed as an approximation of the three elements of Kelley's and Hinde's relationship definitions. Implicit social networks can be extracted from any kind of online communication, as long as sender and recipient are identifiable. An example for this is the study of Bird et al. (2006) on the extraction of social networks from collections of e-mail messages, which lends empirical support to Wellman's earlier claim that various kinds of "electronic groups" implicitly define social networks (Wellman, 1997).

Given that a typical social relationship is enacted by communication through multiple different channels, online and offline, no explicit or implicit social network of a single medium can claim to be an accurate representation of reality. The implicit network lacks relationships for which interaction mainly happens outside of the medium, while the explicit network overstates the importance of relationships that rarely engender interaction anywhere. In a large-scale study of Facebook, Viswanath et al. (2009) find that on average, less than 30% of a user's declared relationships remain active from one month to the next. They explain the change in activity with temporal interaction patterns, and distinguish patterns of frequent and infrequent interaction. In conclusion, the relationships that can be observed among users of an SNS or comparable online service are conceptually heterogenous, and a substantial amount of information is lost when they are represented in a dichotomous way by the presence or absence of an edge in a graph.

In order to move away from a purely graph structural approach to SNA, different methods of augmenting the social network graph with information extracted from communication among the actors have been proposed. We introduce a weight function $f : E \to W$, where W is the set of possible weights, and obtain a *weighted social network graph* (V, E, f). Signed social networks use a basic form of weighting, $W = \{+, -\}$, where the sign corresponds to a subjective valuation of the relationship in terms of emotional valence (Wasserman and Faust, 1994, ch. 4.4). Social balance theory (cf. Tang et al., 2016 and the discussion by Szell et al., 2010 in the context of an online multiplayer game) uses this additional bit of information to predict, via rules like "the friend of my enemy is an enemy", the evolution of the social network over time. Conversely, systematic violations of such common-sense rules may point towards more complex social processes. The next step in representational complexity is the use of real-valued weights. Barrat et al. (2004) demonstrate how to adapt standard methods of SNA to weighted networks with $W \subseteq \mathbb{R}$, and confirm that the augmentation of graph structure by weights affords new insights into the social network. This model can be further generalized to networks with multivariate weights $W \subseteq \mathbb{R}^n$.

A weighted social network represents relationships as points in a one- or multi-dimensional space. A number of theories from the social sciences have attempted to identify general properties of relationships, which could serve as the dimensions of that latent space (Rogers, 1998). For example, Hinde (1995) names ten properties: content, diversity, quality, temporal distribution, and reciprocity of interactions, as well as the subjective perception of the relationship in terms of power balance, intimacy, shared perception of reality, commitment, and satisfaction. Not all of the resulting dimensions lend themselves equally well to computational analysis and measurement in observational data. In Hinde's proposal, the

directly observable attributes of interaction are in stark contrast to the principally unobservable assessment of the relationship by the involved actors. Therefore it is no surprise that empirical research in social computing has focused on those properties of relationships that can justifiably be derived from observable interaction. Following Granovetter's conceptualization of tie strength (Granovetter, 1973), Gilbert and Karahalios (2009) develop a statistical predictive model for measuring the strength of relationships between users of online social network services, which incorporates communication statistics, similarity of user profiles, and the content of messages as features. The predictions of the model are compared to the results of a survey of 35 Facebook users, and are found to be sufficiently accurate for distinguishing strong and weak ties. However, Cronbach et al. (1972), as cited by Kenny (1988), caution that "[t]he claim that [a weighted function of variables] has validity as a measure of some construct carries a considerable burden of proof." Kenny therefore recommends that variables that can be observed separately should also be analyzed separately (Kenny, 1988).

Conceptually similar to tie strength is *social capital*, which Lin (1999) defines as "resources embedded in a social structure which are accessed and/or mobilized in purposive actions", that is, as a property of the social network as a whole. The social capital of an individual is usually understood as the sum of resources he or she can potentially mobilize from neighbors in the graph (Snijders, 1999), decomposing it into a property of the relationships. Schams et al. (2018) find that viewing social capital from the opposite direction, defining it as the sum of resources provided by an individual to the network, simplifies the process of its measurement in observed interactions. While they treat social capital as a property of the individual to facilitate evaluation against ground truth values obtained from a survey among 165 participants, their method can be adapted to yield relationship-level measurements.

A common modeling assumption in SNA is that the relationships within a social network are generally homogeneous in nature, and vary only in a small number of previously specified attributes. This limitation to a single type of social relation has been carried over to most online social network services. Even the ontological vocabulary FOAF, designed to be a formal language for describing potentially heterogenous "social networks of human collaboration, friendship and association" (Brickley and Miller, 2010), can only represent the very general concept of having knowledge of another person (`foaf:knows`). While this design choice makes the language agnostic towards heterogeneity, it does so by hiding differences between individual relationships that might be important for accurate inference from the network structure.

In addition to inter-relationship heterogeneity, individual relationships may be heterogeneous, either because they evolve over time, or because they are consistently enacted in multiple, substantially different ways. The latter condition is known as *multiplexity*. Common forms are multiplexity in the content of interaction, multiplexity in the use of communication channels, and, on a more abstract level, multiplexity in the type of relationship (e.g., siblings who are friends). Networks of multiplex relationships are usually modeled as multigraphs, where two nodes can be connected by more than one edge. Edge-attributed graphs and multi-layered graphs are equivalent to multigraphs in representational capability. Bothorel et al. (2015) discuss how the additional information in edge-attributed graphs can improve the accuracy of community detection. A basic way of modeling temporal variability is to extract a time-indexed sequence of social networks from observations of the

network at different intervals of time. Statistical methods for the analysis of time series data can then be applied to edge weights or to aggregate metrics of the network graph.

Bias of Self-Reports and Trace Data

In statistical terms, the quantitative characterization of relationships can be expressed as a regression problem, where human-generated reference data is used for learning to predict the quantity of interest from observation data. Reference data can be obtained by interviewing a sample of actors about their relationships or by having one or more domain experts annotate a sample of relationships. Since domain experts are involved, either as interviewers or annotators, this process is expensive and therefore only applicable to limited amounts of data. Furthermore, a number of issues may negatively affect the quality of the resulting data (Charania and Ickes, 2006): Interviewees and annotators impart their own subjective perspective on the assessments. Having a relationship described by its two participants yields assessments from two distinct points of view, which have to be reconciled into a single value, and being interviewed about the nature of their relationship may cause them to re-evaluate that relationship. These factors potentially lead to a high degree of uncertainty in the assessments and low annotator agreement. Marsden (1990) summarizes the findings of earlier empirical studies on the accuracy of self-reports as follows: "[...] data on observable features of alters are of high quality, while those on attitudes or internal states are generally poor [...]. Most network data appear to be of better quality for close and strong ties than for distal and weak ones."

In a controlled environment, it is possible to improve the reliability of assessments via sophisticated interview techniques. An example is the method developed by Antonucci and Akiyama (1987) for the elicitation of close relationships in the context of the convoy model. In order to visualize an egocentric network, a visual representation of the interviewee is placed at the center of a number of concentric circles, which represent decreasing degrees of closeness. The interviewee is then asked to name people known to him or her, and place them inside the circles. According to Allan (2006), a benefit of this method is that it "requires respondents to make comparisons about the relative properties and qualities of the different relationships they include." In other words, it encourages comparisons between relationships to obtain more accurate absolute ratings. This suggests that asking for a rank ordering instead of absolute assessments might be a generally viable strategy for improving the quality of assessments.

A new source of bias appears when, instead of collecting assessments in a controlled environment, publicly available trace data is used. The social dynamics of openly declaring one's perspective on a relationship necessitate the employment of strategies for minimizing the chance of a negative outcome. Teng et al. (2010) examine online services that let users rate each other, and find that the distribution of ratings is strongly biased towards the positive end of the scale if ratings are public and raters are identified by name. They attribute this effect to the desire to avoid a negative reciprocal rating. Rating another user, e.g., in terms of trustworthiness, means making a strong statement about the relationship. It is conceivable that weaker public statements about a relationship are also affected by this kinds of bias. For example, declining a friendship request on an SNS like Facebook not only may offend

the requester, the resulting absence of a tie is also visible to all current friends. In certain situations, e.g., when receiving a request from someone who is already well-connected with one's current circle of friends, the decision to accept the request may be strongly driven by wishing to avoid negative social consequences.

In the light of these quality issues, one may attempt to reduce, or eliminate altogether, the dependence on reference data. Earlier research by the author of this thesis (Hauffa, 2010; Groh and Hauffa, 2011) explores the limits of unsupervised quantitative characterization of relationships from observation data alone. Volunteer participants provided a selection of their e-mail messages and were asked to rate their relationships with the senders and recipients on discrete scales of emotional intensity and valence. The resulting dataset contains 399 messages exchanged between 122 actors. The set of ratings has two pronounced clusters: one contains relationships of little emotional intensity and slightly positive valence, the other contains emotionally intense and strongly positive relationships. Features are extracted from meta-data and content of the messages. The prediction of valence is based on sentiment polarity features, while the set of features for the prediction of emotional intensity also includes features derived from message volume and specific characteristics of emotional language in online communication. Predictions are obtained in an unsupervised way by applying dimensionality reduction to the feature vector.

The study compares four methods of transforming the prediction to match different hypothetical distributions of the ratings. The first method uses the minimal and maximal prediction to linearly transform all predictions into the interval $[0, 1]$, while the second method transforms the predictions so that all values within $\pm 3SD$ of the mean end up within $[0, 1]$. The remaining two methods "cheat" by making use of limited amounts of information about the reference ratings. The third method is similar to the first, but transforms the predictions into the interval between the lowest and highest rating. The fourth method applies the transformation of the second to both the predictions and the ratings. The improvement in root mean square error over a baseline predictor that outputs a fixed value of 0.5 is -38.2% for the first, -6.5% for the second, 14.4% for the third, and 32.9% for the fourth method. An improvement of accuracy over the baseline could only be achieved by introducing knowledge about the distribution of ratings, thereby providing a frame of reference for the prediction.

2.1.2 Ontologies and Classification

A first step towards qualitative characterization is to assign a relationship to one of a predefined set of classes or prototypes. Scholars from the social sciences have long argued about the relative merits of such a typology of relationships. An argument in favor of discrete types is that they correspond better to our intuitive conceptualization of relationships. With reference to the empirical work of Haslam (1994), VanLear et al. (2006) state: "When most people think about relationships, they identify them as types or kinds of relationships, not as points along a set of continuous dimensions." They argue that a categorical representation is preferable to a multi-dimensional space if the observed relationships form clearly separated clusters within that space, but acknowledge the counter-argument of Griffin and Bartholomew (1994), who point out the loss of information compared to a vectorial representation. Argyle and Henderson (1985), as cited by Bossert (2010), argue against qualitative

characterization in general: Considering that social relations are so intimately important and emotionally loaded that is hard for humans to describe them verbally, elicitation of the required information with consistently high quality is difficult, so quantitative approaches should be preferred.

If we accept the arguments in favor of discrete relationship types, the question is how to identify and distinguish different types. On a theoretical level, VanLear et al. discuss two mechanisms by which relationship type and observable behavior might be connected: Haslam (1994) hypothesizes that people form distinct internal representations of particular classes of relationships to reduce cognitive load. These representations encode norms and behavioral rules that prescribe how to enact the relationship. Another school of thought is that relationships are constantly evolving, but repeatedly occurring patterns of interactions give rise to a temporally local notion of relationship type (Watzlawick et al., 1967, as cited by VanLear et al., 2006). In either case, the relationship type can be understood as a latent variable that guides observable behavior, so inference is at least theoretically possible.

When trying to formulate the classification of relationships as a machine learning problem, neither supervised nor unsupervised learning algorithms appear to be directly applicable. Supervised classification would require the selection of a meaningful subset from the unknown and potentially unlimited set of possible relationship types. Unsupervised clustering of data pertaining to the social network is not guaranteed to produce clusters that are meaningful to the human observer, and shifts the burden of associating the input data with higher-level concepts to a later stage of analysis. It follows that, as a prerequisite for classification, the space of relationship types needs to be explored and formally specified. In the social sciences, there are many sub-fields that are concerned with the study of specific types of relationships, for example, emotionally close relationships (Kelley et al., 1983). Berscheid (1995) criticizes how "relationship type [is] confounded with disciplinary approach", in other words, the tendency to perceive each relationship type as its own field of research that requires bespoke theory and methodology. She calls for the development of an "overarching theory of relationships" that connects observable relationship phenomena with the "principal relationship types" (Berscheid, 1995), but the review of VanLear et al. (2006), written a decade later, does not name any concrete attempt at formulating a comprehensive typology.

Haslam's idea that relationship types implicitly arise from internal representations can be understood as an argument for a hierarchical typology of relationships. If people can successfully navigate a novel social situation by falling back to generalized knowledge about similar situations, then it is reasonable to distinguish relationship types by their level of abstraction. Koerner and Fitzpatrick (2002), as cited by VanLear et al. (2006), give evidence for a tree-shaped organization of internal representations of social knowledge, with specific relationships as the leaves, general social knowledge being located at the root, and abstracted knowledge about different classes of relationships on the levels in between.

Finally, some theories seek to identify universal, high-level characteristics of relationships. VanLear et al. (2006) use the term "general typologies" for such attempts to "identify the fundamental features of the whole length and breadth of human relationships." An example is the typology of Fiske (1992), who, as cited by VanLear et al. (2006), claims that across all cultures, relationships can be characterized by the degree to which they conform to four basic models: communal sharing (actors are of equal standing within a larger social

Table 2.1: Precision and recall of a naive Bayes classifier for relationship characteristics (Zec, 2008; Hauffa et al., 2011)

class	recall	precision	F-score
professional	0.770	0.656	0.71
miscellaneous	0.592	0.697	0.64
personal	0.627	0.735	0.68
neutral	0.872	0.783	0.82
negative	0.492	0.403	0.44
positive	0.544	0.722	0.62

group), authority ranking (relative rank determines privileges and responsibilities), equality matching (maintaining balance by reciprocal exchange), and market pricing ("investing" in others and expecting proportional return). Fiske's four models are quite removed from the perception of a relationship by its participants, but enable the comparison between different relationship types by expressing them as points in a four-dimensional vector space. Empirical results about the distribution of relationship types within that space could form the basis of a higher-level categorization.

A concrete example for the use of machine learning for supervised classification of social relationships can be found in the work of Zec (Zec, 2008; Hauffa et al., 2011), who uses e-mail communication data to learn about relationship characteristics. Zec hypothesizes that the nature of a relationship manifests in the communication behavior of the involved actors, and the association between communication patterns and relationship characteristics can be learned. Individual messages are classified according to their relationship type, which may be one of "professional", "personal", and "miscellaneous", and the emotional valence ("positive", "neutral", and "negative"). The rating of a relationship is the majority class of its associated messages.

A corpus (i.e., a collection of text documents) for training and evaluation was generated from the CMU version of the Enron corpus (see section 3.3.1) by manual annotation of 1 050 messages. The class distribution is somewhat imbalanced: 31% of messages are of type "personal", 46% are "professional", and 23% "miscellaneous"; 33% are of positive emotional valence, 6% are negative, and 61% are neutral. Features are derived from a vector space model of the message body, ignoring quoted and forwarded parts, after stop word removal and stemming. Multiple classifiers are trained and evaluated on the corpus: k-nearest neighbor, naive Bayes, C4.5 decision trees, and support vector machines. In terms of accuracy, the naive Bayes classifier consistently outperforms all other classifiers, achieving an accuracy of 68.5% for relationship type and 74.0% for emotional valence. The performance in terms of precision and recall is shown in table 2.1. Zec attributes the unsatisfying performance of the classifier in part to the use of rhetorical devices such as irony, sarcasm, and metaphor; important aspects of emotional language that cannot adequately be represented by a simple word-based classifier. In a follow-up experiment on unsupervised learning, neither k-means nor DBSCAN find clusters that correlate with the manually identified relationship types.

Ontology Engineering

In social computing, the development of concrete relationship typologies is typically cast as an ontology engineering task. An ontology, in the computer science sense of the word, is a formal specification of a domain of discourse in terms of objects, their membership in categories, their attributes, and their relationships. The knowledge represented by an ontology can be either abstract (e.g., defining a hierarchical structure of categories) or concrete (e.g., placing a specific object within a category). This duality is reflected by the two main use cases for ontologies: they facilitate the collaborative formalization of domain knowledge, and, within the framework of the semantic web, are the canonical way of building machine-readable representations of domain-specific, heterogeneous data. The vision of the semantic web — although never fully realized — is that machine-readable, interlinked, public representations of knowledge will enable new applications that directly ingest and interact with user-generated content (Sheth and Nagarajan, 2009). We are interested in both use cases of ontologies. Ontology engineering techniques may be helpful for building a consistent formalization of social relationship types and their defining characteristics, and we briefly investigate the value of the semantic web as a source of data about actual relationships.

Extant work on ontology engineering in the social space focuses on ontologies for the semantic web. FOAF (Brickley and Miller, 2010) is an ontology for persons, social groups, their resources and their relationships. It is implemented as an RDF vocabulary, RDF being the canonical format for ontology data on the semantic web. FOAF only specifies a single relationship type, awareness of a person. Relationship (Davis and Vitiello Jr, 2010) extends FOAF by providing a range of additional relationship types. Another extension, MeNowDocument (De Gan, 2004), is mainly concerned with temporally variable properties of individuals ("status updates"), but also proposes some relationship-specific properties for describing the nature of intimate relationships and the emotional valence towards a person. Trust networks can be considered as special cases of social networks. Golbeck et al. (2003) develop an ontology for expressing trust in social networks, specifically addressing interoperability with FOAF. Mika and Gangemi (2004) discuss how further concepts from the social sciences could be integrated into FOAF. Jung and Euzenat (2007) demonstrate the multi-layered reasoning that is enabled by ontologically formalized social networks.

A number of studies give an account of the use of FOAF on the public-facing web at different points in time. Paolillo and Wright (2004) analyze 33 542 RDF documents containing FOAF elements. They find that almost 92% of documents originate from LiveJournal, the first major adopter of the standard. LiveJournal is primarily a blogging service, but it also includes basic social network functionality like user profiles and explicit relationships between users. Similarly, Ding et al. (2005) report on the prevalence of FOAF and perform basic social network analysis. In a preliminary study from 2010, we revisit the usage of FOAF by LiveJournal. Our study is smaller in scale, based on data from only 1 000 users, but, via comparison to the earlier results, allows us to track the development of FOAF usage over time. An RDF property either describes an object or asserts a relationship between two objects. Table 2.2 gives the frequency of five properties relative to the number of "Person" objects. The chosen properties occur in the datasets of at least two studies and contain actual personal information, rather than meta-data related to the user account. Note that Paolillo and

Table 2.2: Relative frequency of selected FOAF properties per "Person" object

property	Paolillo and Wright (2004)	Ding et al. (2005)	2010
interest	122.9%	168.0%	54.3%
knows	96.7%	95.0%	98.7%
dateOfBirth	3.0%	4.6%	0.4%
homepage	1.8%	N/A	0.6%
name	0.7%	N/A	1.3%

Wright do not provide separate statistics for LiveJournal and other data sources.

Obviously, the observed usage of FOAF is to a large degree the result of the involvement of LiveJournal. As the service provider, LiveJournal decides how users may express themselves in their user profile, what information has to be provided mandatorily at the time of registration, and what is ultimately publicized in RDF format. Still, the frequency of FOAF properties reflects to some extent what information is generally considered acceptable to share. While forming explicit connections with others, as represented by property "knows", has been constantly popular, there appears to be a general trend of reducing the disclosure of personal information. Ding et al. (2005) compare the property usage of LiveJournal to other sources of RDF data, and do not find a substantial qualitative difference. While the ranking of the most frequent properties differs, the focus is always on basic personal data and explicit social ties. Little data is available on the dissemination of extensions to FOAF. A study of Finin et al. (2005) briefly mentions "spouseOf" from Relationship as one of 512 properties that are used by 1% or fewer of the examined RDF documents.

These results raise new questions. An important use case of FOAF appears to be the declaration of relationships, yet FOAF by itself only knows a single, general type of social relationship. Richer ontologies of social relationships have been proposed, but failed to gain traction. In order to better understand what users of online services expect from an ontology of social relationships, we look at two studies that attempt to build such an ontology by "crowdsourcing", i.e., delegating the task to a community of laypeople. Liesenfeld (Liesenfeld, 2009; Hauffa et al., 2011) points out that, like FOAF, most online social network services are limited to a single relation that is often referred to as "friendship", but usually conveys a broader sense of acquaintance. While some services allow their users to organize their friends in named lists, this typification is not made public. Given that adding a contact to a specific list or group is comparable to assigning that contact to an ontological category, Liesenfeld hypothesizes that collaborative tagging (Trant, 2009) could be used for bootstrapping the vocabulary of an ontology.

A survey was conducted among users of Facebook, aiming to discover the language people use to describe social relationships. Users were asked to assign tags to their contacts that describe the relationship, i.e., name concepts that apply to the relationship. No constraints were placed on the number or format of the tags. 88 participants provided a usable response. 52% were between 20 and 25 years old, 30% were between 25 and 30 years old, the remaining 18% were older. 88% of the participants tagged fewer than 50 relationships. In total, 5 977

tags were assigned to 3 045 relationships, an average of 4.3 tags per relationship. 24% of the assigned tags are unique, which Liesenfeld attributes to spelling errors and the use of synonyms. We interpret the frequency of a tag as an indicator of the consensus among the participants about the utility or relevance of the tag for the description of relationships.

Table 2.3: Tags most frequently used by survey participants (Liesenfeld, 2009; Hauffa et al., 2011)

tag	frequency	characteristics
friend	57	emotional intensity
friend of friend	31	emotional intensity, social network distance
family	28	biological relation
acquaintance	25	emotional intensity
good friend	24	emotional intensity
best friend	21	emotional intensity
childhood friend	18	emotional intensity, continuity of relationship
colleague	16	social group
school	16	social group
cousin	14	biological relation
university	14	social group
fun	12	personal characteristics, shared activity?
partying	12	shared activity
high school	11	social group
like	11	emotional valence
neighbor	11	geographical distance, social group
dislike	10	emotional valence
fellow student	10	social group
high school friend	10	emotional intensity, social group
old friend	10	emotional intensity, continuity of relationship

The tags used by the survey participants, listed in table 2.3, describe different aspects of social relationships. For example, the tags “acquaintance”, “friend”, “good friend”, and “best friend” differ mostly in terms of emotional intensity. The presence of both “like” and “dislike” suggests that the pair is to be understood as a statement about emotional valence. Other tags name a shared activity that defines the relationship, such as “partying”, or a biological relation (“cousin”). Of special interest is the high number of tags that do not pertain to the relationship at all, and instead name characteristics of the person (“fun”, although this tag could also refer to a shared activity of “having fun”) or a social circle (“university”). Liesenfeld also observed less frequently used tags that represent a person’s function or role in a relationship, e.g., “babysitter”, and resources obtained through a relationship (“information”). She further notes that two people rarely use the exact same tags to describe their relationship, which implies that a relationship is rarely perceived symmetrically.

Participants frequently characterize their relationships in an indirect way: Grouping relationships by shared activities or social circles suggests homogeneity with respect to some

attribute, which is not always stated explicitly. One would have to analyze these groups to determine whether, for example, "university" refers to a specific type of relationship or different types of relationship that are enacted in the context of student life. Furthermore, the descriptions obtained by tagging are nuanced with regard to certain aspects (e.g., compare "childhood friend", "friend", and "good friend"), but superficial otherwise. Certain concepts, such as trust, are not represented at all, even though they are known to be directly relevant to the conceptualization of relationships (Golbeck et al., 2003). User-defined tags meet the users' needs of granularity and information disclosure, but do not provide insight into all aspects of the nature of relationships.

Huston and Levinger (1978) report on an earlier experiment on the classification of social relationships, where participants were tasked with identifying the nature of a relationship by asking a series of yes-or-no questions. About 95% of the questions were concerned with what Huston and Levinger call "general dimensions of relationship": kinship, emotional involvement, gender, age, and content of interaction. There is noticeable overlap between the way participants of Liesenfeld's study describe their relationships and the way participants of Huston and Levinger's study elicit information about the relationships of others. In both studies, the behavior of participants may be affected by a societal consensus on what kind of information about a relationship is acceptable to be disclosed in public.

Bossert (Bossert, 2010; Hauffa et al., 2011) quotes an observation by Hinde (1997), which points towards a different explanation for the results of the two experiments: "[...] whilst we may manage our relationships with moderate success, we are not always adept at pinpointing their special characteristics, describing them to others, or generalizing about them." The reason for the difficulty of describing a relationship may be rooted in its nature as a dynamic construct that arises from communication. The character of a relationship manifests in the use of language, most noticeably and directly in the choice of honorifics and forms of address. However, Argyle (1969), as cited by Bossert (2010) and Hauffa et al. (2011), claims that non-verbal communication is essential for conveying information about emotional aspects of a relationship, since verbal language is optimized for the exchange of factual information. The pervasiveness of emoticons and emoji in electronic communication is an indicator for the necessity of a non-verbal communication channel. Argyle (1975), as cited by Bossert (2010) and Hauffa et al. (2011), states that non-verbal communication can be perceived without "full conscious awareness", therefore it is possible that certain aspects of a relationship are not perceived in a fully conscious way either. Watzlawick et al. (1967), as cited by Bossert (2010) and Hauffa et al. (2011), state that a healthy relationship is defined by its participants without full awareness, while conscious effort is required to agree on a definition of a problematic relationship.

Furthermore, defining a relationship, regardless of whether it happens in private, as the result of internal deliberation, or as a public declaration, may have an effect on the relationship. For example, assigning a relationship to a category has the side effect of invoking all stereotypes, social norms, etc., which are associated with that category. McCall (1988) conceives of the type or class of a relationship as the result of a public negotiation with the social environment: once person a asserts to be in a dyadic relationship of a particular class with b, others assess to what extent the relationship "measure[s] up against the standards set forth in the cultural blueprints", which in turn may cause a and/or b to either justify

or adapt their behavior. The social environment may even be limited to the dyad itself: *a* proposes a relationship definition and *b* disagrees. Limited public disclosure may reflect an unwillingness to enter this negotiation process.

Bossert (Bossert, 2010; Hauffa et al., 2011) conducted follow-up interviews with 10 participants of Liesenfeld's survey. All interviewees see a risk in disclosing information about their relationships and consider it potentially harmful to their relationships. 80% are only willing to provide "superficial information". Many participants were personally acquainted with the interviewer, but only 60% would have provided more information to an entirely unfamiliar interviewer. 30% found it easy to describe their relationships, but only provided a coarse description on the level of the tags listed by Liesenfeld when asked for a demonstration. 50% would have preferred to describe relationships in their native language instead of English, because they felt the need to use a specific vocabulary for expressing fine distinctions. 80% would prefer to select from a list of predefined tags, but are generally aware that the choice of available tags might influence their expression. Some participants believe that explicitly and publicly describing social relationships puts them at a disadvantage, while some state they would perceive the necessity of keeping a public description up to date as a burden. Bossert concludes that even if someone has arrived at a conscious definition of a relationship, he or she may not be willing or able to express it in an accurate and detailed way.

In summary, we can identify three possible reasons for the difficulty of accurately describing a relationship:

1. Unawareness of certain aspects of a relationship that are not perceived consciously, and a lack of vocabulary to describe these aspects;

2. unwillingness to consciously define a relationship out of fear of changing it;

3. a desire for privacy or respect for the privacy of others, driven by a societal consensus on how to talk about relationships.

2.1.3 Affiliations, Tags, and Social Objects

Social scientists have long held that social structure does not arise from dyadic, interpersonal relationships alone. In addition to the relationships among a set of actors, their membership in "overlapping subsets such as voluntary associations, ethnic groups, action sets, and quasi-groups" (Foster and Seidman, 1982, as cited by Wasserman and Faust, 1994) contributes to what we perceive as higher-level social organization. The associations between actors and social groups can be represented by an affiliation network, which is a bipartite (or, equivalently, bimodal) graph. The vertices of a bipartite graph can be partitioned into two types of entities (modes), in this case actors and groups, and edges must connect entities of different types. Entities of the same type are linked by *co-occurrence*, that is, by having at least one common neighbor of the other type. The co-occurrences can be made explicit: Given the adjacency matrix A of a bipartite graph, the products AA^T and A^TA induce two edge-weighted, simple graphs, which are called *projections* of the bipartite graph. Edges in the projections

correspond to co-occurrences in the original affiliation graph. Projecting an affiliation network yields a graph of actors linked by co-membership, weighted by the number of common groups, and a graph of groups linked by overlapping sets of members.

These two graphs have intuitively appealing interpretations. Joint participation in group activities strongly indicates that two actors are either already acquainted or a relationship will eventually develop, so the first graph can be seen as a "potential social network". Overlap between groups enables transmission of information and coordination, and indicates similarity of the groups in purpose, values, and norms (Wasserman and Faust, 1994, pp. 293). In addition to group membership (You et al., 2015), two phenomena that can be modeled with affiliation networks are collaboration (Newman, 2001) and event attendance (Foster and Seidman, 1984, as cited by Wasserman and Faust, 1994). It can be argued that these are special cases of group membership, so the projections of their affiliation networks admit analogous interpretations. Newman and Park (2003) approach the analysis of social structure from the opposite direction. They observe that social network graphs differ from other types of network graphs in the sign of their degree correlation, and hypothesize that the positive correlation (assortativity: nodes prefer to attach to other nodes of similar degree) found in social networks is the consequence of a particular latent structure: An observable social network is the projection of an unobservable bimodal graph of actors and their communities. On that premise, graph clustering and community finding can be seen as attempts to recover the original affiliation network.

Considering the evidence for the presence of a connection between social structure and group membership, it stands to reason that a social relationship between two actors can be characterized in terms of their common group affiliations, either observed or inferred. However, the number of common affiliations is only meaningful in relation to the individual number of affiliations of the two actors. In order to compare two relationships, their co-membership counts have to be normalized. For similar reasons, the identity of common groups is not directly useful as the basis of a qualitative representation of a relationship. The fact that two people are, or are not, co-members of a group is only meaningful within the subset of actors that can — at least potentially — be members of that group. The potential group members are usually homogenous with respect to some attribute, which depends strongly on the group's nature and purpose, e.g., geographical closeness (a neighborhood sports club), distance in the social network (a circle of friends), or political stance. Additional information about the groups is necessary to make relationship descriptions derived from group co-membership comparable. The projected network of groups provides information about their similarity, but overlap in membership is not always a good indicator of similarity: Consider the case of a city that hosts two football clubs. Players, staff, and fans are almost dichotomously divided, yet, conceptually, the two clubs are more similar to each other than to other sports associations.

Explicit and reasonably complete information about group membership is rarely found in social media trace data, so we look at mechanisms for tagging shared resources as an alternative source of information about social relationships. Social tagging jointly creates an implicit social network and a common language to describe the relationships within that network. To understand how this works, one first has to understand in what sense tagging can be a collaborative process. Generally speaking, tags are words or short phrases that

are assigned to larger pieces of information as meta-data. A tag associates an entity with a category, thereby facilitating its retrieval. Tags are closely related to keywords, which have been a popular means of organizing resources even before the advent of electronic databases. In a database system, keywords and tags enable textual search for resources that are non-textual in nature or for which full-text indexing is not feasible. They also enable serendipitous discovery by browsing related resources, grouped by category. There are no generally accepted definitions that would cleanly separate tagging and keyword assignment, and the terms are frequently used interchangeably. Arguably, the main difference between keywords and tags is that keywords are applied to public resources to make them findable by others, while tags are used for the personal organization of resources. People assign tags that describe a resource from their own point of view, using language that feels appropriate to themselves, but when choosing keywords, they try to anticipate the information need and language use of others. Tags and keywords may contain information that cannot be directly obtained from the resources themselves.

The idea of tagging as a collaborative effort was first popularized by Delicious (also known as "del.icio.us"), a now-defunct service for sharing bookmarks, i.e., the titles and URLs of web pages. Users could assign tags to their bookmarks, and browse the bookmarks uploaded by themselves and others by tag. Since a bookmark is uniquely identified by its URL, the tag sets of different users can be merged, and the union of all tag sets that have been applied to bookmark can be seen as a collaborative description (Golder and Huberman, 2006). Marlow et al. (2006) compare a number of web-based services that let users apply tags to shared resources, and find that they differ substantially in the degree and nature of collaboration that is directly supported by the service. For example, users of the photo sharing service Flickr can selectively allow others to tag their photos. A currently popular online service that focuses on social tagging is Pinterest, where users can create "boards", themed collections of images found on the web, and browse the boards of others. Through a process called "re-pinning", users can adopt images from other boards, thereby placing them in new contexts.

On Twitter, an online platform for short-form public discourse (see section 3.1 for a detailed description), tagging has taken on a new role. As the popularity of Twitter grew, users were looking for ways to have continued conversations within the limitations of the medium. Inserting *hashtags*, keywords prefixed with the '#' sign, into messages ("tweets") emerged as the generally accepted mechanism of associating them with a particular topic or an ongoing discussion (Huang et al., 2010). Twitter subsequently began to index tweets by their hashtags, and continuously reports on "trending" hashtags that currently receive attention. A hashtag-based search feature lets users browse recent, topically related tweets. Hashtags blur the line between meta-data and content. Since they are part of the message itself, they are not only relevant for indexing and retrieval; their use also constitutes a communicative act. On a basic level, a hashtag is simply a statement or claim of association with a topic, but Daer et al. (2014) identify several "metacommunicative uses" like self disclosure, providing additional context, or expressing support for a cause. Bruns and Burgess (2011) observe that, especially in political discussions, hashtags tend to create discursive spaces, where people "deliberately engag[e] with one another". Other social networking services, most notably Facebook, have since adopted hashtags and provide related functionality that is inspired by Twitter (Daer et al., 2014).

Arguably, any system for the tagging of public resources, even if it does not directly allow for collaboration, implicitly creates a feedback loop (Zhang et al., 2006): By applying a tag, the tagged resource is placed in the context of other resources bearing that tag. The user can then decide whether the chosen tag fits and should remain as is, the tag should be changed, or additional tags should be applied. The outcome of this decision has the potential to influence subsequent taggers in their decisions. This mechanism gives rise to a community-filtered vocabulary that represents shared concepts and shared norms of talking about these concepts. The term "folksonomy" has been applied to these vocabularies to emphasize their emergence from the actions of a collective of laypeople, which sets them apart from ontologies constructed by domain experts (Mathes, 2004). Evidence for emergent conventions can also be found in earlier work on the extraction of keywords from academic writing: According to Frank et al. (1999, ch. 3.1), a term that frequently occurs within a document is more likely to be in the set of author-assigned keywords of that document if it frequently occurs as an author-assigned keyword of other documents from the same domain.

Mika (2005) formalizes the act of tagging by expressing it as a triple (a, t, r) with $a \in A, t \in T, r \in R$, which are the sets of actors, tags, and resources, respectively. The set of observed triples O can be expressed as a hypergraph with vertices $V = A \cup T \cup R$ and edges $E = \{\{a, t, r\} \mid (a, t, r) \in O\}$. Marlow et al. (2006) have independently proposed an equivalent representation as a bipartite multigraph of actors and resources with tag-labeled edges. Mika describes a scheme for decomposing the hypergraph into three bipartite graphs with weighted, regular edges. The graph G_{AT} of actors and tags is defined by vertices $V_{AT} = A \cup T$, edges $E_{AT} = \{(a, t) \mid \exists r \in R : \{a, t, r\} \in E\}$, and a weight function $w_{AT} : E_{AT} \to \mathbb{N}$ so that $\forall e = (a, t) \in E_{AT} : w_{AT}(e) = |\{r : \{a, t, r\} \in E\}|$. The graph G_{TR} of tags and resources and the graph G_{AR} of actors and resources are defined analogously. By projection of the three bipartite graphs, one obtains a total of six co-occurrence graphs. The graphs of tags that are used by the same actors and tags that are applied to the same resources contain the semantics that emerge from the actions of the community. In a case study of Delicious, Mika finds that the latter graph better reflects the actual conceptual relationships between the tags, while the former is a better indicator of the concepts that the community is interested in.

Just like tag co-occurrence indicates a conceptual relationship, actor co-occurrence may point towards the presence of a social relationship. There is evidence that tagging behavior is, to some extent, determined by the social environment (Rae et al., 2010). However, in contrast to the group membership setting discussed earlier, it is difficult to argue that actor co-occurrence in a virtual space corresponds to a high probability of mutual awareness and interaction. By intersecting the projection with an independently created social network graph, purely hypothetical edges can be eliminated. As before, the number of common tags or resources can, after appropriate normalization, serve as a numeric characterization of the relationship. The tags themselves, in conjunction with their semantic networks, are now useful as qualitative descriptors of the relationship, because they originate from a globally agreed-upon vocabulary.

Building affiliation networks or hypergraphs requires specific kinds of information. Affiliation networks require some notion of group membership, which can either be explicit or inferred from patterns in observed interactions. For example, repeated interaction of

multiple actors with a common object indicates collaboration, while an increased volume of interactions at a particular time and place indicates a social event. Tagging graphs are built from observations of discrete and uniquely identifiable labels being attached to public resources. Knorr Cetina's theory of *objectualization* (1997) may provide a direction for the generalization of these two approaches, so that they can be applied to a more general class of observed social interactions. Knorr Cetina argues that in modern society, objects simultaneously "displace human beings as relationship partners" and "increasingly mediate human relationships". These objects are "knowledge objects", characterized by being perpetually unfinished, incomplete, or unavailable, and therefore thought-provoking. An example for such an object is a scientific theory, which is repeatedly falsified by observations of reality, and subsequently reformulated to be consistent with these observations, but can never claim to fully explain observable reality. Knorr Cetina describes multiple ways in which person-object relationships resemble interpersonal relationships, but also notes that objects can act as focal points for the formation of interpersonal relationships. In parallel to communities and traditions, "objects may play a significant role in constituting such contexts [of belonging]", and therefore may be instrumental in "bringing about [social] integration" (Knorr Cetina, 1997).

In a frequently cited blog post, Engeström (2005) re-interprets Knorr Cetina's theory and applies it to online social networking services. Engeström claims that social relationships are necessarily mediated by shared *social objects*, which resemble the knowledge objects of Knorr Cetina. Social networks should therefore be understood as bipartite graphs of actors and social objects. Despite their structural equivalence to affiliation networks, there is an important difference in interpretation. A social group may or may not have an explicitly stated purpose, but it is always implicitly defined by its members. The opposite is true for social objects: the object is what brings people together, and membership in the resulting group is purely incidental. It follows that a social relationship can be meaningfully characterized by the social objects on which it is founded. Social objects as described by Engeström constitute a useful generalization of social groups in the bipartite affiliation model, and resources in the tripartite tagging model, but the underlying theory offers no guidance on how to identify the social objects that led to the formation of a particular relationship. Considering that social objects manifest in — and become observable through — digital artifacts that are shared and discussed on online platforms, extant empirical research on bipartite graphs of actors and different types of artifacts may point us towards salient social objects, for example, musical genres (Lambiotte and Ausloos, 2005), celebrities (Lim and Datta, 2012), or Twitter hashtags and the discussions that surround them (Bruns and Burgess, 2011).

2.1.4 Social Context

A fairly recent development in the social sciences is the understanding that relationships are dynamic and highly context-dependent constructs. Allan (2006) describes the role of context as follows: "[R]elationships do not occur in isolation, they are structured and framed at least in part by the broader contexts under which they develop, flourish, and eventually end." As an example of how social and cultural context affects people's perception of relationships, consider how the meaning of friendship changes in different contexts: Argyle

and Henderson (1985) note that members of the British working class tend to avoid using the term "friend", even when describing informal relationships with emotional investment. They also find differences in how men and women enact friendship: Women value intimacy, self-disclosure, and social support, while men prefer shared activities. Duck et al. (1997) expand on this example to point out how discrete relationship types, which are usually highly idealized and temporally static, are inadequate for expressing the nature of a relationship:

> *"[R]esearchers typically determine* a priori *the nature of a relevant relationship based on social norms associated with relational labels, and so it is assumed* by definition *that friendship will exclude sexual activity. [...] The social label that is ascribed to a relationship in part preordains a listener's expectations about the processes and sets limits to its features or to its 'fuzziness' by establishing a prototype."*

Generally speaking, the context of a relationship is formed by other relationships (Allan, 2006), including those of neighboring actors in the social network graph, and those that make up the higher-level social structures in which the relationship is embedded. SNA provides tools for the characterization of individual nodes in the context of a network graph, but offers little to learn about the role or function of individual edges: community finding characterizes nodes by their membership in densely connected subgraphs, block models identify groups of nodes that are similar in their connections to others, various centrality measures capture different aspects of a node's connectedness, but edge labels or weights are usually derived from external information. Constructing a *line graph*, where the role of nodes and edges is reversed, allows us to apply each of the aforementioned methods to the characterization of edges (Evans and Lambiotte, 2009). In a line graph, the edges of the original graph are represented by nodes and two nodes are connected if their corresponding edges in the original graph share an endpoint. While the utility of this approach has been demonstrated for community finding, further research is needed to determine if other methods of SNA can be meaningfully applied to line graphs.

Van Dijk (2008), building upon a range of concepts from cognitive and social psychology (Ginsburg, 1988), conceives of social context as a property of *social situations*, which he loosely defines as spatiotemporally separated environments in which dyadic or group interaction takes place. The situation includes everything that is objectively observable, relevant to the interaction, and temporally situated before or during the interaction. Conversely, the context is an individual mental construct of each participant, representing his or her subjective perception and assessment of the situation. The context can be pictured as a series of concentric circles around the situation. The circle closest to the center contains the immediate sensual perception of the situation. With increasing distance to the center, the information in the circles becomes less specific to the situation, shifting from perception to knowledge gained by means of generalization, abstraction, and decontextualization. For example, context may incrementally generalize from information about time, place, and participants of a situation to the intentions, goals, and social roles of the participants, and finally, to general societal norms and values.

If one "unfolds" a relationship into a temporal sequence of situations involving the two actors, each actor's context contains an individual representation of the relationship that

evolves with each new situation. The main mechanism behind the temporal evolution of contexts is the influence of context on language use. Since the communicative content of a social situation in turn affects the context of future situations, there is feedback between context and language, which manifests in dynamic, constantly updated contexts. Contextual information is rarely explicitly verbalized, but rather "signaled" via verbal (Ervin-Tripp, 1996) or non-verbal (Ambady and Rosenthal, 1992) cues, or "indexed", that is, implicitly expressed by reference to shared knowledge or experiences. Inferring an actor's subjective perception of a situation from language use is possible, as demonstrated by the successful applications of sentiment analysis and opinion mining (Pang and Lee, 2008).

A fundamental function of context models is judging the appropriateness of actions in a given situation, thus enabling the production of appropriate language. In consequence, it is possible to make inferences about a person's individual sense of appropriateness from his or her behavior: if person a says x, the context of a must be such that x is true or the act of saying x is otherwise appropriate. The consequences of a's action, i.e., the reactions of the other participants, provide feedback to a: if there is conflict, the contexts of a and the other participants must differ in a meaningful way. It is therefore possible to learn, by observation, about the shared conceptualization of appropriateness that characterizes the relationship.

The duality of context as observable reality (the "situation" of van Dijk) and mental participant construct is reflected by the different approaches to modeling social context in the field of social computing. One way of learning about the context of social interaction is inference about an actor's unobservable mental state from behavioral cues. Vinciarelli et al. (2009) use the term *social signal processing* (SSP) to refer to the automated detection and analysis of the behavioral cues people display in their interactions with others. A behavioral cue may carry meaning on its own (e.g., a facial expression of anger) or may be part of a temporal sequence of cues that possibly involves multiple actors (e.g., two people who turn towards each other, then smile) and has to be understood as a whole. Vinciarelli et al. list the following main categories of behavioral cues: physical appearance, gesture and posture, facial expression and gaze, vocal behavior (including prosody, turn taking, vocal outbursts, and silence), and the relative orientation and distance of multiple actors. A single unit of information about "feelings, mental state, [and] personality" (Vinciarelli et al., 2009), expressed via behavioral cues, is called a *social signal*. Social signals typically convey information about the mental state of individual actors, but some, like turn-taking and congruence of posture, directly describe an interaction.

SSP is limited to non-verbal cues, and is therefore framed as a joint problem of computer vision and the processing of audio and biometrical signals. It is hypothesized that most non-verbal cues are processed unconsciously, as opposed to verbal messages, which are filtered by additional cognitive processes (Vinciarelli et al., 2009). Social signals therefore have a more immediate effect on the perception of social situations, and can be said to contribute to the context of verbal communication. SSP research has long been focused on the analysis of "offline" social situations involving in-person interaction, but is increasingly being applied to mediated forms of interaction (Vinciarelli and Pentland, 2015). The medium of choice might either limit the range of applicable behavioral cues (e.g., to audio in the case of telephony) or necessitate a complete replacement (text messaging). For example, emoticons and avatars can be seen as "attempt[s] to transfer the social signals typical of face-to-face interactions

to the virtual world" (Vinciarelli and Pentland, 2015). Some online platforms invent new signals, such as "liking", others are successful even without specific affordances for social signaling. For example, uploading or sharing a picture acts as a non-verbal social signal with a communicative intent that is related to impression management, that is, influencing others' perception of oneself (Vinciarelli and Pentland, 2015). There is reason to believe that, independent of the medium of interaction, the information conveyed by social signals helps the recipient to understand the social situation.

Another way of learning about the context of social interaction is explicitly gathering information related to human social behavior, either actively, using sensors, or passively, by analysis of trace data. If we define the context of a dyadic relationship in the most basic way, as the set of socially close "co-relationships", we can learn about this set by reconstructing, from trace data, the different implicit and explicit social networks the two actors are members of. Matsuo et al. (2006a) discuss how to reconstruct social networks from trace data that is publicly available on the web, and how to merge them into a unified view. Following the lead of Van Dijk, a social relationship can be understood as a sequence of interactions, each embedded in an observable social situation. Sensors can provide information about these situations. A sensor can be a dedicated piece of hardware carried by an actor, a stationary measuring device, or a network of such devices. Modern mobile phones usually contain an array of sensors that can be queried by applications, including GPS geo-location, accelerometer, microphone, and camera, so they are popular platforms for data acquisition. In the following, we look at two large-scale studies that collect a broad range of contextual data about people's social behavior, and mainly use mobile phones as sensing devices.

The "Reality Mining" study (Eagle and Pentland, 2006) was the first attempt at gathering social context data using mobile phones. The researchers collected proximity (Bluetooth signal strength of other participants' devices) and location data (GPS), interaction data (call records), and application usage statistics from 100 participants. They find that the combination of different kinds of sensor data directly yields insights about the nature of social relationships: Simple time-stamped proximity data already allows to distinguish office acquaintances and friends by considering whether they usually interact inside or outside of working hours, and including geo-location data further increases the accuracy of classification. Building upon this study, the "Friends and Family" study (Aharony et al., 2011) involved the collection data from 130 participants. In addition to the measurements of the earlier study, cell tower and WLAN signal strength were measured for purposes of geo-location, accelerometer and compass readings were collected, the monitored communication activities also included text messages, and information about files stored on the device was collected. Participants could, on a voluntary basis, submit receipts and credit card statements, allow monitoring of their activity on Facebook, and complete personality tests and daily surveys designed to measure "mood, stress, sleep, productivity, and socialization" (Aharony et al., 2011). The collected data was used in an interventional study on the effects of encouraging physical exercise. In both studies, the collected data allows the reconstruction of various densely connected social networks, including self-reported networks of acquaintances and implicit networks of interaction online, by phone, or face-to-face.

The "Copenhagen Networks Study" (Stopczynski et al., 2014) was motivated by the multiplexity of communication channels in everyday life. In the course of its largest iteration

in 2013, data was collected from around 1 000 participants. The application deployed to the participants' mobile phones was the same as in the "Friends and Family" study, so the measurements obtained from the phones are the same as well. In addition, the participants had to complete surveys on personality traits, physical and mental well-being, and their social environment, and were required to consent to monitoring of their Facebook activity. Qualitative data was provided by an anthropological field study, which involved the observation of a randomly selected group of about 60 participants by a researcher, who was embedded in the group. By comparing the communication activity in three implicit networks (face-to-face, phone calls, text messages) at different times of day, the researchers are able to classify relationships according to which communication patterns are considered acceptable, e.g., meeting after work or on weekends, calling or texting late at night, or a general preference for one communication channel. Sekara et al. (2016) find that the high-resolution, longitudinal proximity data from this study facilitates the detection of social groups, enabling the identification of groups that are invisible to community detection in static social networks.

The analysis of rich sensor data points towards previously unexplored sources of information for the characterization of social relationships, mainly the multiplexity of communication channels and spatio-temporal patterns of interaction. However, this kind of data is difficult to obtain due to technical complexity and privacy concerns, as reflected by the low number of participants of the mentioned studies. Stopczynski et al. (2014) note that in "big data" studies of call detail records or tweets, participant counts on the order of 10^5 to 10^8 are not uncommon, while "creating larger [observational] studies, in terms of number of participants, duration, channels observed, or resolution, is becoming expensive using the current approach". To solve this problem, Lukowicz et al. (2012) call for an opportunistic system design: instead of sensors that are "purposefully deployed to satisfy the data needs", one should "exploit devices that just 'happen' to be in the environment." Stopczynski et al. similarly propose the opportunistic usage of data that is already being collected by SNS, media sharing platforms, fitness trackers, etc.

In summary, SSP and the use of sensor data complement each other: the former approach enables inference about the internal state of actors, while the other captures the observable parts of the social context. While there is consensus that context is important for the interpretation of social behavior, further research is needed to understand how longitudinal observations of social behavior, supported by rich context, can be distilled into simpler characterizations of social relationships.

2.2 Content-based Characterization

Orthogonal to the question of how to describe or represent a relationship is the question of where to obtain the information from which such a representation is constructed. As stated at the beginning of this chapter, our work on the characterization of relationships is motivated by the increasing availability of different types of communication trace data, which can be broadly divided into meta-data and content. Due to the inherent complexity of the semantic analysis of unstructured text, many existing studies that work with trace data only make use of meta-data, which is readily available in structured or semi-structured form,

or restrict the analysis of textual content to the extraction of linguistically shallow features. Using communication meta-data as the primary source of information places emphasis on the act of communication, while treating its content as ancillary. An example is the linear model of tie strength of Facebook users, developed by Gilbert and Karahalios (2009), which contains 74 independent variables, 35% of which correspond to features extracted from the unstructured text of wall posts and personal messages. The features are the frequencies of words from particular semantic classes (e.g., emotionally positive, work-related) in the text. The communicative content is therefore effectively reduced to the degree of its association with a number of hand-picked classes. In this section, we investigate if more efficient use can be made of the communicative content.

From the perspective of the social sciences, "communication is the central process giving shape to relationships" (Sillars and Vangelisti, 2006) via the principle of reflexivity: "communication both creates structure and is constrained by it" (Sillars and Vangelisti, 2006). This leads Parks (1997) to conclude that "[t]he information or content that flows through the network structure is, of course, at least as important as the structure itself." Sillars and Vangelisti (2006) identify two patterns of reflexive interaction between relationship and communication: On the surface, the style of communication reflects an agreement among the participants about the type of relationship they are in. For example, research of Goldsmith and Baxter (1996), as cited by Sillars and Vangelisti (2006), has shown that different types of relationships can be distinguished by the occurrence of certain speech events, such as small talk in more causal relationships and gossip in closer relationships. On a deeper level, the definition of a relationship is subject to a constantly ongoing process of negotiation. Watzlawick et al. (1967), as cited by Sillars and Vangelisti (2006), claims that every communicative act, verbal or non-verbal, has a literal meaning, but at the same time constitutes a proposition about the nature of the relationship. The proposition is either accepted by the other participant, or challenged by means of a counter-proposition. This negotiation is usually hidden in non-verbal and the subtext of verbal communication, and we are limited to observing the result, that is, the agreed-upon relationship definition, in terms of its effect on the communicative behavior. We may assume that the nature of a relationship not only affects the form of communication, but also its content, including, but not limited to the range of appropriate conversation topics, which lends some theoretical justification to the word class features used by Gilbert and Karahalios.

All of the previously discussed approaches to the characterization of social relationships suffer, to varying degrees, from a combination of two problems: bias in self-reports and self-presentation, and the need to reconcile the subjective viewpoints of the participants in a relationship. A number of social processes disincentivize people from honestly and accurately communicating their assessments of their relationships. Even if we assume that the true sentiment can somehow be inferred from observable behavior, it is still an individual perspective on a dyadic social construct, and therefore insufficient for its characterization. To address these issues, we propose a representation of social relationships that is based solely on the content of verbal communication. Pentland (2008), as cited by Vinciarelli and Pentland (2015), describes non-verbal social signals as "honest", because they are produced without conscious awareness and therefore do not undergo the same mental filtering as conscious communicative acts. We hypothesize that the aggregated content of communication,

in other words, the subjects, topics, or themes of a conversation, are honest for the exact opposite reason: they have passed the filters of both participants and represent their consensus on "what the relationship is about". This honesty is bought by limiting the scope of analysis to surface-level characteristics of the relationship.

We explore the viability of simple, content-based characterizations of social relationships. One of the most basic forms of content-based characterization is the extraction of keywords from communication trace data. We begin by surveying existing applications of keyword extraction to communication artifacts in section 2.2.1. As discussed earlier, there are two main use cases for the characterization of relationships. One is condensing a large amount of observational data into a concise representation that is accessible to the human analyst. In section 2.2.2, we let human judges evaluate the interpretability of keyword-based characterizations. The other use case is providing an intermediate representation, on which higher-level computational analysis can be performed. A minimum requirement for computational interpretability is that, within the chosen representation, relationships are mutually comparable. In section 2.2.3 we discuss to what extent this is true for keyword-based representations of social relationships, and how topic modeling can improve computational interpretability without sacrificing human interpretability.

2.2.1 Keyword Extraction

Due to its technical simplicity, we use keyword extraction as a starting point for the development of a content-based characterization. In order to understand how keyword extraction can be applied to communication artifacts, we first review the basic principles. The task of summarizing one or more documents by selecting the most relevant words is known as keyword extraction. By formulating keyword extraction as a machine learning task, one can choose from supervised and unsupervised methods. Since the definition of relevance is domain specific, supervised learning in this setting requires a training corpus of messages, in which all keywords relevant to the relationship have been manually identified. However, reports on the construction of corpora annotated according to similarly subjective criteria, e.g., the MPQA opinion corpus (Wiebe et al., 2005), indicate potential problems with annotator agreement. To attain a reasonable level of agreement, a thorough definition of keyword relevance in the context of social relationships is necessary. Such a definition would have to be grounded in sociological theory, empirically validated, and formulated in a way that is comprehensible to the annotators. To the best of our knowledge, such a definition does not yet exist, so we limit the scope of this survey to unsupervised methods.

The unsupervised extraction of keywords from a document is a two step process: First, a set of candidate terms (single words or *n*-grams) is constructed from the content of the document. In a simple keyword extraction system, this might be the set of unique nouns as identified by a part-of-speech tagger. Then, a saliency score is computed for each candidate. The assessment of saliency is usually treated as a domain neutral problem, although there is evidence that domain adaptation can improve performance (Frank et al., 1999; Hulth et al., 2001). Two basic methods of scoring are *tf-idf*, which is the ratio of the term frequency within the document (*tf*) and the proportion of documents from a larger collection that contain the term (document frequency, *df*), and *residual idf*, which compares the document frequency

of a term to the frequency predicted by a simple probabilistic model, usually a Poisson distribution. In *idf*-based scoring, the collection of documents serves as an implicit definition of term relevance. *Tf-idf* rewards terms that are specific to the document, i.e., terms that are frequent in the document under consideration, but rare in the other documents of the collection. The intuition behind residual *idf* is that the document frequencies of relevant and non-relevant terms follow different distributions (Manning and Schütze, 1999). Various methods of scoring have been proposed for the single document setting, where the document frequency is not available. An example is TextRank (Mihalcea and Tarau, 2004), which is described in detail later. Terms are ranked by saliency and discarded if their rank or their score is below a chosen threshold. The remaining terms constitute the keywords of the document.

Since each word (or, equivalently, *n*-gram) is treated as an independent entity, which is known as the "bag-of-words" assumption, the resulting set of keywords is invariant to the order of words in the text. Under the bag-of-words assumption, keyword extraction is similar to dimensionality reduction: Given a vocabulary of n distinct words, a document can be expressed as an n-dimensional vector of word counts. The union of keywords of the documents in a collection induces a new, lower dimensional space, where a document is represented as a binary vector that only encodes the presence or absence of a word in the document's keyword set.

Within this framework, we can obtain lower-dimensional representations of social relationships by aggregating the communication artifacts associated with each relationship into separate documents, and performing keyword extraction on the resulting collection. The use of *idf*-based scoring implies that keyword relevance is determined by comparing language use in one relationship to a broad social context formed by the other relationships. This keyword-based characterization neither has a direct sociological interpretation, nor is it built on a mathematically principled way of dimensionality reduction. Still, the popularity of "tag clouds" in social media indicates that sets of keywords can be effectively visualized and are considered meaningful by humans. Finding relevant keywords that describe a relationship covers the middle ground between supervised learning of a few well-defined concepts (e.g., "professional" and "personal", as seen in section 2.1.2) and unsupervised clustering, which yields arbitrary natured, unlabeled clusters.

The idea of attaching keywords to social relationships has already been explored in a series of studies: Matsuo et al. (2006b) present a system for finding and quantifying arbitrary relationships between individuals on the Web. The system estimates relationship strength from the number of results of specially crafted queries to a web search engine. Mori et al. (2007) extend this system to perform web searches for the actor's names and annotate the relationships with highly ranked keywords from the search results. The keywords are used for clustering the social relationships, and the clusters are evaluated against a manually labeled reference dataset. A related, earlier study by Mori et al. (2004) applies similar methods to augment FOAF descriptions of persons and their relationships with keywords. The earlier study does not evaluate the keywords beyond basic visual inspection, while the later study uses the standard information retrieval metrics precision and recall. Our study complements those of Mori et al. by having human judges perform a subjective assessment of the quality of whole keyword sets.

2.2.2 User Study on the Utility of Keyword-based Characterization

Based on the earlier discussion, we formulate the following research question: It is possible to extract a set of keywords from communication artifacts that is considered a good representation of the associated social relationship by a human observer? The intuition is that, given enough data, a pattern of keywords specific to the relationship will emerge from the background noise of words with mostly conversational function. We apply a keyword extraction system to textual messages exchanged between users of the social network service Facebook. The goal is to find a set of keywords which a human observer would describe as "accurately characterizing the relationship". Evaluation is carried out via a Facebook application that analyses private messages and obtains user feedback on the quality of the extracted keywords.

In order to obtain a representative sample of interpersonal communication, a large amount of messages has to be collected. We use the social network service Facebook (see section 3.2 for a detailed description) as a data source, because it is known for a large user base and network effects ("viral marketing"). Facebook offers an API that allows third-party developers to integrate their applications into the user interface and access user profile data as far as permitted by an individual user's privacy settings. The Facebook application developed for this study performs three distinct tasks: The private messages sent and received by a user are collected via the API, a set of keywords is extracted from the collected messages for each sender-recipient pair, and finally the user is asked to provide an assessment of the quality of the keywords.

The "Talk Doctor" application is designed as a virtual advisor that provides the user with communication statistics as an incentive to use the application and recommend it to other users. Like the underlying keyword extraction system, the user interface is bilingual (German and English). The language is chosen according to the user profile settings. After launching the application, the first screen describes the setting of the survey: "Imagine you're trying to represent the relationship between two people by a few keywords." The following screen lists the user's contacts by name, sorted by message count to emphasize the most salient contacts. When the user selects a contact, keyword extraction is performed on the messages exchanged between the user and the chosen contact. Finally, the screen shown in figure 2.1 appears, visualizing three sets of keywords generated by different scoring algorithms. We chose to generate 20 keywords for each contact to keep the visualization simple and not to overly strain the attention span of the participant. Keywords are displayed in order of their relevance score, which is also reflected by the font size. By clicking a keyword, the user can view it in the context of the messages in which it occurs. Keywords deemed completely irrelevant can be removed. Located below each set of keywords is a slider for rating the quality of the set on a scale from 0 to 100. The slider's handle is a smiley face that changes its expression while being moved. A rating of 100 means that each of the keywords contributes to an accurate representation of the content of the relationship. Lower values indicate a higher proportion of unrelated keywords or a generally lower quality of representation. Using terms of information retrieval, this assessment is closer in meaning to precision than to recall. We are more interested in how well the extracted keywords describe the relationship than in how comprehensively they cover its different facets, since we expect

that most relationships can only be partially observed through online communication.

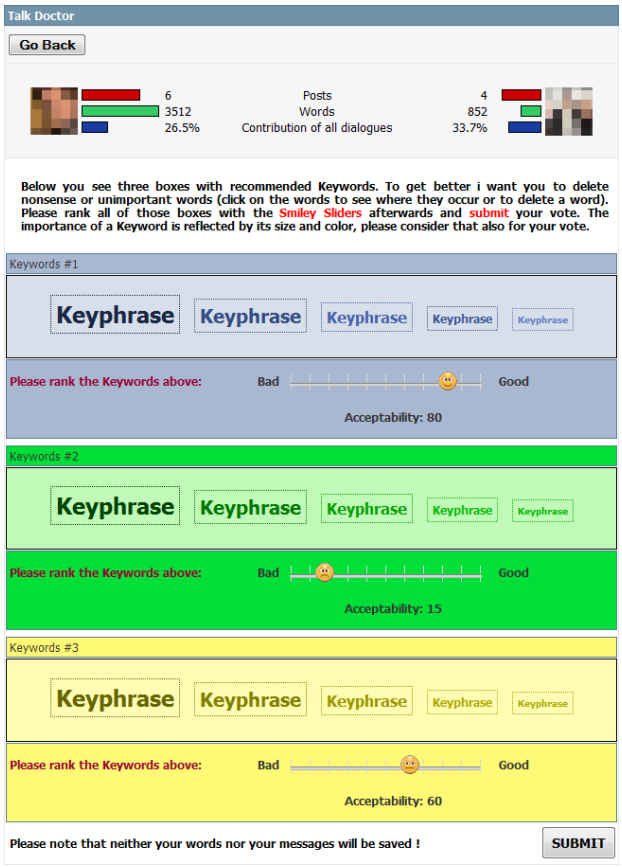


Figure 2.1: The keyword assessment screen

At the time of data acquisition, Facebook provided both long-form, delayed messaging, comparable to e-mail, and real-time chatting. Even though the API offers access to both communication systems, chat messages are ignored due to their conceptually different nature. For the purpose of keyword extraction, all messages of a user pair (the account owner and the chosen alter) are aggregated into one document. The reference corpus for computing *tf-idf* is constructed by aggregating all other messages of the account owner into separate documents by user pair. Incoming and outgoing messages are collected without distinction, so the extracted keywords are associated with an undirected relationship edge. Messages with more than one recipient are discarded, to ensure that all processed messages are equally significant for the relationship. If the number of suitable messages is insufficient, the user is notified and asked to choose another contact.

Measures are taken to protect the participants' privacy: Facebook requires that each application that wants to access private messages is submitted for review and whitelisting. When installing the application, the user has to explicitly grant access to private messages. All communication between the application server and Facebook takes place over TLS en-

crypted channels. The "Talk Doctor" application does not retain copies of the messages after processing, and the extracted keywords are encoded by a MD5 one-way hash function before being stored. The hashes are used in place of the original words in all further processing and evaluation. While the hashing is crucial for preserving privacy, it also limits our ability to evaluate and visualize the results.

The initial set of participants was recruited from the acquaintances of the organizers of the study (Hauffa et al., 2012) and by advertising on university noticeboards and topically appropriate Internet discussion boards. Participants were requested to publicize the application among their Facebook contacts. This amounts to an accidental sampling scheme. As an incentive for participation, all participants were entered in a raffle for a portable media player.

Implementation of Keyword Extraction

The process of unsupervised keyword extraction can be divided into four steps:

1. *Preprocessing:* Tokenization and any per-token processing, e.g., part-of-speech tagging, that is required by the later steps.

2. *Candidate Selection:* Statistically or linguistically motivated filtering to limit the set of candidates to words with a high prior probability of being relevant.

3. *Scoring and Ranking:* Assignment of a numerical rating to the candidates, removal of low-ranked words from the candidate set.

4. *Keyphrase Formulation:* Identification of multi-word expressions composed of highly ranked words.

Although a number of implementations of state-of-the-art keyword extraction systems are freely available, we decided to re-implement three algorithms from Hasan and Ng's survey (2010) to be able to adapt each part of the resulting system to the domain of online social interaction. A small number of preliminary experiments were conducted on data from individual relationships, with the purpose of tuning the system's implicit and explicit parameters. The strategy was to start off with a language- and domain-neutral keyword scoring algorithm and add pre- and post-processing steps as required. In these experiments, straightforward implementations of the keyword extraction algorithms without any further processing performed poorly, and did not produce results that were interpretable in the context of social interaction. Linguistically motivated, and thus language-dependent, pre- and post-processing noticeably improved performance.

Language identification is a requirement for any further linguistic processing. Before any actual processing takes place, all messages not in English or German language are discarded by an n-gram based language classifier. The complexity of implementation rises with the number of languages to be supported, so we restrict the system to the languages most likely to be used by the participants. For each language, a separate 3-gram classifier was trained on the Europarl corpus (Koehn, 2005). Tokenization and part-of-speech (PoS) tagging are performed using the Stanford PoS tagger (Toutanova et al., 2003). PoS tags are used for

candidate selection and as a means of word sense disambiguation. Gimpel et al. (2011) find that the Stanford PoS tagger, being trained on newswire text, performs worse when applied to Twitter messages. We experience a loss in performance consistent with the results of Gimpel et al., which can be attributed to stylistic features not present in the training corpus, mainly words containing punctuation characters (e.g., e-mail addresses) and emoticons.

Candidate selection is performed by a sequence of three filters: A PoS tag filter discards all words that are neither nouns nor adjectives, motivated by the results of Mihalcea and Tarau (2004). A stop word filter discards words that are known to be of little relevance. We augment the stop word lists of the KEA project (Jones and Paynter, 2001) with domain specific vocabulary, such as abbreviations commonly found in social media text. Informal online communication often contains unique spelling variants of stop words that would receive an inappropriately high score by *idf* weighting. To deal with these variants, the stop word filter accepts regular expressions, which are especially useful for filtering out elongated words and words where letters have been substituted with digits of similar shape. For example, `(l|L)+(o|O|0)+(l|L)+` matches variants of "lol" ("laughing out loud"). Finally, a set of heuristics discards very long words ($\geq$ 30 characters), words with a ratio of length to number of unique letters that exceeds a threshold of three, and words containing punctuation characters commonly used as part of emoticons. The remaining candidate words are reduced to their word stems via the Snowball library (Snowball developers, n.d.), which contains an implementation of Porter's algorithm for English and comparable algorithms for other languages. From this point on, two words are considered equal if their word stems and PoS tags are the same. The original word forms are kept for displaying the word to the user.

Three methods of keyword scoring are being compared: *tf-idf*, TextRank (Mihalcea and Tarau, 2004), and a custom variant of TextRank that operates on directed graphs. TextRank is an application of PageRank to a graph constructed by treating words as vertices and adding an edge between two words if they co-occur within a specified distance ("window size"). As social media text frequently contains non-standard punctuation, we deviate from the original algorithm by adding edges between co-occurring words even if they are separated by punctuation characters. TextRank is applied to an undirected and unweighted graph constructed with a window size of two. The damping factor is set to 0.85 and the convergence threshold is 10^{-5}. We also test a variant of TextRank that operates on a directed graph generated according to the rules set forth by Litvak et al. (2011). Directed edges are added between words that occur in direct succession. For both graph-based methods the maximum number of iterations is set to 100 to place an upper bound on processing time.

For the computation of *tf-idf*, we define the inverse document frequency (*idf*) of a word w in a non-standard way as $\log(D/(1 + df_w))$. The document frequency df_w is the number of individual messages, before aggregation into documents of the reference corpus, that contain the word at least once, while D is the number of documents in the reference corpus. This adaption of the *idf* formulation is due to the observation that making df_w grow more slowly produced better keywords. The graph-based methods of scoring do not use the *idf* weighting scheme, so preliminary experiments were performed to test the utility of an *idf*-like heuristic adjustment of word scores as an additional post-processing step: If the ratio of document frequency to word length exceeds a threshold of three, the score is lowered.

This heuristic penalizes short words that occur frequently in the whole corpus. Unexpectedly, it consistently improved the perceived keyword quality, even in conjunction with *tf-idf* scoring, so the heuristic is also used in the main experiments.

Some concepts cannot be appropriately represented by a single word, e.g., place names like "New York City". Therefore, in a final processing step, words that are adjacent in the original document are combined into an n-gram keyword (keyphrase) if certain conditions are met. First, a list of all sequences of candidate words that occur in the messages is complied. The score of a sequence is the harmonic mean of the scores of the constituent words. Using the harmonic mean ensures that keyphrases are only constructed from words that are good keywords on their own, which effectively penalizes longer phrases. From the list of keywords ranked by score, the n highest ranked words are chosen as a representation of the set of exchanged messages and therefore of the relationship as a whole. Multi-word expressions are displayed as they occur in the original message. Word stems that correspond to multiple different word forms in the original messages are represented by the word form with the lowest Levenshtein distance to the stem.

Systems for keyword extraction are usually evaluated by comparing the set of extracted keywords to a manually compiled reference set in terms of precision and recall. However, Turney (2003) remarks that a particular document might be represented equally well by more than one set of keywords, and recommends having the output of the system rated by human judges. This recommendation is consistent with the results of Jones and Paynter (2001), who find a statistically significant agreement on the quality of keywords between different human assessors. Barker and Cornacchia (2000) attribute this effect to keyword coherence: "Judges did not prefer keyphrase sets based simply on the individual keyphrases they contained. A set of keyphrases is somehow more than the sum of its individual keyphrases." Furthermore, as discussed in section 2.1.2, generating an appropriate set of reference keywords is likely to be difficult: An individual's assessment of a relationship tends to be subjective, so people asked to come up with labels for a relationship will focus on functional aspects rather than conversation topics, e.g., emotional intensity ("good friend", "best friend"). For these reasons, the system is evaluated by presenting the participants with the results of different methods of keyword extraction and asking them to rate the quality of each set of keywords on a numeric scale.

Results

Data was collected over a period of approximately 2.5 months. During that period, 98 Facebook users installed the application, and 71 users actually submitted usable data. Assessments were submitted for 275 relationships. The average number of messages associated with a relationship is 20, with an average number of 36 words per message. There is a strong linear correlation between the volume of sent and received messages per relationship, with Pearson's correlation coefficient $\rho = 0.92$ for the number of messages and $\rho = 0.90$ for the word count. This is evidence for reciprocal messaging behavior on Facebook.

We separately analyze the individual assessments and the means of all assessments submitted by one user, to visualize the effect of per-user preferences. The main results are summarized in table 2.4, which lists mean and standard deviation of the assessments for

Table 2.4: Human quality assessment of the keyword extraction system (in percent of a perfect score; Hauffa et al., 2014)

	TextRank (undirected)	TextRank (directed)	*tf-idf*
avg. per relationship	72.43%	67.76%	72.07%
avg. per user	69.86%	65.20%	69.73%
std.dev. per relationship	19.88	20.13	21.56
std.dev. per user	17.51	18.04	20.18

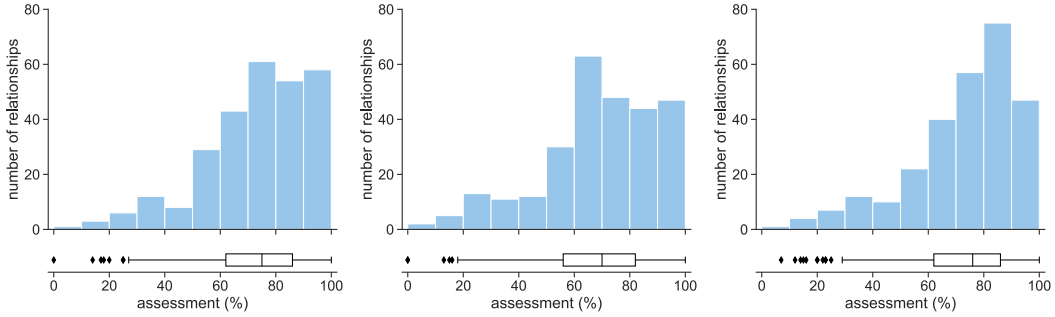
each method of keyword scoring. Figures 2.2 and 2.3 show the distribution of assessments for each method of keyword scoring. In the box plot below the histogram, the whiskers correspond to the 1.5 interquartile range (IQR).

Participants were encouraged to delete keywords they consider to be completely irrelevant. 46 users, 65% of the participants, made use of the facility for removing irrelevant keywords at least once, removing 2.8 keywords on average. As expected, there is a consistently negative correlation between the quality assessment and the number of removed keywords. Comparing the set of words that had their score lowered by the additional *idf*-based heuristic to the set of words deleted by more than one user, we find an overlap of 45% for the German language. Results for English are not representative due to a lack of data. No meaningful correlation exists between the assessment of a relationship and the length of the conversation (word or message count).

The results of the keyword quality assessment are encouraging and show potential for further development. An average quality assessment around 70%, regardless of the algorithm used, indicates that the participants did indeed see value in the selection and presentation of keywords. The frequent deletion of irrelevant keywords indicates that there is still potential for improvement. The distribution of assessments is bimodal: each algorithm produces a small group of keyword sets that are rated noticeably worse than average. While the most highly rated keyword sets of all three algorithms come from the same set of relationships, this is not true at the lower end. Each algorithm has its own "problem cases", for which it generates low-quality keywords, while the other algorithms perform better. Overall, undirected TextRank performs best, but due to the small differences in average performance, the low sample size, and the observation that each algorithm has problems with a different subset of the data, it is not possible to definitely recommend one algorithm over the others. The average per-user assessments show a broad consensus, but a small set of people, different for each algorithm, consistently give low assessments. There appear to be a few cases of distinct personal preference against the results of a particular algorithm.

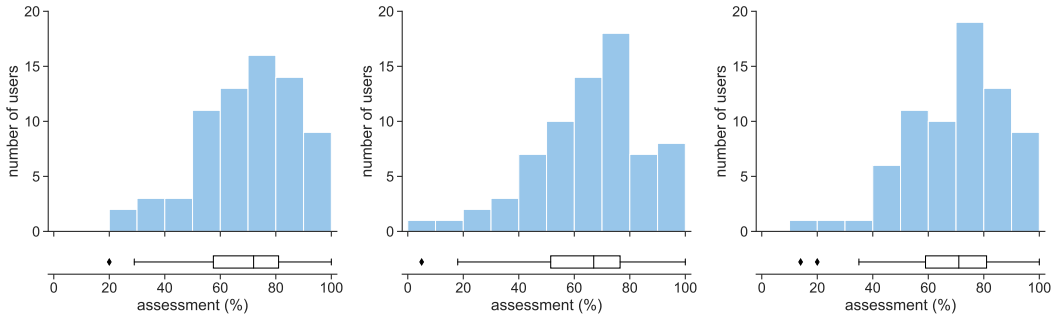
2.2.3 From Keywords to Topics

Given the evidence that keywords extracted from the communication of two individuals convey information about their relationship to human readers, we ask if keywords are also suitable as an intermediate representation for further computational processing. Keyword



(a) TextRank (undirected) **(b)** TextRank (directed) **(c)** *tf-idf*

Figure 2.2: Histograms of the human assessments of keyword sets



(a) TextRank (undirected) **(b)** TextRank (directed) **(c)** *tf-idf*

Figure 2.3: Histograms of the human assessments of keyword sets, per-user average

extraction suffers from a problem known as the *vocabulary gap* (Liu et al., 2012), which refers to a situation where terms that would be appropriate keywords do not occur in the document itself. This problem appears to be highly domain specific: Turney (1999) performs keyword extraction on an e-mail corpus and finds that on average 97.9% of keywords proposed by human annotators occur in the text, compared to 65.3% for a corpus of web pages. The vocabulary gap is indicative of a larger issue, a divergence between the vocabulary used within a document and the vocabulary that is used when talking about the document and its contents. Studies on free-form tagging of web resources (Li et al., 2008) have shown that users prefer a smaller number of high-level abstractions over a complete coverage of all concepts as they occur in the document.

The divergence of content and keywords poses a much bigger problem to the computer than it does to the human analyst. In the user study, if the human judges perceived the keyword sets as accurate representations of their relationships, this is because they were able to interpret the keywords using their knowledge about their relationships and the world in general. Except when working with a small, controlled vocabulary, algorithmically extracted keywords are unlikely to match up exactly with the manually assigned keywords, so

that information about semantic similarity from an external source is required to evaluate the degree of overlap of the two sets. The availability of semantic information is an important requirement for the automated analysis of keyword sets, and therefore, for their utility as an intermediate representation. This information can come from general-purpose semantic databases like WordNet (Miller, 1995), but also from word co-occurrence networks that represent the latent semantic content of a document collection, as discussed in section 2.1.3. WordNet is the product of manual curation, and therefore highly accurate. Latent semantics provide domain- and context-specific information that would be difficult to impossible to obtain elsewhere, but by definition cannot place the documents in a broader semantic context.

Another problem is that, as demonstrated in the user study, the basic keyword extraction algorithms require domain-specific tuning to produce acceptable results when applied to Facebook communication artifacts, and the results are not necessarily transferable to other computer aided communication settings, even ones that are similar in concept. This problem is exacerbated by the introduction of linguistic processing steps, which require training and customization for each language that is to be supported. None of the keyword scoring methods discussed earlier are part of a framework that would allow for controlled domain adaptation and tuning. As an example for the possibilities and limitations of domain adaptation, we briefly examine the adaptability of *tf-idf*, a method of keyword scoring that performed well in the user study.

Robertson (2004) reviews the literature and finds that while *tf-idf* was originally developed as a heuristic for choosing words with maximal discriminative power in a setting where a particular document is to be identified among others, the *idf* formula also occurs as a special case of multiple methods for relevance-based ranking of documents when no external relevance information is available. Even though *tf-idf* is part of a larger methodological framework, this does not provide us with guidance on how to adapt *tf-idf* to our use case. The only tunable parameter, and therefore the only way of incorporating additional information into the computation of the *tf-idf* score, is the composition of the document collection. Since the collection implicitly defines what is relevant, the goals of keyword extraction from a document can be adjusted by placing the document in different contexts. For instance, the social network structure could be incorporated into the keyword extraction as follows: Under the hypothesis that tightly connected groups of enthusiasts tend to be focused on a particular subject matter that is much less relevant to outsiders, one could identify communities in the social network and compare word scores within and outside these communities to recover some information on semantic relatedness.

At this point, our proposed system for the characterization of social relationships involves three distinct processing steps that so far have been treated as independent: keyword extraction, construction of a shared vector space of relationships, and defining a measure of similarity that makes use of latent and/or external semantic information. Is it possible to perform these three steps jointly, and in a mathematically principled way that leaves room for further domain adaptation? We are looking for a low-dimensional representation of textual content that retains the interpretability of keyword sets, but also facilitates comparison in terms of semantic similarity. In the following, we compare three methods for representing documents in a lower-dimensional semantic space built from latent semantic information

provided by word co-occurrence, and evaluate their suitability for the task at hand. Figure 2.4 provides an overview on the different methods.

Under the bag-of-words assumption, a document that contains terms from a vocabulary of size W can be expressed as a W-dimensional vector of term frequencies, and a collection of documents with indices $1 \dots M$ as a $W \times M$ matrix A. This document-term matrix A can either be interpreted as a representation of documents in a high-dimensional vector space, or, in analogy to affiliation networks (see section 2.1.3), as the adjacency matrix of a weighted, bimodal graph. Its unimodal projections $A^T A$ and AA^T represent the relatedness of documents and terms, respectively. Latent semantic indexing (LSI; also known as latent semantic analysis, LSA) is a method for dimensionality reduction of vector space models of text (Deerwester et al., 1990). By means of singular value decomposition (SVD), A can be expressed as the product $U\Sigma V^T$, where U is the $W \times m$ matrix of left-singular vectors, V the $M \times m$ matrix of right-singular vectors, and Σ the $m \times m$ diagonal matrix of singular values, with $m = \min\{W, M\}$. The columns of U and V are linearly independent and therefore can be interpreted as latent *factors* or distinct "artificial concepts" (Deerwester et al., 1990), and the rows as per-document and per-term factor weights, indicating to what degree a document or term is associated with each factor. Conversely, a factor is characterized by the terms and documents it is strongly associated with. In this way, documents and terms are represented in a common, m-dimensional vector space.

Each singular value is associated with a particular column or factor, and can be understood as the contribution of the factor to explaining the data, but also as its importance for accurately reconstructing the original matrix A. It is possible to show that the best approximation, in a least-squares sense, of A with a rank of $K < m$ can be obtained by setting all but the largest K singular values to zero. Equivalently, one can construct submatrices U_K, V_K, and Σ_K by removing the corresponding columns from U and V, and the corresponding rows and columns from Σ, to obtain the approximation $A \approx A_K = U_K \Sigma_K V_K^T$. The dimensionality of the vector space into which documents and terms are embedded can thus be arbitrarily reduced, as shown in the first row of figure 2.4. The notions of document and term similarity represented by the projections $A^T A$ and AA^T can be approximately recovered from the truncated singular vectors. Since $A^T A \approx A_K^T A_K = V_K \Sigma_K^2 V_K^T$, the similarity of documents i and j is approximated by the dot product of the i-th and j-th row of the matrix $V_K \Sigma_K$. A formula for the similarity of terms can be derived analogously.

LSI creates a lower-dimensional "semantic space", in which the distance between two documents is low if they are similar to each other, with similarity being defined by term co-occurrence. Like keyword extraction, LSI can be applied to the analysis of social relationships by aggregating the communication artifacts of the relationships into documents and performing LSI on the resulting collection. The main issue with LSI-based vector representations is that their components express positive or negative association with the different factors, which are expected to correspond to semantic concepts, but the nature of these concepts is opaque. Deerwester et al. (1990) acknowledge that the factors are not amenable to interpretation or verbal description. Xu et al. (2003) attribute this to two differences between the factors of LSI and human understanding of concepts: First, a factor can be negatively associated with certain terms, which contradicts the intuitively appealing notion that combining basic concepts to form more complex concepts is a purely additive process. Second,

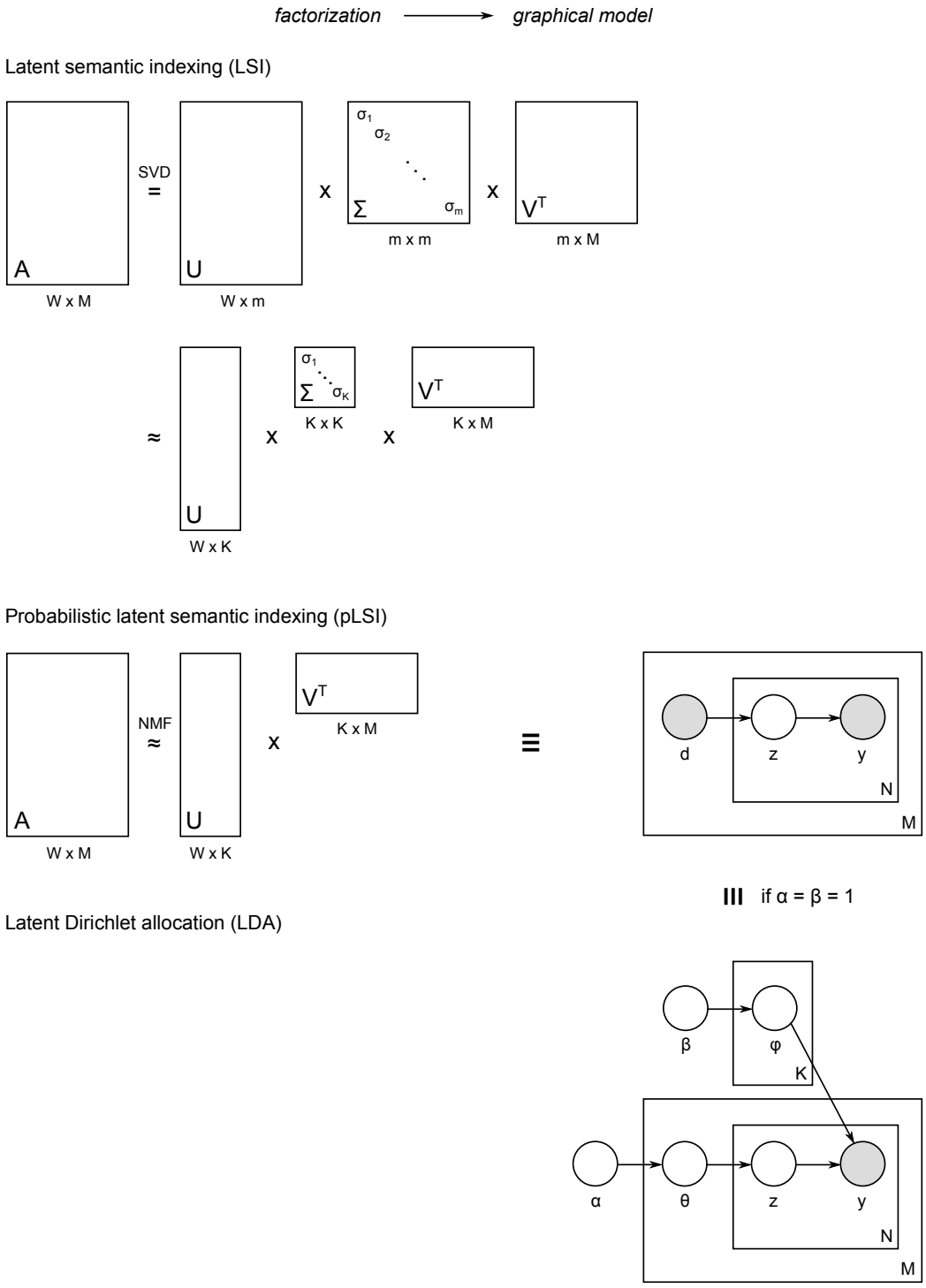


Figure 2.4: Three methods for the representation of text documents in a low-dimensional semantic vector space and their interpretation as matrix factorization, probabilistic graphical model, or both, if equivalent (adapted from Deerwester et al., 1990 and Blei et al., 2003)

SVD forces factors to be perfectly linearly independent, while concepts may overlap. Therefore, by moving from keyword extraction to LSI, we gain a data-driven notion of semantic similarity, but reduce human interpretability.

Non-negative matrix factorization (NMF) addresses both of these interpretability concerns. NMF is a method for directly computing a lower-rank approximation of a matrix by factorization, with the constraint that the elements of the resulting matrices are non-negative (Xu et al., 2003). For a given K, the document-term matrix A is approximately decomposed into a $W \times K$ matrix U and a $M \times K$ matrix V so that $\|A - UV^T\|_F$ is minimal. Since the elements of the two matrices are non-negative, and normalization can be achieved by multiplication with an appropriate diagonal scaling matrix (Gaussier and Goutte, 2005), the representations of documents and terms have a probabilistic interpretation: a document is represented by a categorical distribution over the K factors, while the factors can be viewed as categorical distributions over the vocabulary. The conditional probability $p(z|d)$ of a factor z contributing to document d and the conditional probability $p(y|z)$ of observing a word y given a factor z form a categorical mixed-membership model (Airoldi et al., 2014) that describes the relationship between a document and its words. This relationship can be expressed as a generative process (Hofmann, 2001): Assume, without loss of generality, that we want to generate M documents with an expected length of N.

1. For each of $M \cdot N$ words:
   a) Select a document d with probability $p(d)$.
   b) Choose a factor z with probability $p(z|d)$.
   c) Generate a word y with probability $p(y|z)$.

This model is known as probabilistic LSI (pLSI, sometimes pLSA; Hofmann, 2001). NMF and pLSI are equivalent in the sense that a factorization that minimizes the Kullback-Leibler (KL) divergence to the original matrix can be transformed into maximum likelihood parameters of the mixed-membership model and vice versa (Gaussier and Goutte, 2005). The second row of figure 2.4 compares NMF to the graph structure of the pLSI model in plate notation. Per convention, filled circles correspond to observed variables, hollow circles to latent variables, and arrows indicate conditional dependency, so that, conversely, the absence of an arrow can be understood as an independence assumption of the model (Buntine, 1994). The rectangular "plates" are a shorthand notation for repeated structures, with the number of instantiations given in the lower right corner. An arrow that crosses a plate boundary connects the outside node to each instantiation of the inside node.

The class of models that represent documents as probability distributions over a discrete set of factors or concepts, which are in turn distributions over a vocabulary, has come to be known as *probabilistic topic models*. The discrete factors are called *topics*, as they are close to the human conceptualization of topics of discourse. The interpretability of topic models has been empirically confirmed by a large-scale user study of Chang et al. (2009b), who conclude that "[h]umans appreciate the semantic coherence of topics and can associate the same documents with a topic that a topic model does." Generally, topical representations of documents are at least as expressive as keyword sets, considering that a set of plausible

keywords can be obtained by repeated sampling from the document-topic and topic-word distribution. To determine the semantic similarity of two document-topic distributions, any similarity measure for probability distributions can be used, e.g., KL divergence and its symmetrized variants.

Latent Dirichlet allocation (LDA) is a fully Bayesian variant of pLSI that explicitly models the document-topic and topic-word distributions as latent variables with Dirichlet priors (Blei et al., 2003). By setting the parameters of these prior distributions appropriately, the sparsity of the latent variables can be promoted. Sparsity is usually considered a desirable property; for example, highly sparse topics, which assign a probability of zero to most words, represent cleanly delineated concepts that are easier to visualize and interpret. The *maximum a posteriori* parameter estimate of an LDA model with uniform priors, which have no effect on sparsity, is equivalent to pLSI (Girolami and Kabán, 2003). An arguably highly important contribution of LDA is that it showcases the power of directed graphical models for unsupervised learning from unstructured text, while featuring "generative stories" and plate diagrams as a mathematically rigorous, but accessible language for expressing the model assumptions. This has sparked expansive research on special-purpose topic models, some of which are discussed in the survey of Blei (2012).

One can see that LDA-based topic modeling meets all of the initially stated requirements: it jointly represents documents in a low-dimensional vector space and offers a measure of semantic similarity, its vector-space representations are at least as interpretable as keyword sets, and it is part of a mathematical framework that allows for principled extension and adaptation. When applied to the characterization of social relationships, it also enjoys all the general advantages of content-based analysis: the representation is built from directly observable communicative behavior, and the topics of conversation reflect the participants' consensus on the nature of the relationship.

In order to build a complete system for the analysis of communication on social media, it is necessary to understand the differences between the popular online social platforms and the forms of communication they offer. Therefore, the next chapter is devoted to a detailed exploration of three platforms, the acquisition of communication data from these platforms, and a descriptive analysis of that data. In chapter 4, we return to topic modeling, describe the technical details of LDA, and discuss how topic models need to be adapted to account for the peculiarities of social media data. One question remains: Does a representation of social relationships that is considered accurate by humans also improve the performance of a social computing system? The study of social influence in chapters 5 and following internally uses a topical representation of communication data, and therefore can be viewed as an extrinsic evaluation of topic modeling as a means of content-based characterization.

3 Online Communication Data

The software systems for collecting data from Twitter and Facebook and for processing collections of e-mail were initially designed and developed in the context of the master's theses of Benjamin Koster (2013), Florian Hartl (2013), and Shruthi Padma (2014), as well as the bachelor's theses of Gregor Semmler (2013), Matthias Wadlinger-Köhler (2015), and Felix Sonntag (2015). The graph-structural analysis of the datasets in section 3.6.2 expands upon the bachelor's theses of Sizhe Huang (2014) and Bernhard Schneider (2014). The analysis of temporal interaction patterns in section 3.5.2 expands upon the bachelor's thesis of Matti Lorenzen (2015), and the results have been previously published at the 5th International Workshop on Social Media World Sensors (Hauffa and Groh, 2019). All of the aforementioned theses were jointly supervised by Jan Hauffa and Georg Groh. Verbatim and near verbatim quotations from prior publications (Hauffa et al., 2012, 2014, 2016, 2019; Hauffa and Groh, 2019) are highlighted in gray.

The goals of this thesis are twofold: One one hand, we want to learn about the nature of social relationships from textual artifacts of communication. On the other hand, we want to show that the knowledge thus acquired is useful by applying it to the detection of social influence effects in online communication. This joint task requires longitudinal, observational data of computer-mediated communication. It has been theorized, and subsequently demonstrated via large-scale empirical studies (Eagle and Pentland, 2006; Stopczynski et al., 2014), that human everyday communicative behavior is multiplex in that it is spread across multiple communication channels, traditional and modern alike. It is therefore important not to limit our prospective experiments to a single communication medium, but to make an attempt to replicate our findings across a wide range of different media, to ensure that we learn about general properties of online communication rather than the effects of a particular medium's design.

Here, and in the remainder of this work, we use the term *social platform* to refer to all kinds of online services that enable social behavior. This includes traditional services for computer-mediated communication, social networking services, and social media in general. A social networking service (SNS) is defined as any social platform that lets users publicly and explicitly declare their social relationships, thus creating a social network. An SNS that has characteristics of traditional mass media, for example, a focus on the dissemination of user-generated and external content, is often referred to as a social medium. Finally, there are many systems for computer-mediated communication that do not offer any SNS-like functionality, but still make up an important part of people's online social behavior: e-mail, instant messengers, web-based discussion boards, chat systems, and many others.

According to Flanagin (2017), the rise of SNS and social media has created a new form of communication that blurs the lines between traditional interpersonal communication and mass media. For example, due to viral distribution on social networks, a message may reach an audience of a size that was formerly reserved to newspapers and television, but in ex-

change the original author has to relinquish all control over the presentation and interpretation of the message. An argument can be made that mass media, social media, and interpersonal communication on- and offline are not distinct concepts, but are situated in a continuous space of communication processes. We therefore feel justified in gathering communication data from a diverse range of platforms, traditional and modern alike, and subjecting them to the same experiments.

To be useful for our experiments, a set of online communication data has to meet a number of requirements:

1. Accessibility
    a) Communication needs to be *observable* under reasonable conditions. These conditions include the monetary expenses and the effort required to obtain the data and transform it into a useful representation, but also the ethical acceptability of accessing the data in the first place, and the ethical and legal implications of sharing it with other researchers for the sake of reproducibility.
    b) While the platforms from which data is obtained should be chosen with diversity in mind, the resulting datasets need to be *comparable*. Therefore, the chosen platforms should all occupy the same "communicative niche" as defined by the attributes "synchronicity" and "privacy" of Herring's classification of CMC systems (2007). We focus on platforms where the dominant mode of communication is public, asynchronous (non-realtime) textual messaging.
    c) Observations should to be available in sufficient *volume*, so that our ability to perform inference is not inhibited by a low sample size.
2. Data quality
    a) In the face of social platforms that report hundreds of millions of monthly active users (Kemp, 2020), it is unreasonable to expect being able to obtain a complete dataset of all activity on the platform, or even a representative — in the statistical sense — sample. Instead, we aim for a sample of practically obtainable dimensions, that adequately captures the behavior of at least a sub-group of the general user base. The choice of sampling method has been demonstrated to have an effect on certain downstream tasks, such as the study of information diffusion (De Choudhury et al., 2010). Social context is important for understanding social behavior, so random sampling of nodes or edges is clearly inappropriate. Instead, the sample should be a connected subgraph, so that it forms a core-periphery structure with respect to the entire graph.

        The problem of choosing a sampling strategy that produces a suitable core is a special case of *boundary specification* (Laumann et al., 1983). Ideally, the sample should correspond to a social group or community that is active on the platform. In practice, such groups are not cleanly delineated, and therefore difficult to identify, and the information required for doing so may not be available at the time of data acquisition. For example, application of the Girvan-Newman algorithm for community detection (Newman and Girvan, 2004) would require advance

knowledge of the complete social network graph. We therefore substitute the simpler requirement that the sample graph is at least as dense as the original. If the original graph has a community structure, a dense sample captures more inter- than intra-community edges. Our approach differs from what Leskovec and Faloutsos (2006) call "scale-down sampling": the objective of preserving local structure takes precedence over preserving the global structural properties of the original graph in the sample.

b) The observations should represent the full *variety* of the platform's user base. In particular, all forms of interaction supported by the platform should be observable, the content should not be restricted to a particular subject matter, and observation should not be limited to users from a specific demographic group.

c) An exception to the previous requirement is that communication should be limited to a *single language*, ideally one that is familiar to the researcher. This is necessitated by the use of NLP methods in the experiments and the need for manual inspection and interpretation of the results. Although topic modeling is fundamentally language-agnostic, its performance can be noticeably improved by language-specific preprocessing. This requirement has the side effect of promoting homogeneity, particularly in regard to the geographical distribution of actors.

d) Absence of systematic *errors* and a low rate of random error. Ideally, for each actor, the full content of all communicative acts, together with all relevant metadata, which includes the pseudonymous identity of sender and recipients and a timestamp, should be available in a clean, error-free, machine-readable form.

The first group of requirements mainly determines which social platforms are eligible for data collection, while the second group has implications for the data acquisition process. When acquiring data through an API or by means of web crawling, these requirements have to be considered in the design of the acquisition process. When working with an existing dataset ("found data"), one has to check whether all requirements are met and, if necessary, perform appropriate post-processing. We choose to acquire data from Facebook and Twitter as representatives of SNS and social media, and use two existing e-mail datasets as examples of traditional computer-mediated communication. In the following sections, we describe the social platforms, the data acquisition process, and the resulting datasets in detail. Social platforms keep introducing new features and changing the presentation of content. In this chapter, we describe the state of each platform at the time of data acquisition.

Since we intend to carry out experiments on the different datasets and compare the results, we need to know the differences and similarities of the datasets and the platforms the data was collected from. For that reason, this chapter goes beyond a mere technical description of the data acquisition process. We describe the history and the social conventions of each platform, carry out an explorative analysis of the temporal distribution of messages in each dataset, and compare the social network graphs. A discussion of the platform-specific ethical issues of the use of social media closes this chapter.

Figure 3.1: A typical tweet: Jack Dorsey, co-founder of Twitter, replies to the Twitter public relations team (Dorsey, 2018)

3.1 Twitter

Twitter (https://twitter.com) is both a minimalist social networking service and a platform for public messaging. Users can present themselves with a profile, subscribe to others' messages ("following"), broadcast size-limited messages ("tweets") to their followers and the general public, and forward ("retweet") received messages to their followers. Mentioning others by their screen name ("@-mention") transforms a tweet, at least in intent, from a broadcast to a directed message (Honeycutt and Herring, 2009), and enables direct interaction between users. One major difference between Twitter and other social media is the former's strict message size limit, which at the time of data collection restricted tweets to a maximum of 140 characters (Unicode code points).

There is only one relation over the set of users, "a follows b", whereby a subscribes to messages sent by b. In contrast to most other SNS, following is asymmetric, and does not require confirmation by the user being followed. The asymmetry is reflected in the notation originally used by Twitter: If a is a follower of b, then b is a friend of a. Twitter has since abandoned this notation (Stone, 2007) and uses the more symmetric pair "follower" and "following" on its user interface (UI). Each user is provided with a personal news feed that aggregates the messages sent by friends in chronological order. The news feed is the primary way of being exposed to tweets, though tweets can also be embedded in websites, found through keyword search, and selectively viewed by visiting the profile page of the sender. Figure 3.1 shows a typical tweet[1].

Twitter has a number of properties that make it desirable as a source of data: It is popular, with 340 million active users per month, according to company information (Kemp, 2020). While Twitter supports sending private messages, tweets are public by default, and can be

[1] All screenshots were taken at the time of writing and may show functionality and design elements that were not present at the time of data acquisition.

retrieved via an REST API, which also provides access to user profile data (Twitter, Inc., n.d.d). Finally, Twitter is recognized as a medium of communication by the general public, to the extent that the US Library of Congress has considered building an archive of tweets, arguing that they reflect "today's cultural norms, dialogue, trends and events" (Library of Congress, 2013).

3.1.1 Data Acquisition

Communication data was retrieved from Twitter via version 1.1 of its public API[2] simultaneously by several clients, physically different machines reporting to a central server. The iterative retrieval of data from an online source by visiting an entity, discovering connections to other entities, and repeating the process for a not yet visited entity is called *crawling*. To obtain a dense sample graph, we employ a snowball sampling strategy augmented by computationally inexpensive heuristics: Both the follower network and the implicit communication network are crawled in a manner resembling non-exhaustive breadth-first search (BFS). For each visited user, the names of followers and friends, as well as any other users that are @-mentioned in the retrieved tweets, are sent to the server. The server appends all users that have not been crawled yet to a queue, and assigns them in FIFO order to clients asking for work. Due to the distributed crawling, users are not visited in exact BFS order. For each visited user, all tweets sent between January 1, 2012, and the time of crawling are retrieved. The last day of crawling was June 25, 2012. Tweets are retrieved in blocks of 200, and requesting a retweet also returns the original tweet, so the resulting dataset contains older tweets as well as tweets not authored by any of the visited users.

BFS grows the sample graph by adding nodes in order of increasing distance from the starting node, and can be seen as a compromise between maximizing the density of the sampled subgraph, and thus the number of sampled nodes with a complete neighborhood, and obtaining an unbiased sample. It is known to produce sample graphs that do not share all structural properties of the original, in particular due to a bias towards high-degree nodes (Kurant et al., 2010), which correspond to Twitter users with a high number of friends and / or followers. Alternatives that do not suffer from this bias, such as Metropolis-Hastings random walks (Gjoka et al., 2011), produce much more sparse sample graphs and are therefore not suitable for our purposes.

A well-connected (at least 100 tweets, 1 000 friends, 1 000 followers), English speaking user was chosen arbitrarily as a starting point for crawling. Restrictions are imposed on the crawling process to avoid atypical users: Spam accounts are expected to be reported and deleted shortly after becoming active, so only users that have been registered for more than 10 days and have posted more than 10 tweets are crawled. We note that these heuristics are very basic, compared to the features used by state-of-the art machine learning systems for spam detection (e.g., Varol et al., 2017), so we must expect a certain amount of spam to pass this filtering step. Any user who has fewer than 25 000 followers and friends is considered for future crawling. This arbitrary limit is intended to introduce a bias towards local explo-

[2] Here, and in the following sections, we aim to give a high-level overview of the methods of data acquisition, cleaning and pre-processing. For a more detailed account, the reader is referred to the source code (see section 1.1).

ration and mitigate the effect of "hub" users, e.g., celebrities, who connect otherwise distant parts of the network via unidirectional follower edges. If mutual awareness is unlikely, such edges may not even exhibit the weak tie effects described by Granovetter (1973). To simplify subsequent natural language processing steps (see section 4.2.1), users are excluded from crawling if the language of their tweets cannot be confidently identified as English.

For each crawled user, we store the unique ID and the following values from the JSON `User` object returned by the API (name of the corresponding key in parentheses): account creation date (`created_at`), screen name (`screen_name`), number of followers (`followers_count`), number of friends (`friends_count`), number of tweets sent (`statuses_count`). In addition, we retrieve the IDs of all followers and friends. For each retrieved tweet, we store its unique ID and the following values from the `Tweet` object (name of key in parentheses): date of submission (`created_at`), text (`text`), ID of sender (`user.id`), ID and sender of the original tweet, if it is a retweet (`retweet_of_status_id`, `retweet_of_user_id`), or ID and sender of the parent tweet, if it is a reply (`in_reply_to_status_id`, `in_reply_to_user_id`).

Crawling produced a longitudinal dataset of 358 342 users and 223 million tweets. It contains activity between 2006 and July 2012. Since some of the following processing steps are computationally expensive, the dataset is subsampled to the first 30 000 users in the order of crawling. A development set of 2 500 users for parameter tuning is extracted in the same way.

Content of Referenced Websites

Tweets are limited in size, but frequently reference external resources: about 40% of the collected tweets contain a URL. We hypothesize that the analysis of tweets via topic modeling can benefit from the auxiliary information provided by the referenced websites. To be able to test this hypothesis, we collect the textual content of websites referenced by URLs found in the 30 000 user subset on a best-effort basis. A concrete use case for this data is discussed in section 4.2.2.

We begin by extracting URLs from the tweets. A regular expression is applied to the text of each tweet to identify substrings that are syntactically valid URLs. A URL that occurs at the end of a tweet may be truncated in a retweet, so the original tweet is preferred if available. This process is not exhaustive. Neither does it identify all substrings that are valid URLs according to RFC 3986 (Internet Engineering Task Force, 2005), which would incur a large number of false positives, nor does it fully validate the identified URLs. In particular, we require a URL to start with a scheme identifier associated with the HTTP(S) protocol and do not check the validity of the top-level domain name. Since 2010, Twitter heuristically detects URLs in tweets, even if the scheme is omitted, and replaces them with custom URLs generated by its link shortening and monitoring service `t.co`, which redirects requests to the original URLs (@SG, 2010). The replacement URL always includes the scheme, so we expect our conservative approach to URL detection to be sufficient for the majority of tweets in the dataset.

The collection of website content from a given URL can be divided into two processing steps: First, the document that is referenced by the URL is retrieved. In a second step, the

text body is separated from any HTML markup and miscellaneous text, e.g., navigational elements. To retrieve the document, an HTTP `GET` request (Internet Engineering Task Force, 2014) is issued for the URL. The request header includes an `Accept-Language` field that is set to indicate a preference for English language content. If the response is a redirection to a different location (HTTP status codes 300-399), another `GET` request is issued for the new location. Redirections are frequent: Even before Twitter began to automatically perform URL shortening, external URL shortening services were in widespread use. If the status code of the response indicates success (codes 200-299) and the MIME type specified in the response header starts with "text", the response body is subjected to content extraction. Textual content is extracted using the method of Kohlschütter et al. (2010), implemented by its authors in a Java library called "boilerpipe". Among the different extraction strategies provided by the library, we choose `ArticleExtractor`, which is described as "tuned towards news articles" (Kohlschütter, 2010), a type of content we expect to be frequently shared on Twitter. The extracted content is discarded if its language cannot be identified as English. The final result of this process consists of the original URL as it appears in the tweet, the URL of the final location after following all redirections, and the extracted text.

There are three known problems with this process. Due to the long interval of time between crawling Twitter in 2012 and retrieving the websites in 2015, some content was no longer available or subject to access restrictions like geofencing or "paywalls". For example, some news articles have been moved to an archive that is accessible to subscribers only. To address this problem, no text extraction is performed if the final URL after redirection is a known error page. Furthermore, if the same content is retrieved from many different URLs, it is likely to be a generic error message, login request, etc. If a piece of content is found at more than 10 different URLs, all of these URLs are discarded. Another issue is that EU legislation requires website operators to inform visitors about their usage of HTTP cookies. The chosen text extraction library predates the commencement of this legislation and does not reliably filter out such notices, so the wording used by two frequently referenced websites, Twitter and Instagram, was manually detected and removed. Finally, the process described here cannot, by design, handle dynamic websites that load all of their content via asynchronous HTTP requests.

Data Quality

A general issue with Twitter as a source of communication data is the low information content of a single tweet. At the time of data acquisition, tweets were limited to a length of 140 characters. A tweet may contain elements, such as @-mentions, whose primary purpose is to enact social conventions, rather than transmitting information. After removing these elements and performing further linguistic processing, e.g., removal of stop words and emoticons, it is possible that no text is left.

When reconstructing a social network from the graph of explicitly declared "following" relationships, one has to consider that following a user implies exposure to that user's tweets, but Twitter does not provide any information on which tweets are actually viewed, if any. This is particularly relevant in the case of users who follow many others. Since following does not necessarily entail awareness of the followed user's tweets, it is a weaker indicator

for the existence of a social relationship than explicit ties on other platforms. Furthermore, the API only reports the current presence of a following relationship, so it is not possible to determine how long that relationship has existed. Conversely, relationships that ended before the time of crawling are unobservable. As a side effect, the observed lists of followers and friends of two users are inconsistent if a relationship between them is established or dissolved after crawling one user, but before crawling the other.

In the course of exploratory data analysis, a number of issues with data integrity and consistency were found: First, a number of tweets bear implausible timestamps, some predating the launch of Twitter by decades. Since 2010, the timestamp is stored in a redundant way: as a separate field in the JSON data structure returned by the API and encoded into the tweet's unique ID (@rk, 2010). The tweets predating this storage scheme exhibit no visible inconsistencies, and for all later tweets, an apparently correct timestamp could be recovered from the ID. Second, one particular user is reported by the API as a follower of himself, creating a loop edge in the follower graph. This edge was manually removed from the dataset. Third, there are tweets which, according to the API, are replies to a particular user's tweet, but the reported ID of that "parent" tweet is invalid. If an experiment requires the identity of the parent tweet, these replies are ignored. All of these issues appear to be the result of insufficient server-side validation of submitted tweets.

3.1.2 Social Conventions

The originally envisioned use case for Twitter was posting brief "status updates", also known as *microblogging* (Java et al., 2007). When holding conversations on Twitter became more popular (Honeycutt and Herring, 2009), the community reached consensus on conventions for addressing other users and crediting them with authorship, and for the categorization of messages. These social conventions were subsequently adopted by Twitter and integrated into the UI. A chronology of these developments has been compiled by Halavais (2014). The following conventions are directly relevant to our experiments:

@-mentions Prefixing a user name with the '@' sign anywhere in a tweet causes the specified user to be notified of the tweet. Honeycutt and Herring (2009) identify two main uses of @-mentions: Addressing a message to another user, and referencing a user in a message intended for a wider audience.

Replies By convention, tweets that start with an @-mention are addressed to the mentioned user and should be considered part of an ongoing conversation. Any tweet that begins with an @-mention only appears in the news feed of users who follow both sender and recipient. A conversation can be started manually or by replying to a previous tweet via Twitter's UI.

Retweets Reposting a received tweet under one's own name extends its visibility by exposing it to a new set of followers. Customarily, the tweet is attributed to its original author by prefixing it with a retweet indicator such as "RT" or "via", followed by @-mentioning the author.

Hashtags Prefixing a single word with the '#' sign designates it as a tag or categorical label, which places the tweet into the context of other tweets with the same tag. A tweet may contain multiple hashtags.

Replies and retweets can be constructed manually by following the conventions outlined above, or created with the help of the UI. If a reply or retweet is made via the UI, its parent or original tweet is recorded as meta-data. Twitter also reliably detects and records meta-data for manual replies that begin with an @-mention of an existing user. No meta-data is recorded for non-UI retweets, so these have to be detected heuristically by the API consumer, which fails if non-standard retweet indicators are used. While the UI makes replying and retweeting more convenient, some users value the flexibility that is offered by manual construction of reaction tweets. This has led to the creation of hybrid forms such as the "modified retweet", where the original is quoted selectively and the retweeter may add commentary of his or her own, and the "public reply", where the @-mention is prefixed with a period, so that the tweet is visible to all followers.

Kooti et al. (2012) analyze the process that led to adoption of the "RT @" prefix for retweets and find that it can be explained by social influence originating from a small number of well-connected core users. Despite this evidence for high-level social processes, it is difficult to quantify to what degree Twitter is a network of peers, as opposed to a bipartite network of content producers and consumers (Huberman et al., 2009). Kwak et al. (2010) find that the structure of the follower network differs from known social networks, and attribute this to the low amount of user pairs in a reciprocal following relation. Wu et al. (2011) describe a class of "elite" users, mostly comprised of celebrities and Twitter accounts associated with traditional news media, who get retweeted a lot, but rarely interact directly with regular users. This is consistent with our earlier observation that the edges of the follower graph only convey a weak notion of social relatedness. Even reciprocal following is not necessarily indicative of anything more than superficial awareness: Gabielkov and Legout (2012) describe a class of users who "follow back" everyone that follows them.

Among the 17.3 million tweets of the 30 000 user dataset, 40% contain at least one @-mention, 28% are replies, 18% are retweets, and 46% are devoid of any conversational features. 85% of retweets (15% of all tweets) are made via the UI, while the remaining 15% are identified as retweets by the presence of one of the retweet indicators listed above. 8% of replies are users replying to their own posts, presumably to present a group of related tweets in proper sequence. The most striking difference to other social platforms is that messages without a clearly defined addressee or target group make up almost half of the overall communication volume.

3.2 Facebook

With more than 2.4 billion monthly active users in 2019, according to company information (Kemp, 2020), Facebook (https://www.facebook.com) is currently the most popular social networking service. Facebook focuses on interaction with a circle of friends. It defines "friendship" as a symmetric relation, and accordingly the establishment of a friendship tie requires the consent of both users. Compared to the unidirectional following on Twitter,

Figure 3.2: A typical Facebook post: Mark Zuckerberg, co-founder of Facebook, talks about plans for his AI assistant, and actor Robert Downey Jr. comments (Zuckerberg, 2016)

friendship on Facebook is more likely to be accompanied by an offline social relationship. While these offline relationships may differ in emotional and geographical closeness, Facebook does not explicitly distinguish close friends from brief acquaintances. However, family members and spouses can be, after mutual confirmation, labeled as such. Facebook is frequently used as the primary tool for maintenance of relationships between geographically distant individuals (Bryant and Marmo, 2009).

On their profile page, users can opt to upload a picture and provide information about themselves in a structured way. Among the predefined fields are date of birth, gender, sexual orientation, "relationship status", language proficiency, education and work history, religious confession, political leaning, and various contact details. Each user has an individual timeline, a page where text or various kinds of media (pictures, video, hyperlinks, etc.) can be posted, and the content of other users can be republished ("sharing"). A news feed that shows recent activity of friends in reverse chronological order. Users can interact with others by posting on their timeline, commenting on their posts, or by exchanging private messages. A typical Facebook post is shown in figure 3.2. Facebook has repeatedly increased its upper limit on the length of posts, having a limit of 63 206 characters at the time of data collection (Lavrusik, 2011). It is unknown whether this limit is in units of bytes (i.e., single-byte encoded characters) or Unicode code points. The maximum length of a comment appears to be 8 000 bytes (Web Applications Stack Exchange, 2012).

Users can be mentioned in a post by explicitly tagging them via the UI, which causes Facebook to insert a header line of the form "*a* is with *b*" or "*a* is with *b* at *location*". If a user is mentioned by name in the post text, the name is converted to a hyperlink to that user's profile. Since 2013, Facebook supports hashtags in a way that is similar to Twitter: a word prefixed with the '#' sign is converted to a hyperlink that, when clicked, initiates a search for other recent posts that contain the tag. Compared to the other datasets described in this chapter, the Facebook dataset is the only one that is both sufficiently recent and informal in register to contain a non-negligible amount of emoji. Facebook also supports a small

set of proprietary graphical emoticons that predate the Unicode standardization of emoji. The handling of these graphical characters in the context of topic modeling is described in section 4.2.1.

A major difference between Facebook and online services that, like Twitter, are oriented towards public discourse, lies in its privacy settings, which give users fine-grained control over the visibility of their data. Users can control the visibility of their profile, their list of friends, and individual posts on their timeline, including messages left by others. Visibility can be unrestricted, limited to friends, limited to a specific group of users, or fully restricted, in which case the content is invisible to everyone but the owner. In effect, what is visible to an outside observer may only be a subset of a user's profile data and interactions. What is visible depends on the user's desire for privacy, and therefore varies from person to person.

3.2.1 Data Acquisition

The dataset was acquired by BFS-style crawling of the friendship graph. For each crawled user, all publicly visible friends, profile details, and posts on the timeline, along with their comments, are retrieved. The dataset therefore only contains content that is accessible to all Facebook users. The crawler has a client-server architecture similar to the Twitter crawler. Unlike Twitter, however, Facebook does not provide unconditional API access to the profiles and timelines of arbitrary users. In order to obtain the data we are interested in via the official API, users would have to manually install a crawling "app" and explicitly grant it permission to access the relevant parts of the profile. Our experience with the user study on keyword extraction from private messages (section 2.2.2) shows that this approach does not scale: in the absence of viral spreading, which cannot be reliably provoked, the acquisition of N participants requires proportional effort, and is unlikely to produce a dense network.

Instead, we adopt a screen-scraping approach. The crawler clients use the Selenium WebDriver framework (The Selenium Browser Automation Project, 2020) to automate a web browser (here: Mozilla Firefox) and interact with Facebook. Being a web-application, Facebook delegates certain computational tasks to the viewer's PC to improve UI response time, and loads data on demand. On request of the crawler server, a client logs into a Facebook account, directly navigates to those parts of a user's profile and timeline that can be reached through a URL, and simulates interaction with navigational elements. This triggers execution of the associated JavaScript code, which then loads additional data via asynchronous HTTP requests. The three main forms of navigation are explicit pagination, where the currently displayed data is replaced by new data after clicking a "next" button, "infinite scrolling", where new data is appended to the end of the page when scrolling down to a certain position, and collapsed sections that can be expanded by mouse click, inserting new data into the current page. After all interactions are completed, a HTML representation of the current DOM tree (WHATWG, 2020) is sent to the crawler server. The server archives the HTML data, transforms it back into a DOM tree, performs XPath queries to extract the relevant content, and stores the resulting structured data in a relational database. The server performs language detection on the textual content of posts and comments to ensure that only the friends of English-speaking users are considered for future crawling. An arbitrarily chosen English-speaking US resident with at least one publicly accessible post and a public

friend list was chosen as the starting point for crawling.

For each user, we store the unique ID, basic demographic data (gender, birth year, place of residence), the number of friends as reported by Facebook, and as many IDs of friends as can be retrieved by paging through the list of friends. For each element on a timeline, either a post or a comment, we store its ID, timestamp, the ID of the sender, and any textual content. If the ID of the element is not exposed through the DOM, a synthetic ID is generated. For comments, we additionally store the ID of the parent post. For posts, we store the IDs of mentioned users and whether or not they contain shared content. If a post contains shared content, and the ID of the original post is exposed, we store it as well.

This crawling strategy has several issues: First, it is slow. Each time the client triggers the load of additional data, the DOM tree is modified, and has to be rendered by the browser. Facebook displays the timeline of a user on a single page, which contains many navigation elements that have to be triggered, which causes the page to be rendered many times over. The overhead adds up and limits the number of users that can be crawled under given time and resource constraints. Second, Facebook frequently makes small changes to the page layout, some of which may necessitate an adaptation of the XPath queries for data extraction. If these layout changes affect the location of a mandatory element in the DOM tree, e.g., the timestamp of a post, extraction will fail immediately. However, many elements are allowed be absent, e.g., not every post has comments. It is not possible to distinguish the absence of an element from an outdated XPath query. Therefore, this problem has to be addressed by manual monitoring. If certain pieces of information start to be consistently missing, the structure of the DOM tree is investigated. If necessary, the XPath query is adapted, and the missing data points are recovered from the archived HTML. This process is labor-intensive and prone to human error.

In addition, some issues were found with the data that was successfully scraped from Facebook. Facebook allows users to arbitrarily backdate posts to commemorate important past life events, so timestamps do not always correspond to the actual date of posting. Some observed inconsistencies cannot be explained by backdating, for example, comments predating the post they are associated with, so there appear to be rare cases of unsystematic corruption of timestamps. Not all inconsistencies can be clearly attributed to either Facebook or the crawler: There is a discrepancy between a user's friendship count as reported by Facebook and the number of friends that are actually retrieved. While it is possible that the number of friendships changes between the retrieval of the friendship count and the retrieval of the actual friend list, this is unlikely to be a sufficient explanation. On average, we retrieve about 8% fewer friendships than reported, there is no case of retrieving more friendships, and the number of missing friendships is roughly proportional to the number of reported friendships.

Another class of inconsistencies arises from the fact that entities like posts, comments, pictures, etc., presumably have a unique identifier in Facebook's internal data model, but this ID is not consistently exposed through the DOM of the website. A direct consequence is that there is often no way to unambiguously associate a piece of shared content with the post it originally appeared in. Also, a single object in Facebook's internal data model may show up on the timeline more than once. A photo album appears on the timeline each time new photos are added to it, and each appearance is accompanied by all comments that the album

received over its whole lifetime. If the ID of the album is not exposed, these comments have to be heuristically deduplicated. Since we do not know in response to which appearance of the album a particular comment was written, the best possible approximation is to associate each comment with the closest previous appearance of the album. If no suitable parent post can be found, usually due to incomplete crawling, no parent is recorded. Even though attempts were made to identify and correct systematic errors, we expect the Facebook data to be more noisy than data obtained through a documented API.

The set of crawled posts and comments may contain duplicates for several reasons: Due to crawler error, multiple copies of the same post may have been retrieved. A user may have accidentally submitted a post or comment twice. Finally, a post that references other users may, at the discretion of Facebook, appear on the timelines of some or all of these users. We address this issue with a two-step process. First, the set of crawled posts and comments is exhaustively scanned for duplicates. Two posts are considered identical if they have the same content, timestamp, and sender. Then, from each set of duplicates, one canonical post is generated by merging the lists of recipients.

Finally, there is one social convention to be aware of: A commenter may call the attention of other users to a post by mentioning their names in the comment, which causes them to be notified. Frequently, such comments contain no textual content beyond a list of names. We therefore remove the names of mentioned users from the text body as a general processing step that is applied to all posts and comments. If no text remains, the post or comment is removed by the deduplication process.

The resulting dataset contains profile data from 16 834 users and activity on their timelines between 2005 and March 2015. Among the 79% of crawled users who specify their gender and make it visible to the public, 45% are female and 55% male. Only 3% of users provide their date of birth. Their age distribution is shown in table 3.1. 57% of users specify their place of residence, but Facebook frequently only displays the name of the city, which may not be sufficient to unambiguously identify the geographic location. For only 31% of users, the place of residence that is displayed on the profile page includes the name of a federal state, province, or country. 99% of these users resided within the USA at the time of crawling. Their locations, grouped by state, are shown in figure 3.3. Due to the low sample size, a breakdown of the remaining 1% by country is omitted. The geo-spatial distribution exhibits a noticeable bias towards the state of Massachusetts (MA) and the surrounding states. Other states are represented roughly proportional to their population. This geo-spatial distribution suggests that BFS-style crawling does not only promote the density of the resulting sample graph, but also the geographical closeness of its nodes. Limiting the crawling to English speaking users further contributes to this effect. Even though the geographical location of the seed user, i.e., the starting point of the crawling process, is not known, one may reasonably assume that he or she is located in or close to MA.

For their 2012 study on information diffusion, Bakshy et al. (2012) were able to acquire a large amount of user profiles directly from Facebook, and report basic demographic data. With a sample size on the order of hundreds of millions of users, their dataset can claim to be reasonably representative of the general population of Facebook users. By comparing our sample to that of Bakshy et al. (specifically, the “no feed” subset), we can get an impression of the bias incurred by our crawling procedure. With 52% of users specifying their gender

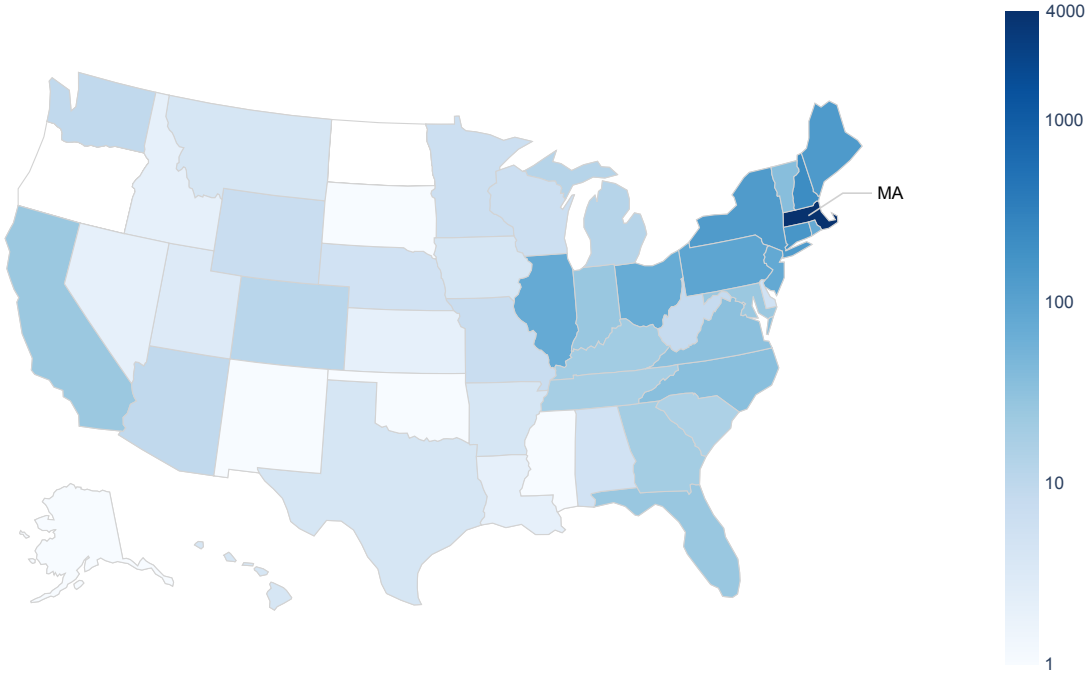


Figure 3.3: Geo-spatial distribution of crawled Facebook users with residence in the USA, grouped by federal state. Color intensity is logarithmically proportional to the absolute number of users crawled.

Table 3.1: Age distribution of the crawled Facebook users

	our data	Bakshy et al. (2012)
n	560	$> 218 \times 10^6$
< 18	0.0%	13.1%
18–25	53.6%	36.1%
26–35	40.9%	26.9%
36–45	2.7%	12.9%
≥ 46	2.9%	10.9%

as female, the gender distribution of Bakshy et al. is closer to uniform than ours. The age distribution, shown in table 3.1, differs from ours in two ways: The complete absence of users below the age of 18 indicates that the accounts of children are specially protected and therefore inaccessible to our crawler. Users above the age of 35 are underrepresented in our dataset. Since the sample of Bakshy et al. is neither limited to users within a certain path distance from a seed node nor restricted to English speakers, its geo-spatial distribution is much less concentrated, with only 29% of users residing in the USA.

Among the 5 149 768 posts and comments that were retrieved, 131 847 (2.6%) duplicates were identified and removed. Some posts contain media such as photos, photo albums, videos, or URLs, which are not accompanied by any textual content, and are therefore not relevant for our experiments. The final dataset contains 1 032 826 posts with textual content and 2 693 783 non-empty comments.

3.3 E-Mail

Electronic mail (e-mail) is a decentralized system for the delivery of textual messages. In contrast to the previously discussed social platforms, it is not operated by a single service provider. The delivery of e-mail is handled by a network of independently operating mail transfer agents (MTA). E-mail is an open system, in the sense that specifications are freely available and various implementations of the client and server components exist. A simplified model of the e-mail system can be described as follows: Via a piece of client software called the mail user agent (MUA), users connect to the specific MTA that is responsible for their e-mail account. After successful authentication, the MUA collects incoming mail and submits outgoing mail to the MTA for delivery. If a message cannot be delivered locally, i.e., is not addressed to an account that is directly handled by that MTA, it is passed on to a different MTA that is closer to the destination.

The scheme for the addressing and routing of messages is closely tied to the domain name system (DNS). An e-mail address usually consists of a local and a global part, separated by the '@' character. The global part is a domain name, and the MTA that is responsible for mail addressed to that domain can be identified by querying the DNS. Upon receipt of a message, the MTA will then look at the local part, an arbitrary string that identifies the account, and decide if the message can be delivered locally, or has to be forwarded to another MTA. A direct consequence of this design is that the identity of all recipients has to be known to the sender. Communication among dynamic groups is usually implemented by means of "mailing lists": a central server maintains a list of the addresses of group members, regularly polls a specific mailbox, and forwards incoming mail to all members.

The structure of an e-mail message is defined by RFC 822 (Internet Engineering Task Force, 1982). While this standard has been repeatedly superseded by later RFCs, the basic structure of a message remains unchanged: the header, an order-invariant sequence of key-value pairs ("fields"), is followed by the message body. The body can either be a single block of plain text, or a sequence of MIME parts, as specified by RFCs 2045 to 2047 (Internet Engineering Task Force, 1996b,c,a). MIME parts have their own headers and can be nested to an arbitrary depth. Notable use cases of MIME are attaching files to messages, and providing

```
Date: Fri, 11 Jan 2002 20:34:11 +0100 (CET)
Message-ID: <8B8F71DE55D07C4B94CEA9EC69A2572824D3E9@NAHOU-MSMBX03V.corp.enron.com>
MIME-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
Microsoft Mail Internet Headers Version 2.0
X-MimeOLE:  Produced By Microsoft Exchange V6.0.5762.3
content-class:  urn:content-classes:message
Subject:  Kudos!!
X-MS-TNEF-Correlator:
 <8B8F71DE55D07C4B94CEA9EC69A2572824D3E9@NAHOU-MSMBX03V.corp.enron.com>
Thread-Topic:  Kudos!!
Thread-Index:  AcGa1vII1vfRssMVSDeMZQKC+fYzFQ==
From: "Fogo  Georgia" <Georgia.Fogo@ENRON.com>
To: "Lay  Kenneth" <Kenneth.Lay@ENRON.com>

Congratulations on the successful auctioning of the trading operations.  There
now seems to be a positive buzz among the employees here in Houston -- things
are looking up.  Let's keep the momentum going!

P.S.  Hang in there with all the other stuff going on -- you have my support.
```

Figure 3.4: A typical e-mail message, addressed to Kenneth Lay, CEO of Enron. Taken from the "EDRM v2" version of the Enron corpus. Meta-data that was inserted for archival purposes is omitted, and lines longer than 80 characters are wrapped.

variants of the message body in different formats, usually HTML and plain text, so that a MUA that does not support one format can fall back to the other. The header contains the message ID, a unique identifier that is either generated by the MUA or inserted by the first MTA that handles the message. Other relevant header fields identify sender and recipients and document the date of sending. It is customary for the sender to provide a brief summary of the message content in the "subject line", which is also transmitted as a header field. The message header, as well as the header of each MIME part, has a `Content-Type` field that specifies the format of the data in the body. An example of an e-mail message including the header can be found in figure 3.4. There is no theoretical limit to the size of an e-mail, but an MTA may reject overly large messages. In practice, the maximum acceptable size is in the tens of megabytes (Zapisotskyi, 2019).

Due to the open nature of the e-mail system, a diverse range of client software (MUA) exists, but these applications have over time converged to a common set of core functionality and common principles of UI design. Most clients divide the interface into three parts: a list of mailboxes and folders, a (reverse) chronological list of messages in the current folder, and the content of the currently selected message. Users can create a hierarchy of folders for the topical organization of their mail, with some folders having a special function. Usually the client saves copies of outgoing messages in an "outbox", stores messages that have not yet been sent in a folder for drafts, and moves deleted mail to a "trash" folder before it is actually deleted. A new message can be sent on its own, as a reply to a previously received message, or by forwarding a received message to someone else. There are two common

approaches to the storage of mail. Some clients store the mail that was received from the MTA locally, others, especially clients that are designed for use in a corporate environment, keep all messages stored on the MTA, and only maintain a local cache. This facilitates the coordinated backup and archival of all users' mail.

3.3.1 Data Provenance

We work with two collections of e-mail: The Enron dataset consists of about 150 mailboxes of employees of Enron Corporation, which were made public in the course of legal proceedings (Diesner et al., 2005). The HackingTeam (HT) dataset is derived from data that was "leaked", i.e., made available to the public despite its confidential nature, from an Italian IT company specializing in surveillance software in 2015 (Greenberg, 2015). The leak includes data that was stored on the company's e-mail server, including the mailboxes of 60 employees. In contrast to the social media datasets discussed previously, the two collections of e-mail are found data. Each collection is a snapshot of the state of e-mail communication within a company at a specific point in time. Since we had no control over the process of data acquisition, it is all the more important to be aware of the historical context of the data and all processing steps that have already been performed by previous custodians.

A number of other collections of e-mail were considered for use in the upcoming experiments, but ultimately rejected:

- The *IIT CDIP Test Collection* consists of documents that were released in the course of legal proceedings against US tobacco companies, including more than 300 000 e-mail messages (Eichmann and Chin, 2007). These messages were reconstituted from printouts by means of OCR and manual data entry. The majority of the meta-data that is usually found in a message header is not preserved in print, and is therefore missing from the dataset. The quality of the recovered text is negatively affected by OCR errors. The work of Padma (2014) and Wadlinger-Köhler (2015) illustrates the difficulty of unambiguously identifying sender and recipients in this setting.

- The *Avocado collection* (Linguistic Data Consortium, 2015) contains anonymized messages from 279 mailboxes of employees of an unnamed, now-defunct IT company. The messages cannot be freely redistributed, and can only be used for research purposes under restrictive legal constraints that guard the privacy of the former employees. Its proprietor, the Linguistic Data Consortium, levies a distribution fee that constitutes a significant barrier to smaller research groups.

- Over the past decade, *WikiLeaks* has published leaked e-mail from various political figures, government bodies, and companies (Wikipedia, 2020b). The list of individual US politicians includes Sarah Palin, former governor of Alaska, John Podesta, chair of Hillary Clinton's 2016 presidential campaign, Hillary Clinton herself, and staff members of the Democratic National Committee. Their datasets are comparatively small and yield egocentric social networks that are not amenable to conventional social network analysis. The "Syria Files" are a collection of approximately 2.4 million messages from Syrian politicians and ministries. Any analysis beyond the purely statistical is

made difficult by the fact that the predominant languages are Arabic and Russian. The "Global Intelligence Files", 5.5 million e-mail messages of intelligence company Stratfor, have not yet been used in social computing research. It is not known whether the dataset is a reasonably complete sample of communication among Stratfor's employees.

- *GMANE* is a service that archives the messages exchanged on public mailing lists, mostly those of open source software development projects. It provides a convenient NNTP interface for bulk retrieval of messages (Marek-Spartz et al., 2012). The context of the messages leads us to expect a set of actors that is divided into core developers, who interact frequently and over a long period of time, and occasional contributors, who are only active for brief periods of time. In terms of content, we expect a narrow topical focus and interactions that are mainly driven by technical, rather than social, concerns. These unique features would make a comparison to other collections of e-mail difficult.

Enron

The history of Enron, as related by Diesner et al. (2005), can be summarized as follows: In 1985, Enron was formed through a merger of a Houston utility company and a gas pipeline company, and has been based in Houston, Texas, ever since. Initially, Enron bought wholesale electrical power and sold to industrial and retail customers. Following the deregulation of the US energy market, Enron positioned itself as an energy broker, and later expanded into new commodity markets such as TV ad time and Internet bandwidth. In October 2001, after being presented with internal accusations of improper accounting and business practices, CEO Kenneth Lay had to publicly announce huge financial losses that had occurred over the previous five years. The stock price dropped, Enron became insolvent, and had to file for bankruptcy in December 2001.

Following the bankruptcy, the US Federal Energy Regulatory Commission (FERC) commenced investigations into Enron's business practices. In April 2002, FERC made a contract with Aspen Systems (now part of Lockheed Martin) to collect e-mails and databases from Enron's computer systems (Bartling, 2006). Bartling, "hired by Aspen Systems as an independent contractor to lead and manage the collection" of data, specifically recalls copying "email PST files", where PST refers to the storage format used by Microsoft Exchange for data exchange and client-side storage. The collected messages contain evidence that Enron originally used Lotus Domino and Notes for their internal e-mail infrastructure, but switched to Exchange and Outlook in May 2001[3]. Both are CSCW systems of comparable functionality, providing e-mail, contact management, calendars, and other tools for collaboration. The first of each pair is the server software that acts as MTA, the second is the client software that acts as MUA.

In the course of its investigations, FERC decided to put the collected data in the public domain. In May 2002, a collection of e-mails from 156 employees (as counted by Zhou et al.,

[3] See file `native_000/3.161429.NMHYAIKYOAIW2SRBIPGOS3MQUZOSJRXIB.eml`, archived in `edrm-enron-v2_skilling-j_xml.zip` in the "EDRM v2" version of the Enron corpus.

2007), mostly senior management, was publicly released (Diesner et al., 2005). The e-mails could be individually accessed via a web interface and the whole set was available for purchase on physical media (Bartling, 2006). After receiving complaints that the publication of their e-mail communication violated the privacy of Enron's employees, FERC removed 141 379 e-mails pending further review (Krasnow Waterman, 2006). E-mails that were found not to be entitled to removal were successively reinstated. Different sources estimate the total amount of published e-mails between 1.3 and 1.6 million (Krasnow Waterman, 2006; Cormack et al., 2010).

Multiple parties are known to have purchased the Enron e-mail data from Aspen Systems in aggregated, machine-readable form, but apparently none of them have published the data in its original form, without any additional processing. A letter from Aspen Systems Corporation (2006) to an employee of the University of Maryland, who handled the purchase of the data, is our only source of information on the nature and format of the data released by FERC[4]. The letter mentions four datasets, two of which contain e-mail: the "Enron Email database" and the "Enron Email (.pst) database". For clarity, we will refer to the former as the "non-PST" and to the latter as the "PST" database. Each is subdivided into a main part and a "re-released" part, which most likely contains the messages that were reinstated in the aforementioned review process. A database consists of data in a CSV-like, tabular format, and a set of external files that are referenced by name from the data table. Each row of the table corresponds to one message. Both databases contain the usual e-mail meta-data like subject line, sender, and recipients. The fact that the non-PST database has a column named `OCR_TEXT`, and e-mail attachments are only provided as TIFF images, suggests that the e-mails in this database were reconstituted from paper printouts by means of optical character recognition (OCR) and (possibly manual) meta-data extraction. The PST database has additional columns for the message headers and body, and a column `FILENAME`, which is preserved in some published collections of the Enron data. As we will see, this column contains evidence that Aspen Systems collected data not only from the Exchange server, but also from the decommissioned Notes-based system. In summary, Aspen Systems received e-mail data in different formats (possibly including printouts), extracted individual e-mails and their meta-data, and stored them in their own database format.

Two parties ended up publishing processed versions of the data they had received from Aspen Systems, thus creating two branches of what is now generally known as the "Enron (e-mail) corpus". Each of these branches has in turn spawned a number of variants that differ in the format of the data and the additional processing steps that were taken. The family tree in figure 3.5 shows the relationships between the major variants.

The first branch originates with the CALO project at SRI International. Researchers at the Massachusetts Institute of Technology purchased the data from Aspen Systems and passed it to the CALO group, who processed the data and "corrected several integrity problems" (Diesner et al., 2005). The dataset is currently hosted and maintained at Carnegie Mellon University (CMU), but the processing steps that were performed on the data obtained from Aspen Systems are not documented beyond some brief remarks on the CMU website (Cohen, 2015). CMU provides its version of the data in the form of EML files in a directory structure

[4] Credit for discovering the letter goes to the EnronData Project (2016).

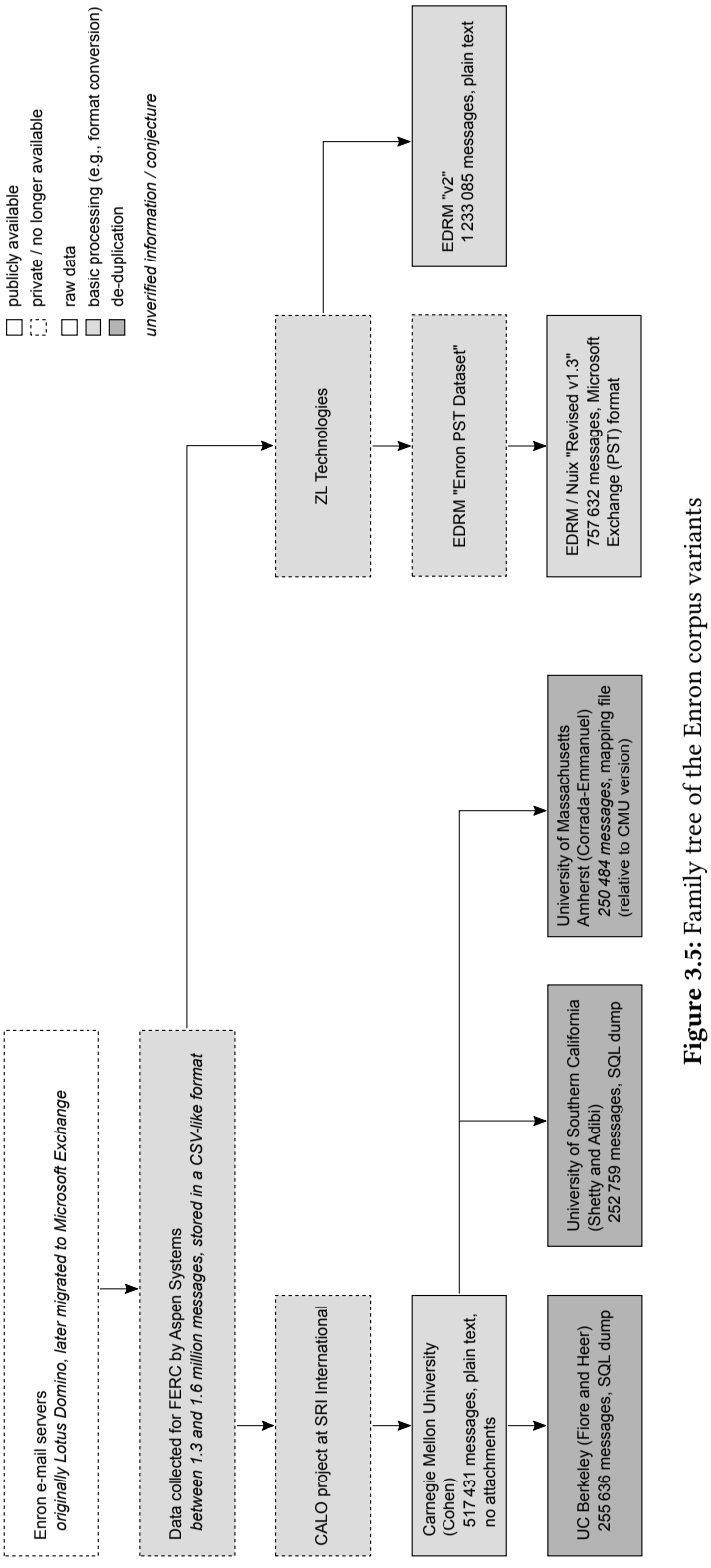


Figure 3.5: Family tree of the Enron corpus variants

that mirrors the original mailboxes and folders. EML is a plain text format that stores the e-mail data as it is received from an MTA: a set of headers and the message body, formatted according to RFC 822 (Internet Engineering Task Force, 1982). Additional meta-data, likely originating from the Aspen Systems database, has been inserted into the header in the form of non-standard fields, identified by a name starting with "X-". A field named `X-FileName` presumably contains the name of the file in which the message was originally stored. Different messages reference files with the extensions "NSF" and "PST", which are associated with the storage formats of Lotus Notes and Microsoft Exchange, respectively, which suggests that the data originates from two different software systems. The CMU version of the Enron corpus contains 517 431 messages[5] without their original attachments.

The second branch of the corpus is the result of a collaboration between industry and academia in the context of the legal track of the Text Retrieval Conference (TREC). To obtain data for the 2010 and 2011 legal tracks, the organizers cooperated with ZL Technologies, who had "acquired the full collection of 1.3 million Enron email messages from Lockheed Martin" (formerly Aspen Systems; Cormack et al., 2010). This dataset was later released by EDRM, an industry association of companies specializing in computational processing of text documents for the preparation of legal proceedings, as the "EDRM Enron Dataset, version 2" (Cormack et al., 2010). Containing a total of 1 227 255 messages, the EDRM v2 dataset is substantially larger than the CMU version and preserves the attachments. E-mails have been augmented with meta-data similar to what is present in the CMU dataset. Consistent with previous findings, a header field `X-Filename` references NSF and PST files.

As implied by the suffix "v2", EDRM had, at an earlier time, released a different version of the Enron data under the name "Enron PST Data Set". According to EDRM, this earlier dataset has also been compiled by ZL Technologies (Socha, 2010). The exact differences between the two versions are not documented, but an FAQ posted by EDRM describes v2 as "more complete" (Socha, 2010). EDRM has stopped providing access to both datasets over privacy concerns, but v2 can still be downloaded from the Internet Archive (2011). In 2013, EDRM partnered with Nuix to address these concerns by removing personally identifiable information from the Enron data, using the "PST Data Set" as a basis (Cassidy and Westwood-Hill, 2013). The result, called "Revised EDRM v1.3", is hosted by Nuix. According to Hermans and Murphy-Hill (2014), it contains 752 605 messages.

As a result of the decentralized delivery and storage of messages, any collection of e-mail may naturally contain duplicates: a message appears in the inbox of each recipient, and a copy is usually retained in a "sent messages" folder of the sender. Duplication may have also occurred at the time of data collection. Looking at the meta-data that originates from the `FILENAME` column of Aspen Systems' database, it appears as if in most cases, the messages belonging to a particular user were originally distributed across multiple Notes and Exchange data files, possibly including backups or snapshots taken at different times. Together with previous evidence that data was migrated from the Lotus-based to the Microsoft system, this makes a certain degree of overlap between these files seem likely. Finally, Klimt

[5] A conference paper by Klimt and Yang (2004), researchers at CMU's Language Technology Institute, counts 619 446 messages. Not having access to the exact version of the dataset they used, we are unable to explain the difference.

and Yang (2004), as well as Shetty and Adibi (2004), attribute a large amount of duplication to what they call "computer-generated folders". This is most likely an artifact of the unusual way Lotus Notes handles folders: a message can be simultaneously present in more than one folder. In addition, a message may be shown in one or more "views", virtual folders that are populated with the results of a database query. For example, any message, sent or received, can always be accessed via the "All Documents" view (HCL Technologies Ltd., 2019). An exporter that naively iterates over the structure of views, folders and subfolders, and exports each message it encounters as a separate entity, potentially creates a large amount of duplicates.

Several researchers have released deduplicated versions of the CMU dataset. Fiore and Heer (n.d.) apply an unspecified deduplication strategy and retain 255 636 messages. Shetty and Adibi (2004) remove a number of folders that are known to correspond to Lotus Notes "views" and keep 252 759 messages. Corrada-Emmanuel (2004) identifies 250 484 unique messages by their MD5 hash. Cormack et al. (2010) built a list of duplicates in the EDRM v2 dataset for the TREC 2010 legal track and count 455 449 unique messages. While the approaches differ, the results of the various deduplication efforts are consistent if the origin of the data is taken into account. The CMU dataset has a duplication rate of approximately 50%, while the EDRM v2 dataset has a somewhat higher rate of 63%.

Even after deduplication, EDRM v2 is almost twice as large as the CMU dataset. A shallow comparison between the CMU and EDRM v2 datasets by counting how often a particular file appears as the original source of a message shows almost complete overlap of the sets of source files. This observation admits two hypotheses: Either the two parties received the same data originally, but the CMU data was subjected to stricter filtering, or the CMU dataset was built from smaller, less complete versions of each source file. An in-depth comparison of the datasets that could rule out either hypothesis is out of scope for this thesis. There is no evidence that any e-mails that were reconstituted from printouts are included in any publicly available version of the corpus. They are not mentioned in the literature and the frequency distribution of words in the EDRM v2 dataset does not have the typical long tail one would expect in the presence of OCR errors.

Having to resort to conjecture about the nature of the data highlights the main issue with the Enron corpus. The variants of the data that are currently available to researchers clearly have a history of multiple data conversions and other preprocessing. Even if we assume that all custodians of the data intended to preserve the original data as well as possible, each additional processing step may cause a loss of fidelity. It is unfortunate that key processing steps have not been sufficiently documented. Proper, uninterrupted documentation of the chain of processing would give users of the data confidence that there are no undiscovered, systematic deviations from the original data. Consequently, for our experiments, we would like to use the dataset that is closest to the data originally collected by FERC, and perform our own preprocessing. Based on what we know about the provenance of the different versions, we choose the EDRM v2 dataset[6].

[6] See section "Legal Notices" at the end of this document.

HackingTeam

Compared to the Enron corpus, the provenance of the HackingTeam dataset is less complex. Internal company data was leaked to the public Internet (Greenberg, 2015). This data includes the mailboxes of 60 employees in PST format, presumably exactly as they were stored on the company's e-mail server. Due to the controversial and politically sensitive nature of the services provided by the company, the leaked data has been widely circulated and subjected to public scrutiny. While it is theoretically possible that the person who leaked the data also manipulated it with malicious intent, neither the public examination of the data nor our experiments have uncovered any inconsistencies or other traces of manipulation. Since HackingTeam is an Italian company, the dataset contains e-mail in English and Italian language.

3.3.2 Cleaning and Preprocessing

The process of converting the EDRM v2 and HackingTeam datasets into clean databases of messages and their meta-data can be broadly divided into two steps. The first is building a list of e-mail address aliases, the second is the extraction and cleaning of the textual message content and the relevant meta-data. In the EDRM v2 dataset, messages are stored in individual EML files in RFC 822 format, while the HackingTeam data is distributed in PST format. Due to the peculiarities of each format, the processing differs in some details.

Detection of Address Aliases

It is a common occurrence that two different addresses refer to the same mailbox. For one, multiple domain names may be handled by the same MTA. For example, HackingTeam uses the international domain "hackingteam.com" and the national "hackingteam.it" interchangeably. For the other, the MTA is free to map addresses with different local parts to the same mailbox. For example, former Enron CEO Kenneth Lay was using at least 13 addresses, ranging from the brief "klay@enron.com" to "kenneth.l.lay@enron.com". Microsoft Exchange adds a special case: If a message can be delivered locally, sender and recipients are not identified by their e-mail addresses but by the LDAP identifier of their user accounts. Furthermore, an actor may use multiple, separate mailboxes, for example, a company-provided e-mail account and a personal account provided by the home ISP. When constructing a social network from a collection of e-mail, it is therefore necessary to decide if two addresses belong to the same actor. To that end, we treat one address as canonical, and all other addresses used by the actor as aliases of that address. Whenever an alias is encountered during the processing of a message, it is replaced by the canonical address.

The identification of e-mail aliases in a collection of messages can be reduced to a binary classification task: Does a pair of addresses belong to the same actor? For this classifier, the uniform cost model is not appropriate, since the effects of a misclassification on the social network graph differ in severity: If two addresses are misclassified as belonging to separate actors (false negative), the actor is represented by two nodes, which are likely to be structurally similar, in the sense that their neighborhoods overlap to a high degree. If the addresses of two separate actors are misclassified as belonging to a single one (false

positive), the neighborhoods of two unrelated actors are merged, which may create short paths between actors that are distant in the true network, and therefore has the potential to significantly distort the network structure.

We therefore apply a heuristic method that is known to have a very low false positive rate at the cost of a higher chance of false negatives. We start from an initial set of aliases that reflects our prior knowledge. For the HackingTeam data, this is a list of manually identified equivalences, for the Enron corpus, a list of aliases compiled by the SNAT project (Google Code Archive, n.d.). For detecting aliases in the Enron dataset, we make use of correspondence between certain message headers and external meta-data. For each message in the dataset, addresses are extracted from the header fields `From` and `X-ZL-From`. The content of the latter is most likely derived from a column of the database provided by Aspen Systems, which in turn may be derived from meta-data that was originally stored in the PST file the messages were extracted from. If the two addresses differ, they are treated as aliases of each other. Since `X-ZL-From` appears to always contain a real e-mail address, this simple approach is sufficient for building a complete mapping between LDAP identifiers and e-mail addresses.

For the HackingTeam data, we use a method similar to that of Zhou et al. (2007), which makes use of external meta-data provided by the PST file. If a message is stored in a folder called "Sent", "Outgoing", or "Drafts" (or their Italian language equivalents), then the sender address must be the address of the mailbox owner. Under the assumption that each PST file corresponds to the mailbox of a single actor, the sender addresses of messages in these folders can be treated as aliases. Sometimes, a message has been mistakenly placed in a folder for outgoing mail by the user, and must not be used for the purpose of alias detection. If a message contains a `Received` header field, which is inserted by an MTA at the time of delivery, it must have been received by the mailbox owner and is therefore ignored. In both datasets, some (manually identified) messages and addresses must be ignored in order to avoid false positives, e.g., generic addresses that are not linked to a particular employee ("support@hackingteam.com").

The correspondence between PST meta-data and certain transport headers can be exploited to find even more aliases. Two heuristics are used: First, if the sender address of a message, as recorded in the PST meta-data, does not have a HackingTeam domain name as its global part, then a list of addresses is built by collecting this sender address, the "representing e-mail address"[7] from the PST meta-data (present if the sender is acting on behalf of someone else), and all addresses than can be extracted from the header fields `From`, `Sender`, and `Return-Path`. If this list contains more than one distinct address, these addresses are treated as aliases. Second, for each recipient of a message, the "e-mail address" (an arbitrary address, which may be an LDAP identifier) and "SMTP address" (strictly an e-mail address), as recorded in the PST meta-data, are compared. If they are different, they are treated as aliases.

In addition to these unsystematic aliases, there are a number of systematic variations that

[7] Here, and in the remainder of this section, quotation marks indicate that a term is used as defined by Microsoft in the PST file format specification (Microsoft Corporation, 2020) and the suite of Exchange Server protocol documents (Microsoft Corporation, 2019).

can be applied to an e-mail address without changing the mailbox it refers to. To avoid having to explicitly list all possible variants, we transform addresses into a canonical form before comparing them. First, all letters are transformed to lower case. Second, all domain names that are known to be used by Enron and HackingTeam are replaced with the strings "enron" and "hackingteam", respectively. For example, the address "Kenneth.Lay@ect.enron.com" would turn into "kenneth.lay@enron". When dealing with e-mail addresses from the Enron corpus, we also remove certain invalid characters and restore addresses that were mangled, most likely in the course of the migration from Lotus Notes to Exchange. Mangled addresses start with "imceanotes-", followed by a string that has been encoded with a custom variant of percent-encoding (Internet Engineering Task Force, 2005, section 2.1), which uses '+' instead of '%' as the escape character. Decoding that string yields the original content of the `From` header field.

Cleaning of Textual Content and Meta-Data

For the actual processing of messages, we iterate over all files in EML or PST format that belong to the dataset. A PST file contains the aggregated data of a single Outlook user: different kinds of entities (messages, appointments, contact details, etc.), arranged in a tree structure of folders. PST files are processed by walking the tree of folders in breadth-first order, and performing further processing on each entity that can be identified as a message by having a "message class" that starts with "IPM.Note". One specific file, "support.pst" of the HackingTeam dataset, is skipped entirely, because it is known to contain mail that is not associated with a particular actor, but rather with a role (technical support) that may be filled by different actors concurrently or in rotation. An EML file, by convention, contains a single e-mail message, but may also contain other data when used as a format for data export from Outlook / Exchange. To exclude any non-message entities from further processing, an EML file is skipped if a header field named `Content-Class` is present and its value is not "urn:content-classes:message". From each individual message, the relevant textual content and meta-data is extracted.

An EML file contains the raw message as delivered by the MTA. If the message has a simple (i.e., non-MIME) message body, it is processed directly. Otherwise, we walk the tree of MIME parts, process each part individually, according to its format, and concatenate the extracted text. Content of a type that does not start with "text/" is ignored. If a MIME part contains multiple alternatives, plain text is preferred over HTML. Messages inside a PST file have already been preprocessed by Exchange. Upon receipt of a message, Exchange separates the header from the body, aggregates the textual content from all MIME parts, and stores them in a common format, either HTML or text. Styled text sent between local users of the Exchange server is stored as HTML encapsulated within an RTF document for easier rendering on the client side. The original HTML is preserved inside of special sections that are ignored by regular RTF processors, and can be recovered without loss of information (Microsoft Corporation, 2016).

The textual content of an e-mail message is not entirely unstructured, but can be divided into several parts with different communicative functions (Lampert et al., 2009). Not all of them are directly relevant to expressing the sender's communicative intent. In a corpo-

rate environment, it is common practice to have the e-mail system automatically append a signature to outgoing messages. This signature may contain the company name, a slogan, contact details of the sender, or a legal statement. Even if the content of the signature is left to the discretion of the employee, and therefore reflects the personality of the sender to some extent, its contribution to an individual communicative act is negligible. Furthermore, statistical methods of NLP that are based on word frequency are susceptible to assigning undue weight to repeated content. Signatures are not the only source of duplicate content; a similar problem is posed by a convention for replying to earlier messages: When replying, it is common to extensively quote from the original message, often to the point of including the whole text of the original message verbatim. Over the course of conversational back and forth, the amount of duplicated content grows. It is therefore highly important to clean the message body of signatures and quotations from earlier messages. The steps we take to that end depend on the format of the text.

If the body is plain text, we iteratively examine each line. A signature is often separated visually from earlier parts of the message, e.g., by a sequence of dashes or asterisks. If the current line matches one of these patterns, this line and all following lines are discarded[8]. Forwarded or quoted message content is usually introduced by a specific phrase, e.g., "On February 29, John Miller wrote:". Note that there are two conflicting conventions for quoting from earlier messages: If the reply precedes quoted content, this is known as "top quoting" or "top posting". The older convention of placing the reply below the quoted content is known as "bottom quoting / posting". There are no recent studies on the prevalence of each quoting style, but considering that Outlook and many mobile e-mail clients default to top quoting, it is reasonable to assume that this style is more common in a corporate environment. We attempt to account for both styles by dividing the text patterns that introduce quoted content into two sets. The first set of patterns is used by e-mail clients that let users choose between top and bottom quoting, and prefix each quoted line with the character '>' as an indicator. On encountering one of these patterns, we only discard the matching line and all following lines that start with a quotation indicator. The second set of patterns is used by clients that enforce top quoting, and therefore do not specifically distinguish quoted from original content. On encountering a pattern from the second set, all following lines are discarded unconditionally.

Content in HTML format is identified by a content type that starts with "text/html". In some cases, HTML content was found to be mis-specified as plain text. Therefore, if the body of a plain text message begins with `<html` or a DOCTYPE declaration, it is treated as HTML. The HTML data is processed by constructing a DOM tree (WHATWG, 2020), iterating over its nodes in depth-first order, and appending the content of text nodes to an output buffer. When visiting an element node with a tag name of `br`, `p`, or `span`, a line break is appended to the output. Signatures and quoted content are identified heuristically. If an element has a tag name of `blockquote`, `pre`, or `div`, and has an attribute named `type` or `class` with a value that is associated with unwanted content (e.g., element `div` with attribute `class`, value `gmail_quote`), the element's children are ignored. As a notable exception, we keep `div` el-

[8] A positive side effect of this heuristic is that it also removes the legal notice that EDRM and ZL Technologies have seen fit to append to each and every message in the EDRM v2 dataset.

ements with an attribute `type` of value `moz-cite-prefix` or `moz-forward-container`, which were found to sometimes contain a mixture of original and quoted content. The detection of quoted content in a HTML-formatted message generated by Outlook is particularly difficult, as the generated elements do not provide any hints about their role or function. We identify the horizontal line that separates original from quoted content, a `div` element with an attribute of name `style` and value `border-top:solid`, and stop the traversal of the DOM tree if this element is encountered, effectively discarding all text below that line. After all nodes have been visited, the extracted text is subjected to the plain text cleanup procedure described above.

From the message header, we extract the following meta-data: sender (field `From`), recipients (`To`, `Cc`, `Bcc`), unique message identifier (`Message-Id`, or randomly generated if the field is not present), and timestamp (`Date`). We keep a list of addresses that are ignored if they appear as sender or recipient of a message. This list contains senders of automated messages and addresses that are jointly used by multiple actors. Messages that do not have a sender and at least one recipient are discarded. The subject line (`Subject`) is prepended to the text extracted from the message body. Messages from the bilingual HackingTeam dataset are then subjected to language detection.

Messages that reply to or forward earlier messages are detected via two independent heuristics. If its subject line, transformed to lower case, starts with "re:", "fw:" / "fwd:", or their Italian language equivalents, a message is correspondingly marked as a potential reply or forward with unknown parent. The message ID of the parent, i.e., the earlier message that is replied to or forwarded, can be extracted from a number of different header fields. These are, in order of precedence, `X-Forwarded-Message-Id` (forwarding only), `In-Reply-To` (replies only), `References`, and `Thread-Index` (a proprietary extension; Microsoft Corporation, 2018, section 2.2.1.3). While replies can be detected reliably, forwarded messages are not always linked to the original (Mozilla Corp., n.d.).

To the extracted messages, we apply the same deduplication procedure as to the Facebook data, with the added condition that two messages with the same ID are considered equal. The message ID is generally considered to be reliable, but, like all natural IDs, its uniqueness cannot be guaranteed. In the HackingTeam dataset, a single case of two different messages with the same ID was found. When comparing two pieces of text extracted from a message body, all whitespace, including line end markers, is ignored, to account for a variety of ways a message may be wrangled during transfer, by the receiving client application (conversion between DOS and UNIX style line endings, line wrapping), or during pre-processing (conversion of RTF to HTML to plain text). If two or more messages are deemed identical by the deduplication algorithm, but have different lists of recipients, they are merged into a single message, addressed to the union of recipients.

The final Enron dataset consists of 153 552 messages from 158 mailboxes, sent and received between 1998 and June 2002. Among 253 052 valid messages extracted from the EML files, 0.5% were empty, 37.8% were discarded as duplicate, 0.6% were merged due to content-wise equivalence, and 0.4% were merged by ID. The final HackingTeam dataset consists of 350 338 messages from the mailboxes of 60 employees, sent and received between October 2005 and July 2015. Among 1 069 687 valid messages extracted from the PST files, none were empty, 65% were discarded as duplicate, 1.4% were merged by content, and 0.8% were merged by ID.

The content-based merges can possibly be attributed to groups of automated messages with the same content that are sent at the same time, but separately, to different recipients. The ID-based merges point towards undiscovered address aliases, but may also be caused by the use of BCC. Recipients on the BCC list are invisible to anyone but themselves, so any BCC recipient sees a different subset of the true recipient list.

While the Enron mails are expected to be mostly written in English language, the HackingTeam dataset contains a large proportion of Italian messages: 45.0% of the messages are English, 50.5% are Italian, and the remaining 4.5% cannot be confidently assigned to either language. To comply with the single-language requirement, we split the dataset by language ("HT-en" and "HT-it"), but keep the original, multi-lingual dataset ("HT") for use in any language-agnostic analysis.

Compared to the message volume of the various deduplicated versions of the Enron corpus discussed earlier, the amount of messages in our version is unexpectedly small. The origin of the difference is that of 1 233 085 EML files provided by the EDRM v2 dataset, 687 335 (55.7%) could not be processed, because they lack a valid sender, recipient, or both. Upon further investigation, the problematic messages can be divided into two groups: In one group, sender or recipients are only specified by their display name (e.g., "John Miller"). Without an e-mail address or other unique identifier, actors of the same name cannot be distinguished. In a small sample of manually inspected messages, this issue was limited to messages imported from Notes. Another group of messages appears to be calendar entries / appointments stored in EML format, which do not specify any sender or recipient at all. Unambiguous information about the identity of sender and recipients is required to reconstruct the social network, so both groups of messages cannot be used in our experiments. Comparing a small number of problematic messages to their equivalents in the CMU version of the Enron corpus, we find that the CMU version does contain valid e-mail addresses. It is unclear how these addresses were obtained, and Zhou et al. (2007) identify some data quality issues specifically with the e-mail addresses in the CMU dataset. Still, the CMU dataset might have yielded a higher amount of usable messages than EDRM v2, despite being smaller.

3.4 Common Representation of Data from Different Platforms

In spite of their outward differences, the three social platforms discussed in the previous sections offer a common set of core functionality. The first step towards being able to perform the same experiments on each of the datasets is defining a set of terms that allow us to talk about online communication data in a generic, platform-independent way.

The message is the basic unit of online communication. On all of the social platforms under investigation, the sender, the textual content, and the date of submission (or publication) of a message are clearly identifiable, but the definition of what constitutes a recipient is more involved. To better understand the difficulty of identifying the recipients of a message, we draw upon two basic characteristics of messages, *addressivity* and *originality*. In case of *addressive communication* (Honeycutt and Herring, 2009), the sender of a message explicitly designates one or more recipients, thus demonstrating individual awareness of the recipients. One may reasonably assume that the sender considers the message content to

be relevant to the recipients. Conversely, *non-addressive communication* constitutes "broadcasting" a message to an undisclosed group of people. Since even the sender may not be aware of the composition of this group, there can be no individual awareness. On most platforms, the addressivity of a message affects its visibility to others, but addressivity and visibility are distinct concepts: a message may be visible to a large group of people, but still be written with specific recipients in mind.

Another attribute of messages, independent of their addressivity, is whether they convey *original* content or re-transmit information received from others. For example, on Twitter, retweeting is a means for information sharing within the medium, while tweeting URLs introduces content from the outside. The act of re-transmitting, usually referred to as "sharing", is a compromise between appropriating someone else's content by re-publishing it under ones own name, and keeping distance to the content by attributing it to its original author. Sharing a message is a complex social signal with three distinct aspects: It confirms that the sharer has read the message and attaches some kind of relevance to its content. If the platform notifies the original author that the message is being shared, then the author, depending on the context, may treat this as positive or negative feedback. Finally, as a consequence of sharing, a new group of users is exposed to the message.

We argue that addressivity and originality are fundamental concepts of computer-mediated communication, and are supported, in some form, by all online social platforms. E-mail communication, for example, is addressive by design, yet non-addressive communication can be achieved via mailing lists or sufficiently large recipient lists (e.g., team- or company-wide mails), and mail forwarding and attachments enable information sharing. Conversely, most "modern" social media are non-addressive by default, and do not require messages to be explicitly associated with a group of recipients. For each platform, we can define criteria for classifying a message as non-addressive or addressive and its content as either original or shared, and define rules for deriving a set of recipients from the observed data.

For a given dataset, we distinguish between *core*, *non-core*, and *unobserved* users. A user is counted as a core user if we have reason to believe that the dataset contains a representative sample of that user's communication. This is the case if the user has been explicitly visited during the crawling process, and messages were successfully retrieved. When working with found data, the criteria depend on the dataset. Non-core users are known implicitly, e.g., by being listed among the recipients of a message, or by being neighbors of a crawled user in the explicit social network graph. All other users of the social platform are unobserved. A *core message* is one that is sent by a core user, and is either non-addressive, or has at least one core user as a recipient.

The specific rules that we use for determining addressivity, originality, and the recipients of messages on each social platform, as well as the "core-ness" of its users, are as follows:

Twitter Each tweet has a unique author, but lacks an explicit list of recipients, so the set of intended recipients has to be reconstructed heuristically. Depending on whether or not a tweet is a reply, it is visible to different groups of users in different ways. While, technically, all tweets are public and can be accessed via the author's profile page, Twitter's search facilities, etc., we assume that a user is highly unlikely to be aware of a particular tweet if he or she is not explicitly exposed to it in some way. We call

a tweet *visible* to user a if it is shown on that user's news feed. If a replies to a tweet of $b \neq a$, the reply is only visible to a, b, and users who follow both a and b. All other tweets are visible to all followers of the author.

Given these visibility rules, we classify any reply as addressive communication with the author of the original tweet, so that each addressive tweet has exactly one recipient. A tweet that is not a reply is classified as non-addressive. Its visibility to a potentially large group of followers, who the sender may not even be individually aware of, makes it unlikely that the tweet was written with particular recipients in mind. Although a regular tweet may contain @-mentions, it is difficult to determine whether the sender's intention is to mention these users to his or her followers, or to actually address them (cf. Honeycutt and Herring, 2009). For example, people use the hashtag "#ff" ("Follow Friday") to recommend interesting users to their followers (Leavitt, 2014). A typical "Follow Friday" tweet simply consists of the hashtag and a list of @-mentions, which clearly have no addressive function.

Retweeting maps directly to our definition of sharing, so any non-retweet (i.e., a regular tweet or a reply) is treated as original. Due to the difficulty of unambiguously identifying manual retweets, only retweets made via the UI are counted as shared content. Any user that was visited by the crawler is counted as a core user, but for computational reasons we usually work with a subset of the data that only includes the first 30 000 core users.

Facebook As in the case of Twitter, the author of a post or comment on Facebook is uniquely defined, but the recipients have to be determined heuristically, using visibility as a proxy for the author's intentions. We call a post that is authored by user a visible to user $b \neq a$ if it is explicitly posted to b's timeline or b is mentioned in it, and therefore receives a notification about it. The visibility of comments is harder to define. The author of a post is notified of all comments. If user a comments on a post authored by someone else, a is notified of all subsequent comments on that post. A possible explanation for this notification scheme is that, at the time of crawling, comments could not be nested, that is, comments are always associated with a post, and comments cannot receive comments of their own. Since there is no reliable way of telling if a comment is addressing the post or one of the earlier comments, all commenters are notified to enable their further participation in the discussion.

Facebook does not provide a general mechanism for addressive communication like "replying" on Twitter, so we have to be careful to identify all commonly used expressions of addressivity. Posts to the timelines of others and comments on the posts of other users are treated as addressive. Posts on a user's own timeline are treated as addressive if they mention other users, and non-addressive otherwise. If a post mentions more than 10 users, it stands to reason that the author is no longer individually aware of each of them, so we also count such posts as non-addressive. While one could, in accordance with visibility, extend the set of recipients of a comment to include all earlier commenters, this would make the number of recipients dependent on the attention a post receives. Particularly in the case of posts with a large comment volume,

it is unlikely that every commenter has read all previous comments. Since we can be sure that every comment is somehow related to the original post, but cannot reliably tell if a comment is a reaction to one of the earlier comments, we only consider the author of the original post as the intended recipient.

Facebook provides a function for sharing other users' content, which matches our definition of sharing. Any user that was visited by the crawler is counted as a core user.

E-Mail Electronic mail has an explicit sender and at least one explicit recipient. An arbitrary number of recipients can be addressed directly (via the header field `To`), or listed as secondary recipients that receive a copy of the message. The identity of secondary recipients is either visible to all other recipients (`Cc`) or hidden (`Bcc`). Otherwise there is no difference in the delivery of messages to the three types of recipients, so we make no distinction between them. Since there is no explicit indicator of non-addressivity, we use an ad-hoc threshold of 10 recipients to decide whether a message is addressive or non-addressive in intent.

The mechanism that is closest to our concept of sharing is the forwarding of a received message to someone else, so any forwarded message is counted as shared content. If the dataset contains the complete mailbox of a user, that user is counted as a core user.

After applying these rules to the four datasets, we see pronounced differences in the ratio of addressive to non-addressive messages, while the proportion of shared content is consistently low. In the Twitter dataset, approximately 75% of messages are core messages. Among these, 5% are addressive and 80% are original. In the Facebook dataset, 47% of messages are core messages; 47% of these are addressive and 99% are original. The high proportion of original messages reflects the technical difficulty of identifying shared content when retrieving data from Facebook via screen scraping. In the Enron corpus, only 6% of messages are core messages; 82% of these are addressive and 94% are original. We have to assume that the bulk of messages in the corpus are interactions with non-employees or employees whose mailboxes were not archived in the course of FERC investigations. In the HackingTeam dataset, 48% of messages are core messages; 94% of these are addressive and 95% are original.

A direct consequence of defining originality, addressivity, and the identity of recipients via custom heuristics for each platform is that inherent characteristics of the platform are difficult to separate from effects of the choice of heuristics. Still, it is reasonable to assume that e-mail, by design, favors addressive communication, while current SNS encourage non-addressive "broadcasting" by providing each user with a bespoke audience of followers or friends. From a different point of view, one might conclude that addressivity and originality are not truly dichotomous: messages can be completely non-addressive, or exhibit varying degrees of addressivity, from making a passing reference to someone in a public statement up to being confidential communication between sender and recipient. In the process of sharing, the source material may be modified or augmented. However, a more fine-grained modeling of these aspects is out of scope for this thesis.

3.4.1 Data Quality Issues

From the discussion of the different social platforms and the respective mechanisms for data collection, one can identify a number of general issues that may negatively affect the quality of the collected data and inferences drawn from that data:

Missing data Due to the limited amount of resources that can be spent on crawling, an exhaustive observation of interactions on a social platform is impossible. Some interactions are principally unobservable, because they are not publicly accessible (e.g., due to restrictive privacy settings) or not being archived anywhere (e.g., e-mails that have been deleted by the sender and all recipients).

Corrupted data Inconsistencies in the data may be the result of intentional manipulation, like in the case of backdated Facebook posts or falsified timestamps of spam e-mail. In all collected datasets, we also observed a proportion of unsystematically corrupted records, possibly due to software defects.

Systematic bias The available data may not be representative for the medium as a whole. Both e-mail datasets originate from a workplace communication setting, while the observable Facebook communication is limited to a subset of posts that were explicitly made accessible to the public. The datasets were collected over different periods of time, which may also have a biasing effect due to global shifts in the way people use online social media. Other, yet undiscovered bias may be present. None of these individual datasets can claim to be representative of human social behavior in general.

Differences in presentation While the examined social platforms offer a common set of basic functionality, differences exist in the visual presentation of messages, the visibility of messages to other users, and in how interacting with messages affects their visibility. These differences have to be considered in any comparative analysis of the data.

Three specific issues deserve further elaboration: cultural bias, the volatile nature of explicit relationships, and automated messages. Cultural and societal bias, i.e., the overrepresentation of certain demographic or social groups in a dataset, threatens the generality of claims that are founded on inference from the data (Olteanu et al., 2019). Some of this bias can be attributed to the platform itself and the people it attracts, while some has to be attributed to the crawling process and the restrictions placed on the crawler (starting point, sample size, desired languages). Due to homophily (the tendency to associate with people that are similar to oneself), a subgraph obtained by BFS-style crawling may exhibit more homogeneity than a random sample in personal attributes such as gender, ethnicity, age, economic status, or political stance. The Facebook dataset exhibits a measurable bias towards younger, male users, who are located in the north-western USA (see section 3.2.1), and all datasets almost exclusively represent Western culture.

A problem of any platform with explicitly declared relationships is that the network of explicit relationships is dynamic, i.e., nodes and edges may be created or removed at any time, but one can usually only observe the presence or absence of a relationship at the exact

moment of crawling a particular user profile. Neither Twitter nor Facebook report when a relationship was established (cf. section 3.1.1, "Data Quality"), and it is generally impossible to know for how long it persisted after crawling. For the sake of simplicity, all upcoming experiments assume, where necessary, that any observed relationship has existed and will persist indefinitely.

The data quality issues discussed up to this point are either inherent to the medium or associated with the process of data collection and analysis. A new kind of issue arises from users of a social platform who operate their accounts in a semi- or fully automated way. Recently, the term "social bots" has been applied to these accounts (Bessi and Ferrara, 2016). Varol et al. (2017) estimate that 9–15% of active Twitter accounts are social bots, and distinguish spammers (bulk senders of unsolicited commercial offers), self-promoters, political actors, and connected applications that send messages on behalf of a real user (e.g., a weight loss app that automatically posts a weekly progress report). Other social platforms are similarly affected by bots. Some types of bots are ubiquitous on all platforms (spammers), others are highly specific to the platform (computer viruses spreading by e-mail). Automated messages are a source of systematic bias: regardless of their nature as spam, manipulation attempt, or legitimate information, they are typically sent in high volume and their content follows repetitive patterns. Furthermore, bots differ from real users in their interaction patterns (Varol et al., 2017).

For the task of bot detection on Twitter, Varol et al. (2017) propose a supervised classifier that bases its decision on more than one thousand features, and achieve an accuracy of 0.85 AUC on unseen data. Bot detection clearly is a complex problem, so in this work, we deliberately limit our efforts to basic preventive measures. In the acquisition of Twitter data, some measures are taken to avoid crawling accounts that belong to spammers. In the preprocessing of the HT e-mail data, certain mailboxes that are known to be sources of high-volume automated messages (e.g., nightly status reports of various software systems) are skipped. These efforts are not subjected to any quantitative evaluation.

3.5 Temporal Characteristics

The acquisition of data from a social platform is usually constrained by the available resources, and the resulting dataset can only be a sample from a much larger population. Often, the acquisition is not only limited in the number of user profiles that can be visited, but also in the number of messages that can be retrieved from each profile: the Twitter API imposes a rate limit on requests, Facebook's privacy settings allow users to selectively hide messages, and e-mail users may have different retention strategies for old messages. Some users indiscriminately keep all messages, some delete messages they no longer consider relevant, and others delete all messages above a certain age. Consequently, the dataset is not only a sample from the user base of the platform, but also a sample from the stream of messages over time.

To be able to perform experiments on datasets from different social platforms in a way that yields comparable results, differences in the temporal distribution of messages have to be taken into account. The interval of time that contains the bulk of messages and the presence

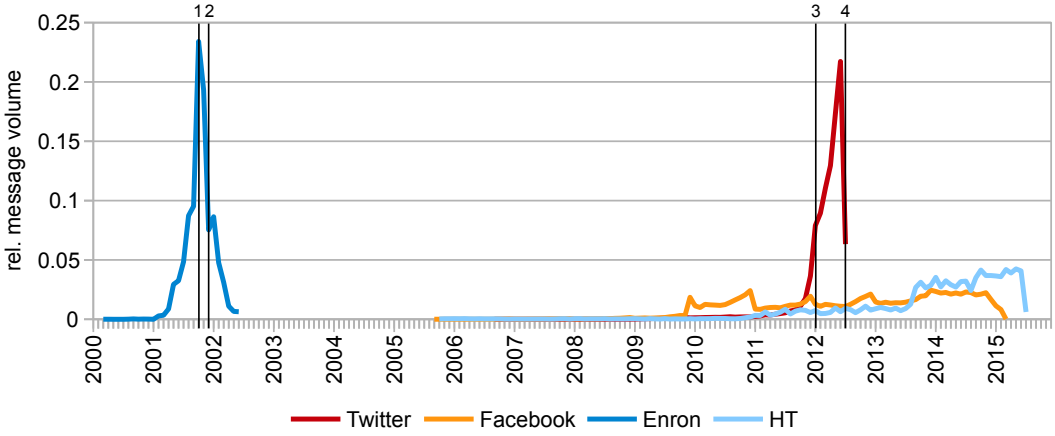


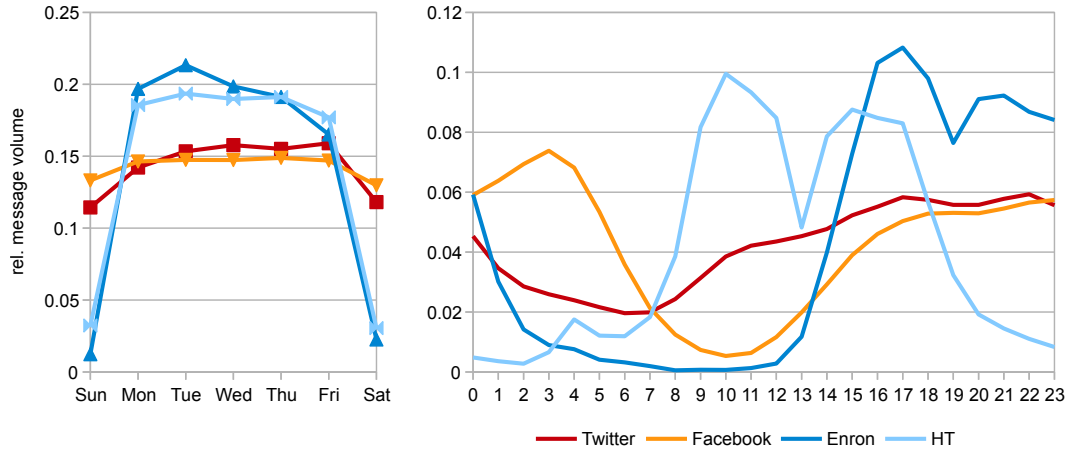
Figure 3.6: Monthly relative message volume

or absence of regular daily and weekly fluctuations can be determined by simple visual inspection of the temporal distribution of message volume. Beyond that, we are interested in the temporal rhythms of human communicative behavior. Our experiments on social influence (chapter 5) involve the comparison of behavior in two successive intervals of time, in other words, a temporal quantization of the observation data. The pertinent question is whether suitable quantization intervals can be derived from patterns found in the data.

3.5.1 Message Volume

Figure 3.6 plots the distribution of message volume of each dataset over time. The Enron corpus is the oldest dataset, containing messages sent between March 2000 and June 2002. The amount of messages dated before 2001 is negligible. The message volume peaks in October 2001 (labeled '1'), around the time when the accounting scandal of Enron Corporation was publicized, and sharply drops in December 2001 ('2'), following the bankruptcy of Enron and the subsequent lay-off of thousands of employees (Wikipedia, 2020a). Diesner et al. (2005) note that year-end vacations may have also contributed to the reduction in volume. The HT data is more uniformly distributed over time, with most of the messages dating between 2011 and 2015, and a long tail of older messages going back to 2005. The level of activity remains constant for most of the time, but rises sharply in January 2011, mid-2013 and mid-2014. Without manual inspection, one can only speculate about possible causes of this step-wise increase in activity, for example, the hiring of new staff, or the introduction of a new internal software solution that sends automated messages.

The posts and comments collected from Facebook are rather uniformly distributed over the years 2010 to 2014, preceded by a long tail of messages reaching back to 2005, and an increase in volume in the second half of 2009, which matches the growth in popularity of the SNS at that time. There is a visible yearly trend of increasing message volume that coincides with the US "holiday season", which includes Thanksgiving, Christmas, and New Year, followed by a drop in activity in January. People may be generally more active on Facebook over the holidays, but the communal nature of the festivities may also invite a



(a) weekly **(b)** daily (UTC+0)

Figure 3.7: Weekly and daily relative message volume

high amount of "season's greetings" posts that are publicly shared rather than kept private, and are therefore accessible to our crawler.

In the case of the Twitter dataset, the temporal distribution of tweets reflects the crawling policy as well as particular internal mechanisms of Twitter. For each user, the crawler retrieves all tweets between the time of crawling and January 1, 2012, via a series of API requests. Since each request yields up to 200 tweets, and retweets include a copy of the original tweet, a certain amount of older tweets may be fetched as well. This accounts for the long tail reaching back to the very first tweet in 2006, and the gradual increase in message volume leading up to the start of the crawling period ('3'). Within the crawling period ('3', '4'), one would expect an approximately uniform temporal distribution. The observed burst of activity towards the end can be attributed to two factors: The presence of accounts that were created after the start of the crawling period implies an increase in message volume towards the end of the period. However, this class of users only accounts for 13% of the overall message volume. Due to unanticipated behavior of the Twitter API, the retrieval of tweets was terminated early for a subset of users. Since tweets are retrieved in reverse chronological order, this results in a bias towards more recent tweets. Finally, Twitter has placed an arbitrary limit on the API endpoint for retrieving a user's past tweets: only the 3 200 most recent tweets can be accessed (Twitter, Inc., n.d.b).

Figure 3.7 shows the average weekly and daily (UTC+0) message volume. In a corporate setting, as reflected by the Enron and HT e-mail datasets, the message volume falls close to zero on weekends, while on social media there is only a slight reduction in volume. The Enron dataset exhibits a gradual decline of message volume towards the weekend, while, conversely, the message volume on Twitter slightly increases. Apart from that, the message volume is fairly evenly distributed over the working days in all datasets.

As one might expect, given their origin in a workplace setting, the daily patterns of the Enron and HT dataset are very similar: a sharp onset of activity in the morning, a dip at

lunchtime, followed by gradual decline until the end of regular working hours, at which point we see a steeper decline to nocturnal levels. The nightly activity level in the Enron dataset is close to zero, and somewhat higher in the HT dataset. This may be, in part, attributable to different working habits in the smaller, more tech-focused company, but may also be indicative of an increase in the volume of automated messages between 2000 and 2010. Similarly, the small peak at 4 AM may be caused by regular internal automated messages, e.g., server status reports.

While the employees of Enron and HT were expected to be physically present at the companies' offices, the Facebook and Twitter users that were crawled to obtain the respective datasets are known to be geographically spread across multiple time zones. Krasnow Waterman (2006) reports that "the majority of the emails were sent or received in Texas at the Enron headquarters". After geocoding 8 139 users of the Twitter dataset, Grob (2013) finds that 84% are located in the USA and the UK. Only a small amount of reliable geodata is available for the Facebook users (see section 3.2.1). Due to the choice of an US resident as the crawling seed and the restriction to English speakers, it is likely that the majority of users in the dataset is located in the USA. The temporal distribution of Facebook posts reaches its peak in the evening and has a minimum when it is nighttime for most users. The distribution of tweets still has a recognizable day-night rhythm, but is much closer to uniform due to the broader geo-spatial distribution.

Having explored the temporal distribution of individual messages, we turn to the analysis of temporal patterns in users' engagement with discussion threads.

3.5.2 Rhythms of User-Content-Interaction

In informal comparisons of online social media to traditional communication and news media on- and offline, the former are commonly said to be "faster". In this context, speed is not clearly defined, but may refer to the lifetime of content published on the medium: How long does a particular event, discussion topic, or trend receive attention? We propose a novel method of characterizing the *content lifetime* using a Hidden Markov Model (HMM), apply it to the comparative study of our social media datasets, and identify common time scales of social interaction. The presence of these time scales across the different social datasets motivates the use of fixed-length temporal quantization for the aggregation of message content in subsequent experiments.

Most online social networking and communication platforms offer a core functionality that can be described by a simple model of user-content interaction: Each user is provided with a custom, inverse chronological news feed that lists individual *content* items. The nature of the platform determines the type of content (messages, status updates, news articles, etc.) and the type of distribution (e.g., explicit recipients, push / pull models). Users interact with content items by replying to them or sharing them with others, thus increasing their visibility. In consequence, each item has an *audience* that grows monotonically over time. On the news feed, content is ordered by the time of the last interaction, so items under active discussion are at the top, and move downwards as interest fades. However, even old content can be located at any time, e.g., by a full-text search, and interacted with, bringing it back to the top. The temporal characteristics of content lifetime are therefore to a large

degree determined by the style of presentation and the possible ways of interaction (Szabo and Huberman, 2010).

We pose the following research questions: How can a meaningful definition of content lifetime in a social medium be derived from the temporal distribution of interactions? How do traditional and modern social media compare in terms of content lifetime? How does sharing affect content lifetime? To answer these questions, we analyze the data that was collected from different online social platforms.

User behavior in online social networking services can be studied on different socio-structural levels. At the micro level, timestamped interaction sequences of individual users reveal global rhythms of sleep, work, and daily life (Golder and Macy, 2011) as well as medium-specific usage patterns (Guo et al., 2009). The aggregated interactions of a group of users with one or more topically related content items are studied under the moniker *collective attention*. At the macro level, peaks in collective attention can be attributed either to endogenous build-up of attention through information diffusion within the network, or exogenous effects such as unanticipated real-world events (Lehmann et al., 2012). Our study of content lifetime sits in between these levels: While we analyze the collective attention a content item receives from its audience, we expect audiences to be small to the point of not being a representative sample of the platform user-base. Our work is similar in nature to that of Clauset and Eagle (2007), who identify a "natural time scale" that is optimal for the temporal quantization of dynamic social networks for the purpose of tracking the evolution of network structure.

Related studies of the aggregated interactions of multiple users with particular classes of content commonly use Twitter data due to the ease of acquisition via an officially supported API. For example, Kwak et al. (2010, section 6.2) provide descriptive statistics of inter-event times in retweet cascades, while Chalmers et al. (2011, section 5) analyze conversations among dyads and triads and manually identify patterns of activity on multiple time scales. Concrete definitions of content lifetime have been proposed in the context of online journalism (Castillo et al., 2014) and diffusion cascades (Lerman and Ghosh, 2010), which track the unmodified retransmission of information. A related problem is the prediction of the longer-term popularity of a content item from an initial sequence of interactions (Bandari et al., 2012).

Data Preparation

From each dataset, we extract items of user-generated content and the associated interactions. As illustrated in figure 3.8, the resulting subset of data has a tree structure, which is induced by the different types of interaction: Sharing creates a copy of the original content that is visible to a new audience, while replying creates new content that is linked to the original content, and can in turn be shared and replied to. Each interaction refers to exactly one parent item. The tree is flattened by arranging the nodes in temporal order. Representing the sequence in terms of the inter-event times δ makes the temporal dynamics of different sequences comparable. We only consider sequences of length two or above. In the sequel, we briefly describe the social platforms from which the datasets were obtained, and map platform-specific terms to our model of user-content interaction.

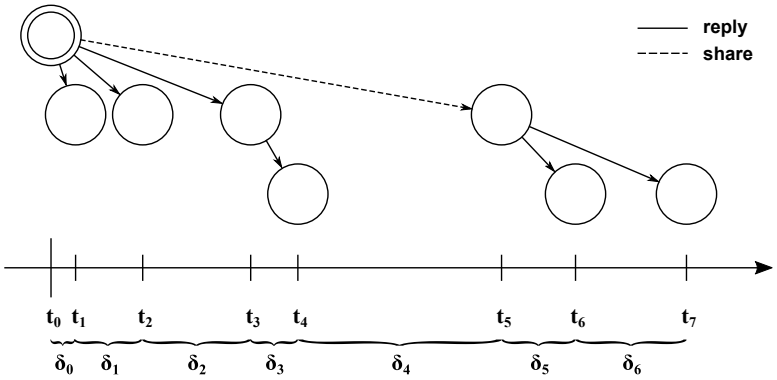


Figure 3.8: Transforming a tree of interactions into a sequence of inter-event times (Hauffa and Groh, 2019)

Twitter Tweets are public, but are only shown on the news feed of followers of the author. Sharing ("retweeting") exposes a tweet to the followers of the retweeter, while replies are visible to users who follow both the author of the reply and the author of the original tweet. Replies can form arbitrary long chains, but retweets cannot: if a' is a retweet of a, retweeting a' will create another direct retweet of a. Therefore, any tweet that is not a reply or retweet is counted as a content item, while retweets of the original tweet, replies at any distance to the original tweet, and retweets of these replies count as interactions. While the news feed of a user generally follows a chronological order, particular types of content are prioritized. The exact criteria are not published, and may include user preferences learnt via feedback mechanisms and intermixing of paid advertising (Twitter, Inc., n.d.a).

Facebook Posts are only visible to friends unless specifically made public, and shared content is only publicly visible if it was designated as public by both the original poster and the sharer. Like Twitter, the presentation of content on the news feed may deviate from chronological order. Posts can be shared and replied to in the form of comments that are directly associated with the post. Comments cannot be shared individually, and cannot be nested. Any original post is counted as a content item. Sharing that post and commenting on the post or any of its shared copies is counted as an interaction.

E-Mail The presentation of incoming messages to the user depends on the client software, but an inverse chronological, conversation-centric view has emerged as the usual form. Each message has a unique identifier, which is used to link replies to the original message. Sharing is realized as *forwarding*, where a received message is passed on to a different recipient in unmodified form. Any mail that is neither a reply nor a forwarded copy of another mail is counted as a content item. Replying to or forwarding of that mail or any of its replies or forwarded copies counts as an interaction. The two e-mail datasets, HT and Enron, are analyzed separately, and the original bilingual HT dataset is used.

A major problem for this study is posed by interaction sequences that cannot be fully

observed and remain incomplete or fragmented. Interactions are usually expressed as backlinks (e.g., from a reply to its parent). A limited notion of completeness of a sequence can therefore be established by following the links and checking for the presence of a root. If missing data separates a branch of the interaction tree from its root, or the whole branch is missing, the resulting state of incompleteness cannot be detected. In consequence, the collected data may underrepresent the amount and frequency of interaction, and the chosen method of analysis needs to be able to cope with incomplete sequences. Another data quality issue that directly affects this study is corruption and inaccuracy of timestamps. To eliminate inconsistency, we ignore interactions that predate the content item they refer to. We manually verified that all remaining timestamps are plausible, that is, they are neither before the launch of the platform, nor after the date of crawling. A peculiarity of Facebook is that users can arbitrarily backdate their posts. Timestamps of backdated posts always lie on a 30- or 60-minute boundary, and backdating only affects the root of an interaction tree, so it is possible to detect and ignore such posts with a low probability of false positives.

Finally, we note that due to limited resolution of the timestamps, the observed temporal distance between two interactions may be zero. In the Enron dataset, most of these simultaneous actions turn out to be automated "out of office" replies, which are generated by the server immediately upon receipt of an e-mail. The timestamps published by Twitter and Facebook are only accurate to whole seconds, so, given the large potential audience of a post, most simultaneous interactions can be attributed to chance. In a manually examined sample of tweets, all simultaneous replies appear to be natural. For example, a prominent user posting a controversial tweet may invite many angry replies, often simultaneous. Similarly, most simultaneous retweets in the sample are natural, and often revolve around topics of popular interest, but there is some evidence of spam activity. A group of (possibly automated) accounts was found to tweet promotional messages. Another group of accounts retweeted these messages as soon as they appeared, ostensibly in an effort to boost their visibility. In all three types of datasets, simultaneous interactions occur naturally, so we preserve them by representing them as distinct actions with a temporal distance of zero.

Table 3.2 compares the volume of data collected from the different social platforms. The first part of the table describes the volume of raw data. The second part of the table describes the extracted sequences of inter-event times. On all platforms, the sequence length has a long-tailed distribution, as evidenced by a mean and median close to the lower limit and a high maximum, making us expect an exponential decay in frequency with increasing length. Compared to Twitter and Enron e-mail, the Facebook and HT datasets noticeably favor longer sequences. The different characteristics of the two e-mail datasets may be due to differences in company culture or change in e-mail usage over time. The rarity of content sharing in the Facebook dataset is caused in part by the visibility rules for shared content, but also by difficulties of reliably determining the origin of a shared item in the data available to us.

Figure 3.9 compares the distribution of inter-event times in the different datasets, obtained via kernel density estimation with Gaussian approximation of bandwidth. For each dataset, we observe power-law behavior within specific lower and upper bounds, and exponential truncation at the upper bound. In each case, the upper bound matches the age of the platform or the timespan of collected messages, respectively. Twitter and Facebook favor shorter

Table 3.2: Descriptive statistics of the user-content interactions (Hauffa and Groh, 2019)

	Twitter	Facebook	E-Mail	
			Enron	HT
users	358 342	16 834	158	60
messages	223 013 844	3 912 832	153 552	350 340
sequences	5 796 548	465 869	1 031	25 409
% incomplete	5.7%	1.0%	25.2%	30.4%
mean (median) length	3.3 (2)	5.3 (4)	3.2 (2)	5.2 (3)
maximum length	3 578	3 428	44	157
interactions	19 223 692	2 462 609	3 306	132 192
% sharing	61.3%	0.2%	20.2%	8.0%

inter-event times below one hour, while the e-mail datasets, particularly HT, contain more longer inter-event times between a day and several months. Both e-mail datasets have a peak around intervals of one day, since most of the users are employees working within the same time zone and therefore likely to have similar daily rhythms.

Measurement of Content Lifetime

From a mathematical point of view, a sequence of interactions can be characterized by its inter-event times, i.e., the temporal distance of each interaction to its predecessor. In the absence of any limiting assumptions, events may be treated as uniformly and independently distributed over time and as such are suitably approximated by a Poisson process, which predicts exponentially distributed inter-event times. However, Barabási (2005) finds that in many settings, the inter-event times of individual human behavior follow a more heavy-tailed distribution, which manifests as *bursts* of high activity separated by longer periods of inactivity. In studies of collective behavior, the amount of attention a particular content item receives has been empirically linked to its visibility (Hodas and Lerman, 2012). Heavy-tailed distributions of inter-event times are inherent to systems where "high-priority tasks will be executed soon after their addition to the list, whereas low-priority items will have to wait until all higher-priority tasks are cleared" (Barabási, 2005). If, in the case of aggregated human behavior on a social networking platform, priority can reasonably be equated with visibility, a bursty temporal distribution of events is to be expected. Ghosh and Huberman (2014) examine aggregated interactions of users with different types of online content and observe a non-Poisson distribution of inter-event times, which they attribute to human activity on multiple time scales. We aim to identify these time scales using a Hidden Markov Model.

Fitting a Hidden Markov Model According to our model of user-content interaction, the frequency of interaction with a content item depends on its current visibility, which reflects the collective interest in the content. Visibility and collective interest are not directly

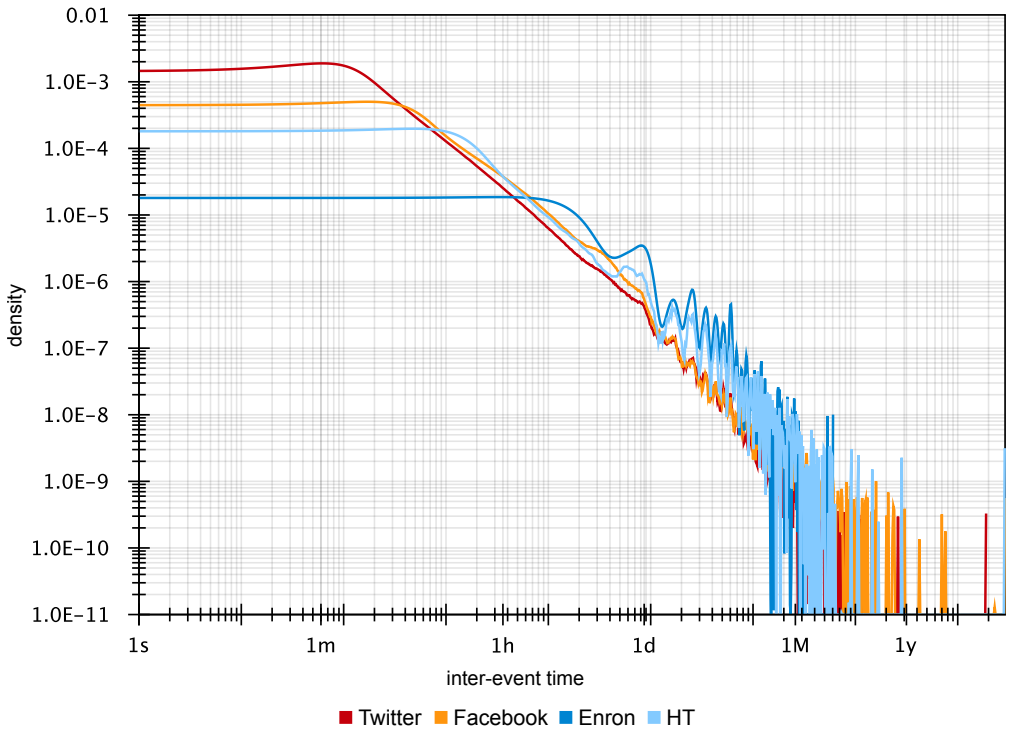


Figure 3.9: Kernel density estimates of inter-event times (Hauffa and Groh, 2019)

observable, and may interact with each other. We therefore collapse them into a single latent variable, *activity*, as illustrated in figure 3.10. Its joint distribution with the inter-event times can be described by an HMM if activity is represented as a quantity on a bounded ordinal scale, and if its temporal development can be reasonably approximated by a low-order Markov process. A HMM with Gaussian emissions is a generalization of the Gaussian mixture model, which is able to approximate arbitrary smooth probability densities (Juang et al., 1986). We therefore choose the normal distribution as the emission distribution of each state, so that the model learns to approximate the heavy-tailed distribution of the inter-event times via state transitions. Modeling inter-event times with an HMM represents the assumption that the interaction of collective interest and visibility produces discrete states of higher and lower activity, with each state producing interactions on a different time scale, and each state containing sufficient information to predict the next state. Our reasoning up to this point is very similar to that of Kleinberg (2002), who also uses an HMM-like model for the identification of bursts, but attempts to recover the hierarchical structure of activity.

The data is split into 10% development, 80% training, and 10% test set. The development set is used for determining the optimal number of hidden states according to the Bayesian Information Criterion (BIC) via 10-fold cross-validation (CV). Because of the high computational cost of HMM parameter fitting, the portion of the data that is held-out for hyper-parameter search is kept small. We observe that, for all datasets, the unpenalized CV likelihood continues to increase with a rising number of states, so the BIC penalty term is necessary for

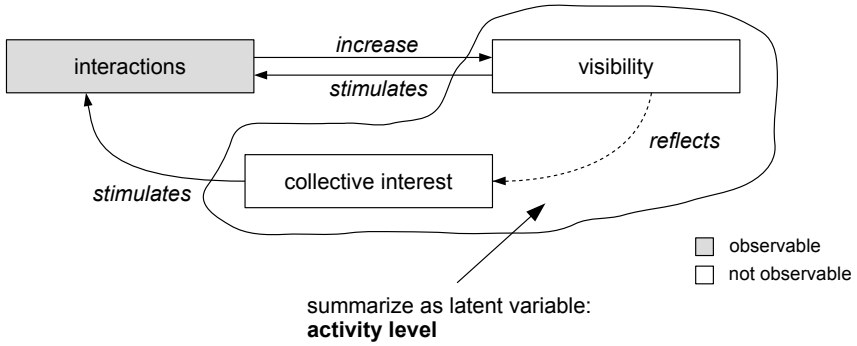


Figure 3.10: Modeling the interplay of user-content interaction and content visibility as a latent variable

the selection of a parsimonious model. By comparing the perplexity of training and test set, the ability of the model to generalize to unseen data is verified.

A first-order HMM with Gaussian emissions is fit to the sequence data via expectation maximization (EM) in log-space (Rabiner, 1989; Mann, 2006). When fitting an HMM to the training (development) set, the EM algorithm is stopped after 20 000 (100) iterations or when the per-iteration gain in log-likelihood falls below 10^{-6} (0.1). The EM algorithm is modified so that incomplete sequences do not cause updates to the initial probabilities. Due to the limited Markov horizon, no changes to the update rules of the transition probabilities are necessary. To avoid pathological output distributions (variance tending to zero), their initial parameters are obtained by k-means clustering of the inter-event times followed by fitting a normal distribution to each cluster.

Analysis of HMM Parameters Within our model of user-content interaction, it is impossible to be certain that an interaction sequence has concluded, in the sense that no further interactions will occur. A period of inactivity, regardless of its length, may possibly be followed by renewed activity. Therefore, it is not appropriate to simply fit an HMM with an absorbing final state to the data, and define the content lifetime as the expected time to reach that state. However, since each state of the HMM represents a particular level of activity, we can propose a hypothetical minimum level of activity, below which no further interactions happen, and compute the expected time until activity drops below that level. By computing the expected lifetime for different minimum levels of activity, we obtain the typical lifetimes of short- and long-lived content.

Given an HMM $H = (A, B, \pi)$ with a vector of initial probabilities π, a matrix of transition probabilities A and output distributions B, we first construct an isomorphic HMM $H' = (A', B', \pi')$ by permutation of H so that the states $1..N$ are ordered by ascending expected values of their output distributions. For each state $1 \leq i < N$, the expected lifetime $L(i)$ for a minimum level of activity represented by state i is computed as follows: We construct a sub-HMM H'_i, where all states $j > i$ have been turned into absorbing states by setting $A'_{j,k\neq j} = 0$, $A'_{j,j} = 1$, $\pi'_j = 0$ (see figure 3.11 for an example), and re-normalizing π'. H'_i is characterized by its fundamental matrix $N = (I - Q)^{-1}$, where Q is the upper-left $i \times i$ sub-matrix of the

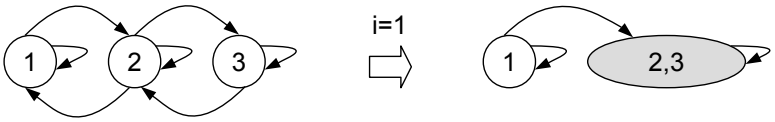


Figure 3.11: Example for merging multiple states of an HMM into a single absorbing state

transition probabilities A'. $N_{i,j}$ is the expected number of times visiting a non-absorbing state j when starting in state i, so the expected counts of visiting each non-absorbing state given the initial probabilities π' are $x = \pi'^T N$ and can be obtained by solving $(I-Q)^T \cdot x = \pi'$. By linearity of expectation, the expected lifetime $L(i)$ (equation 3.1) can be derived from the expected visit count and the expected value of the corresponding output distribution:

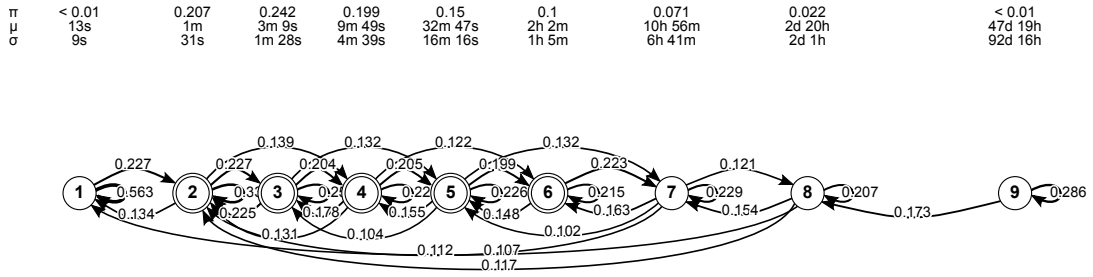
$$L(i) = \sum_{k=1}^{i} x_k \cdot E[B'_k] \tag{3.1}$$

Analogously, the expected number of interactions $C(i)$ can be defined as the component-wise sum of x.

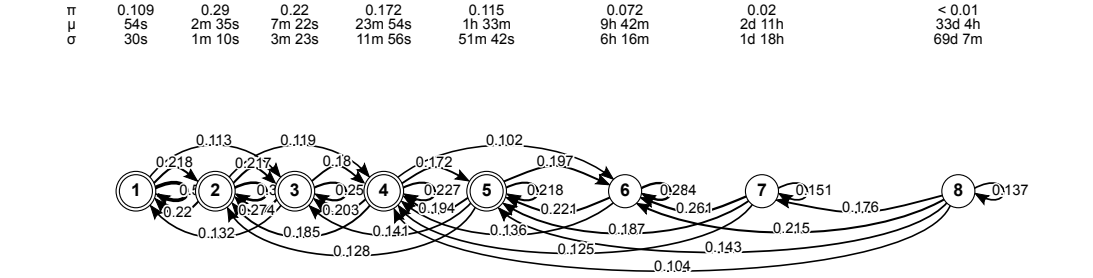
This procedure is sensitive to outliers that cause structural change to the HMM. If outliers end up being represented by a dedicated state, the estimated probabilities for entering that state will be very low, leading to obviously unrealistic lifetime predictions if the state is chosen as the only absorbing state. This is the case for two strong outliers in the HT dataset (a user forwarding two mails from nine years ago), which we elect to exclude from this study following manual investigation.

Results

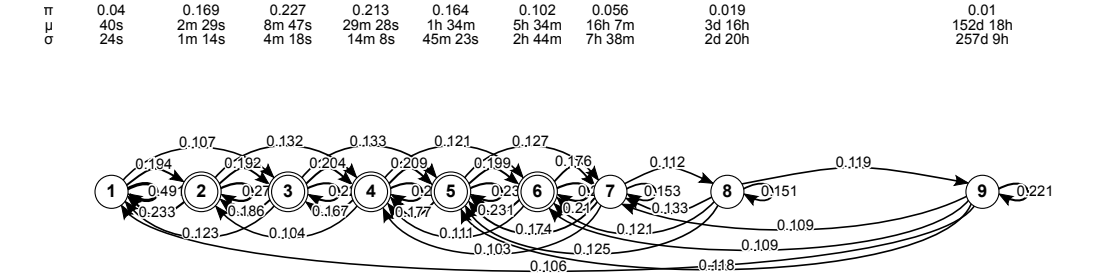
The fitted HMMs generalize well, as can be seen by comparing the perplexity of training and test data given a particular model. For all datasets except the HT e-mails, the test set perplexity is slightly below the perplexity of the training set. In the case of HT, the test set perplexity is 13% higher. By comparing the HMMs fit to the different datasets, one can identify a number of general patterns. Figure 3.12 shows the HMM fit to the Twitter data and visualizes the output distributions on a doubly logarithmic plot, while figures 3.13 and 3.14 compare the HMMs fit to interaction sequences from the different datasets. Across the HMMs, the number of states chosen by the CV procedure grows with the amount of training data. In each HMM, the means of the output distributions (indicated by the horizontal position of the states) are roughly equidistant on a logarithmic scale, approximating the heavy tail of the empirical distribution of inter-event times. The probability of self transition decreases with the level of activity, while the probability of transitioning to a different state is roughly proportional to the difference in activity levels. An exception is the comparatively high probability of transitions from states of low activity to states of high activity. These observations are consistent with our earlier hypothesis of how the temporal distribution of interactions is shaped by the presentation of content: waning collective interest results in a gradual decrease of activity, but a return to high activity may be triggered whenever a new interaction temporarily makes the content visible to others.



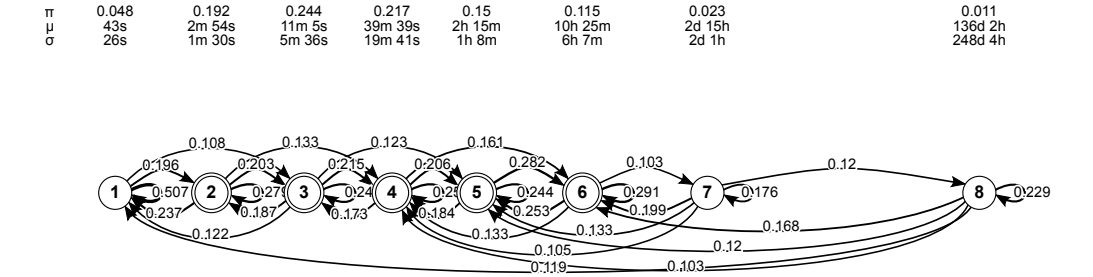
(a) Twitter (reply + share)



(b) Twitter (reply)

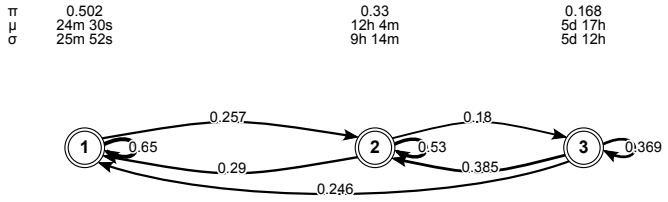


(c) Facebook (reply + share)

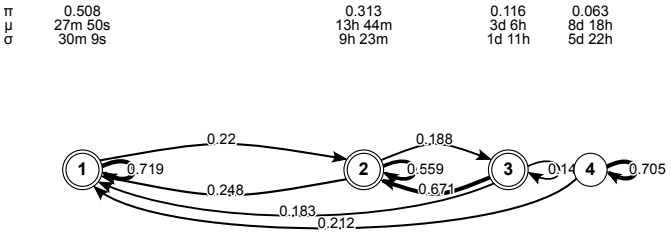


(d) Facebook (reply)

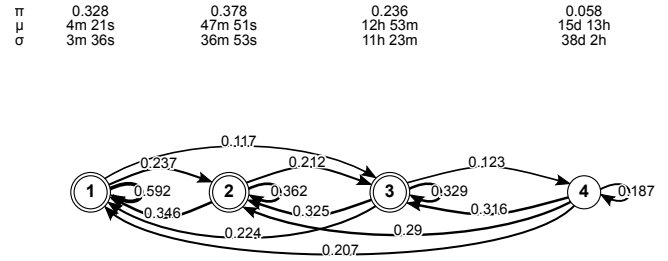
Figure 3.12: Parameters of HMMs fit to inter-event times of Twitter and Facebook interaction sequences. Double circles indicate $\pi > 0.1$. Transitions with a probability below 0.1 have been elided for visual clarity.



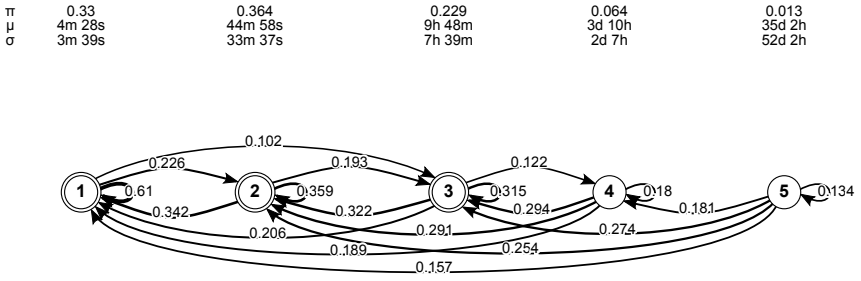
(a) Enron (reply + share)



(b) Enron (reply)



(c) HackingTeam (reply + share)



(d) HackingTeam (reply)

Figure 3.13: Parameters of HMMs fit to inter-event times of e-mail interaction sequences. Double circles indicate $\pi > 0.1$. Transitions with a probability below 0.1 have been elided for visual clarity.

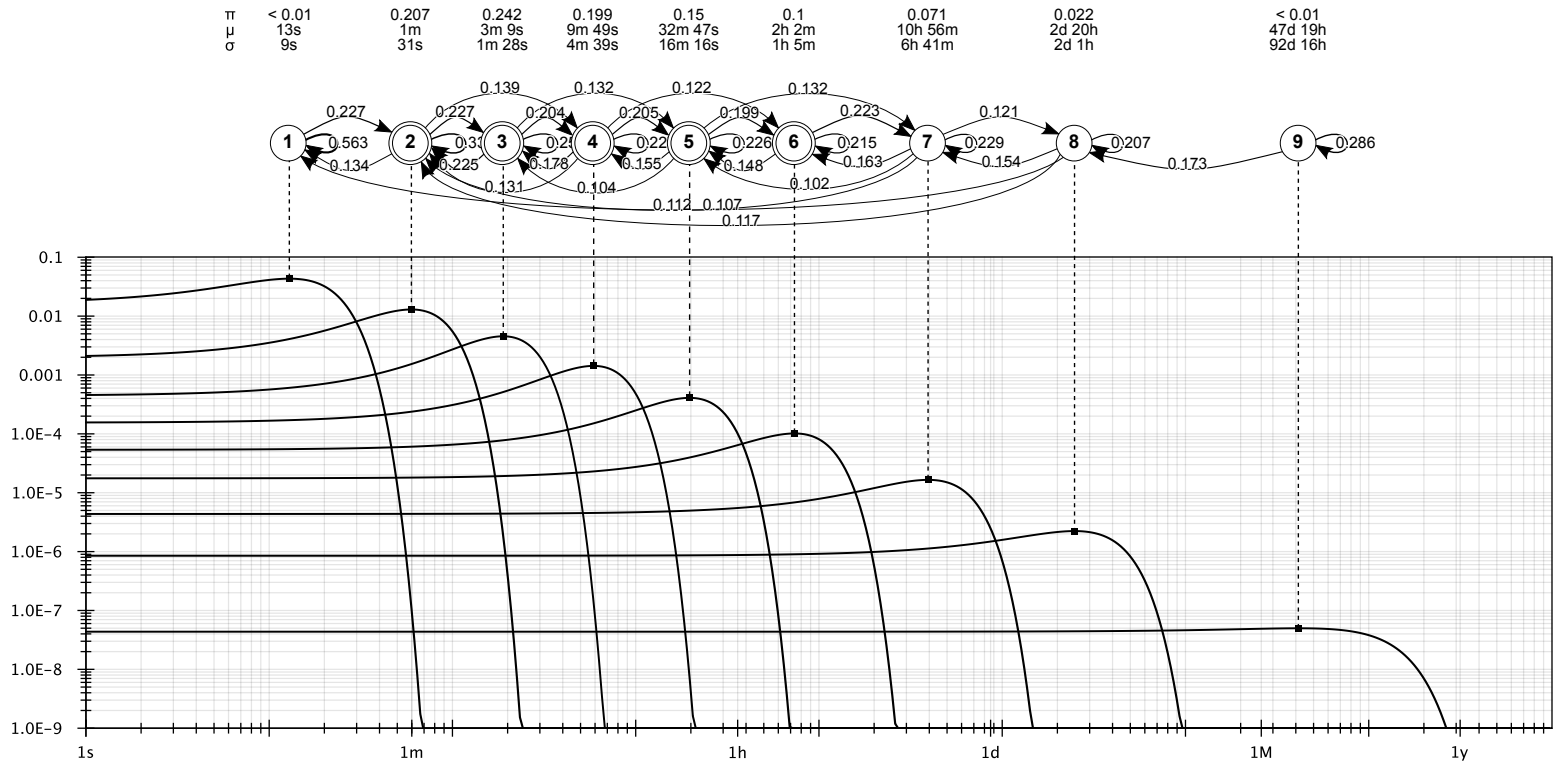


Figure 3.14: Transition and output probabilities of a HMM fit to inter-event times of Twitter interaction sequences (reply + share; Hauffa and Groh, 2019). Double circles indicate $\pi > 0.1$. Transitions with a probability below 0.1 have been elided for visual clarity.

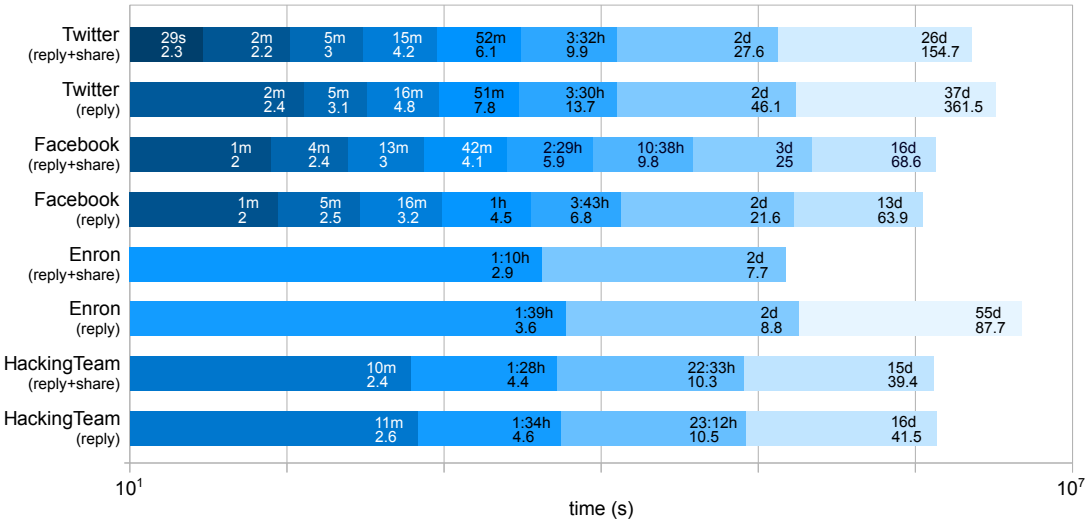


Figure 3.15: Expected lifetimes $L(i)$ and expected interaction counts $C(i)$ per level of activity i for each dataset (Hauffa and Groh, 2019). Lightness is proportional to lifetime.

Figure 3.15 shows the expected lifetimes $L(i)$ on each social platform. One can observe four main time scales, one of short term activity for at most 15 minutes, two classes of medium term activity for 1–1:30 hours and 1–3 days, respectively, and finally a class of long term activity taking up two weeks or more. While medium term activity is fairly consistent across the social platforms, the characteristics of both short and long term activity are more variable and dependent on the medium. In order to evaluate the effect of content sharing on lifetime, we compare the expected lifetimes obtained from the complete dataset of a particular platform ("reply+share") to the expected lifetimes computed from a restricted dataset from which all sharing interactions have been removed ("reply"). Sharing generally shifts the model parameters towards shorter inter-event times. On Twitter, the presence of an additional short term state points to a class of content that is rapidly shared as long as it is highly visible, while on Facebook, an additional medium term state indicates that sharing may also prolong the lifetime of particular types of content.

On each of these time scales, Twitter outpaces Facebook in terms of expected interactions $C(i)$ per unit of time, while Facebook is comparable to modern corporate e-mail communication as represented by the HT dataset. Enron is noticeably slower, which matches earlier observations about the differences in the two e-mail datasets.

We have proposed a characterization of content lifetime that arises naturally from the interaction of collective interest in the content and its visibility on a social media platform. This characterization is not only relevant for learning about human behavior and the temporal characteristics of social media. When performing an experiment or observational study on social media data that involves temporal quantization, i.e., the partitioning of an observation period into intervals of equal length, our method can inform the choice of interval size. Too fine or too coarse quantization may render temporal phenomena invisible (Krings et al., 2012). The time scales identified by our model correspond to interval sizes at which

Table 3.3: User activity within the observation period on different social platforms

	observation period	nodes (non-addr.)	edges (addr.)
Twitter	Apr. & May 2012	81.7%	41.0%
Facebook	Apr. & May 2014	43.5%	8.0%
e-mail (Enron)	Aug. & Sep. 2001	52.4%	37.4%
e-mail (HT-en)	Apr. & May 2015	45.7%	40.1%
e-mail (HT-it)	Apr. & May 2015	62.0%	49.3%

distinct effects can potentially be observed.

3.5.3 Choosing an Observation Period

Although we are trying to treat all datasets equally, some concessions have to be made to the origin and nature of the collected data when using it in an experiment. The experiments on social influence conducted in this thesis are fundamentally predictive in nature: observations from one interval of time are used to make predictions about future intervals. It follows that an important data-specific decision is the choice of an appropriate observation period, defined by its start date and its length. The observation period must be long enough to allow for a subdivision into a sufficient number of intervals. However, the period should not be too long, in order to avoid the adverse effects of conceptual drift: long-term changes in trends, conversation topics, and social behavior in general, that let the predictive value of data deteriorate over time.

In the previous section, the analysis of temporal rhythms in social media data identifies multiple classes of discussion lifetime that are to some degree consistent across different media. The longest consistently observed lifetime was around 14 days, so we use this value as an upper bound for the interval size in our experiments. Using an observation period of 56 days (approximately two months) yields time series with a minimum length of four. To maximize the amount of available data, the observation period should coincide with the period of highest recorded activity (see section 3.5.1). While most of the datasets have their activity peak in different years, data from the same months is used where possible to avoid confounding seasonal effects. Due to the comparatively small size of the Enron dataset, its observation period is chosen by activity alone.

Table 3.3 shows the chosen two-month observation period for each dataset and the activity within that period. For the sake of interpretability, we specify the observation periods in terms of calendar months, while the actual period is the 56-day interval that ends with the last day of the second month (inclusive). An active node is a user who has sent at least one non-addressive message, while an active edge is a pair of users (a, b) where a has sent at least one addressive message to b. Table 3.3 gives the ratio of nodes and edges that are active within the observation period to nodes and edges that are active within the whole time frame spanned by the collected data.

Despite choosing periods of high activity, restricting the datasets to observations from a two-month period has a strong sparsifying effect on the implicit social networks induced by

communication. We refer to this spatio-temporal sparsity of interactions in a social network as *temporal sparsity*. The inequality of the distribution of message volume across individual users provides a different point of view on this phenomenon. In the Twitter dataset, the 10% of users with the highest individual message volume contribute 43% to the overall amount of messages. Conversely, 90% of messages are authored by the 54% most active users. The Gini coefficient is 0.58; a uniform distribution would be characterized by a Gini coefficient of zero, while the coefficient approaches one with increasing inequality. The other datasets exhibit an imbalance of similar magnitude. All three e-mail datasets have a Gini coefficient between 0.66 and 0.67. With a coefficient of 0.73, Facebook is the most imbalanced dataset. The top 10% of users are responsible for 58% of messages, and 90% of the message volume can be attributed to the top 35% of users. In all of these cases, a large amount of messages is generated by a comparatively small group of highly active users.

Within the scope of this thesis, we can only hypothesize about the mechanisms that cause temporal sparsity. In section 3.5.2, we establish that the temporal distribution of user behavior exhibits burstiness, i.e., periods of frequent activity interrupted by longer periods of inactivity. The activity statistics indicate that both addressive and non-addressive communication are highly bursty: If up to 89% of otherwise active users and user pairs are inactive during the observation period, a likely explanation is that the communication behavior of many users is characterized by long gaps between periods of activity. Another contributing factor is that nodes and edges can become permanently inactive, for example, if users quit the platform or user pairs stop interacting.

Temporal sparsity points towards inefficiency in our data acquisition process: Traversing the explicit network graph in BFS order enables a comparison between the explicit and implicit network, but the visited nodes are not guaranteed to be active in the observation period. Traversing only the implicit network ensures that all nodes, and at least one incident edge per node, are active, and would have therefore constituted a more efficient use of our limited crawling resources.

3.6 Structural Characteristics

Social network analysis distinguishes two kinds of network: Explicit networks are made up from relational ties that have been explicitly reported by — or elicited from — the actors. Implicit networks are induced by the actors' behavior towards each other and can be reconstructed from observations of said behavior. Among the datasets presented in this chapter, only the Facebook and Twitter datasets include an explicit social network graph. Facebook provides an undirected graph of "friendship" between users, while Twitter provides a directed graph of users subscribing to others' content by "following". There are many conceivable ways of recovering an implicit network graph from observations (cf. section 2.1.1 and Tsur and Lazer, 2017). Here, and in the following experiments, the implicit network graph is constructed by inserting a directed edge for each ordered pair (a, b), where a is actively communicating with b in the observation period. Isolated nodes, i.e., actors that are not actively communicating with at least one other actor, are removed. On both platforms, users are able to addressively communicate with others even in the absence of an explicit

social tie, so the implicit graph is not necessarily a subgraph of the explicit graph.

The structure of these social network graphs can help answer a number of questions about the dataset: Is the crawling process biased towards a particular group of users? Does the crawled subgraph meet the requirement, formulated at the beginning of this chapter, of being sufficiently dense? Are the implicit and explicit graphs of the various social platforms similar in structure? Do the graphs exhibit structural characteristics that are commonly associated with "real-world" social networks, i.e., social networks formed by human actors outside of online spaces? In particular, can they be decomposed into meaningful (to be defined) overlapping communities?

3.6.1 Graph Metrics

A network graph can be characterized by various summary statistics, commonly known as *graph metrics*, which capture particular aspects of the graph structure. For a directed graph G with $|V|$ nodes and $|E|$ edges, the metrics of interest are defined as follows:

Mean and variance of the node degree The degree of a node is the number of its neighbors in the graph. In a directed graph, one distinguishes the number of nodes connected via incoming edges (in-degree) and outgoing edges (out-degree). In- and out-degree have the same mean, but may differ in variance. If all nodes are isolated, the mean degree $\bar{D}$ is zero, and if the graph is fully connected (complete), $\bar{D}$ attains its maximum of $|V| - 1$. In a uniformly connected graph, the variance of the node degree σ_D^2 is zero, while a high variance indicates the coexistence of highly connected nodes and nodes with few or no neighbors.

Such heavy-tailed degree distributions have been observed in a diverse range of large graphs, including collaboration networks and the network of interlinked websites. If the distribution can be modeled by a power law $P(D = k) \sim k^{-\gamma}$, the graph is called *scale-free* (Barabási and Albert, 1999). For typical exponents $2 \leq \gamma < 3$, in the limit $|V| \to \infty$, the variance of the power law distribution diverges, so that the degree of a randomly chosen node may exhibit an arbitrarily large deviation from the mean (Barabási, 2016c). Conversely, high empirical variance of the node degree is indicative of scale-freeness. In graphs that grow over time, a possible explanation for the emergence of scale-freeness is *preferential attachment*, which is a tendency of new nodes to associate with existing nodes of high degree (Barabási and Albert, 1999).

Beyond graph theoretical considerations, the number of simultaneous social relationships that humans are able to actively maintain is believed to be limited. This capacity, known as Dunbar's number and estimated to be in a range of 100–300 (Gonçalves et al., 2011), effectively places an upper bound on the node degree in social networks derived from human interaction.

Density and normalized degree variance Since the node degree is bounded above by the size of the graph, we report not only the unnormalized mean and standard deviation of the node degree, but also normalize mean and variance to the interval $[0, 1]$ to enable the comparison of different graphs. Normalizing the mean degree yields the

density of G, which is more conventionally defined as the ratio of its edge count $|E|$ to the edge count of a complete graph with the same number of nodes: $d = \frac{|E|}{|V|(|V|-1)}$ Naturally occurring graphs are typically sparse in the sense that $d \ll 1$ (Barabási, 2016b, section 2.5). Smith and Escudero (2020) define the normalized variance as $\frac{\sigma_D^2(|V|-1)}{|V||E|(1-d)}$.

Connected components G is called *weakly connected* if any two nodes a and b are connected by a path in the undirected graph $G' = (V, E')$ that results from replacing all directed with undirected edges: E' contains all pairs of nodes that are connected by at least one directed edge in G. A *weakly connected component* (WCC) is a maximal connected subgraph, that is, a subgraph that cannot be grown by adding another node from the original graph without becoming disconnected.

Of particular interest is the number of components and the relative size of the largest component. One of the most simple random graph models is the Erdős-Rényi (ER) model, where an undirected edge is inserted between two nodes with a constant probability p. A graph generated by the ER model almost surely develops a giant component as $|V| \to \infty$ if $p\,|V|$ is held constant so that $p > 1/|V|$. If a giant component is present, no other component contains more than $O(\log(|V|))$ nodes. A similar result can be obtained for directed graphs generated by the configuration model, a generalization of ER (Newman, 2003). In general, deviations from the metrics predicted by the ER model indicate that the graph has a more complex formation process than random chance.

Clustering coefficient The clustering coefficient C measures the tendency of a graph to contain subgraphs of high density. Multiple definitions exist. The one used here is the ratio of the number of closed triplets to the total number of triplets. A triplet is formed by a node a that is connected to each of $\{b, c\}$, with $a \neq b \neq c$. The triplet is closed if b and c are also connected. The number of closed triplets in a graph is three times the number of its triangles (fully connected subgraphs of size three). This definition is valid for undirected graphs. For the purpose of computing the clustering coefficient of a directed graph G, an undirected graph G' is constructed, as defined previously, by replacing directed with undirected edges (Opsahl and Panzarasa, 2009).

The clustering coefficient can be interpreted as the conditional probability of two nodes being connected if they have a common neighbor. In an ER random graph, the probability p of the presence of an edge is constant and therefore independent of all other edges, so $C = p$. Social networks have been found to have a substantially higher clustering coefficient than predicted by simple random graph models (Newman and Park, 2003). Opsahl and Panzarasa (2009) review the literature and list a number of social mechanisms that may be responsible for this phenomenon, but note a lack of consensus.

Assortativity Generally speaking, assortativity is a manifestation of homophily: the tendency of nodes to be connected to other nodes that are similar with respect to some attribute. When the scope of analysis is limited to the graph structure, the node degree is a natural choice for this attribute. In this case, Pearson's correlation coefficient of

Table 3.4: Metrics of Twitter's explicit social network (follower graph)

	working set	entire crawl	complete network (largest WCC; Kwak et al., 2010)
vertices	30 000	358 342	41 652 230
edges	3 825 022	151 463 754	1 468 365 182
type	directed	directed	directed
mean degree	127.5	422.7	35.3
in-degree std.dev.	313.2	1 228.1	354.7
out-degree std.dev.	163.6	976.2	2 419.7
density	4.25×10^{-3}	1.18×10^{-3}	8.46×10^{-7}
norm. in-degree variance	2.58×10^{-2}	9.97×10^{-3}	8.57×10^{-5}
norm. out-degree variance	7.03×10^{-3}	6.30×10^{-3}	3.99×10^{-3}
WCC count	11	36	
largest WCC	99.97%	99.99%	
clustering coefficient	1.05×10^{-1}	7.96×10^{-2}	8.46×10^{-4}
assortativity	-0.127	-0.113	-0.051

the degrees of all pairs of nodes connected by an edge is a measure of assortativity (Newman, 2003) with a range of $[-1, 1]$. By convention, graphs with an assortativity of zero are neutral, graphs with an assortativity above zero are called assortative, and graphs with an assortativity below zero disassortative. In an assortative graph, nodes tend to be connected to other nodes of similar degree, while in a disassortative graph, connected nodes tend to differ in degree. An ER random graph is neutral (in the asymptotical case of $|V| \to \infty$; Newman, 2002a), since each edge exists with a probability that is independent of all other edges. Newman (2003) observes that "essentially all social networks measured appear to be assortative, but other types of networks (information networks, technological networks, biological networks) appear to be disassortative."

With the exception of the explicit network of Facebook friendships, the networks analyzed here are directed graphs. The normalization factor for the degree variance is defined in the same way for directed and undirected graphs. For the other metrics listed above, a corresponding definition for undirected graphs can be formulated by constructing — implicitly or explicitly — an equivalent directed graph by replacing each undirected edge with two opposing directed edges.

Table 3.4 compares relevant metrics of our differently sized samples to the complete Twitter follower graph of July 2009. The follower graph was obtained by Kwak et al. (2010), who claim to have "crawled the entire Twitter site" by distributed BFS-style crawling and made the resulting graph publicly available. This graph does not contain isolated nodes and consists of a single weakly connected component. Cha et al. (2010) take a different approach:

after exhaustively checking all user IDs in the range of zero to 80 million in August 2009, they encountered 54 981 152 valid accounts and 1 963 263 821 following ties. They report that in addition to a large connected component containing 94.8% of users, 0.2% of users belong to smaller components, while 5% are isolated.

If an explicit network graph is sampled by BFS-style crawling, the resulting graph is weakly connected by construction. In practice, however, the crawling process described earlier may produce a disconnected graph, because of crawling errors (friends or followers of a user could not be retrieved) and, specifically in the case of Twitter, users discovered via @-mentions that are not connected to the follower graph. As expected, each of the sample graphs has a single giant WCC that contains the majority of nodes. All of the remaining components are isolated nodes. Even though @-mentions are systematically explored as part of the crawling process, the follower graph and the graph of @-mentions appear to be mostly overlapping.

The metrics confirm that BFS exhibits a clear bias towards high-degree nodes, but still improves data quality for our use case by yielding subgraphs that are more dense than the original graph by orders of magnitude, which is also evidenced by the clustering coefficient. The sample graphs strongly overestimate the in-degree variance, which indicates, in conjunction with the high mean degree, that users with many followers are over-represented. With increasing sample size, graph metrics that do not depend on the graph size (density, normalized variance, clustering coefficient, assortativity) appear to converge towards the metrics of the complete network. The graph is weakly disassortative, so highly connected nodes tend to avoid each other. This paints the picture of a medium where celebrities connect with their fans, but not with each other, which is also reflected by the degree variability. More generally, one can distinguish highly connected hubs and regular users.

Table 3.5 compares our sample of the Facebook friendship graph to the complete friendship graph of May 2011. Ugander et al. (2011) had direct access to Facebook data, analyze a snapshot of the friendship graph, and report a number of graph metrics. In a later study, Backstrom et al. (2012) analyze the same snapshot and provide additional metrics. Both studies only consider active users who have logged into Facebook at least once in the 28 day period preceding the crawl and have at least one friend, so their graphs do not contain any isolated nodes. Neither group has publicly released the graph data.

As in the case of Twitter, BFS-style crawling produces a sample that is denser than the original graph by orders of magnitude. However, the sample underestimates the mean degree. One possible explanation is that users can opt to hide their friend list, which makes it inaccessible to the crawler. Another possible cause is that Facebook does not have hub users with a disproportionally high degree: Facebook limits its users to a maximum of 5 000 friends and suggests that celebrities, brands, and other entities that intend to broadcast to a large audience create “fan pages” that exist outside of the friendship graph (Facebook Help Team, 2015). Still, the complete Facebook graph has a much higher mean degree than the Twitter graph. The average user appears to be more discriminating when choosing whom to follow than when deciding who to accept as a friend. Finally, the sample preserves the assortative tendency of the original graph. Well-connected users are likely to be connected, which is more typical for human social networks than the disassortative tendency of Twitter.

An issue with comparing the crawled sample graphs to substantially larger reference

Table 3.5: Metrics of Facebook's explicit social network (friendship graph)

	crawl	complete network (Ugander et al., 2011; Backstrom et al., 2012)
vertices	16 834	~ 721 100 000
edges	608 054	~ 68 700 000 000
type	undirected	undirected
mean degree	72.2	190.4
degree std.dev.	91.4	"high"
density	4.29×10^{-3}	2.64×10^{-7}
norm. degree variance	1.38×10^{-2}	-[1]
WCC count	1	> 100 000
largest WCC	100.00%	99.91%
clustering coefficient	2.95×10^{-1}	-[2]
assortativity	0.189	0.226

[1] Cannot be computed without knowing the exact degree variance.
[2] Cannot be computed without access to the network graph. Ugander et al. (2011) use a definition of the global clustering coefficient that differs from ours.

graphs is that the two available reference graphs predate our crawling efforts by years. The Twitter graph of Kwak et al. was captured during a period of fast growth. Gabielkov et al. (2014) report that "the total number of accounts went from 4.265 million in January 2009 to 67.487 million in January 2010". Gabielkov et al. (2014) replicated the crawling procedure of Cha et al. (2010) three years later, and find a noticeable shift in the macrostructure of the graph: a higher number of hub users with many followers, but few friends, and also more passive consumers of content, who mainly follow hubs and have no followers of their own. Similarly, Backstrom et al. (2012) observe that certain characteristics of the Facebook friendship graph have changed over time, for example: "During the fastest growing years of Facebook [...] density was going down steadily." In consequence, it is unclear how meaningful a comparison with older reference graphs can be. Under the provision that changes in the graph metrics over time are mainly driven by the growth of the graph, and are therefore correlated with its size, older snapshots of the graphs can still be useful points of reference.

An alternative way of evaluating the quality of the crawled sample that does not require any information about the original graph (beyond what is known at the time of crawling) can be obtained by asking how much of a node's direct neighborhood is present in the sample. We define a node's *local completeness* as the ratio of the number of neighbors that are present in the sample to the number of neighbors that are known to the crawler, which includes nodes that had not been visited yet when the crawling process was terminated. To account for the fact that in a graph obtained by BFS-style crawling, each node has at least

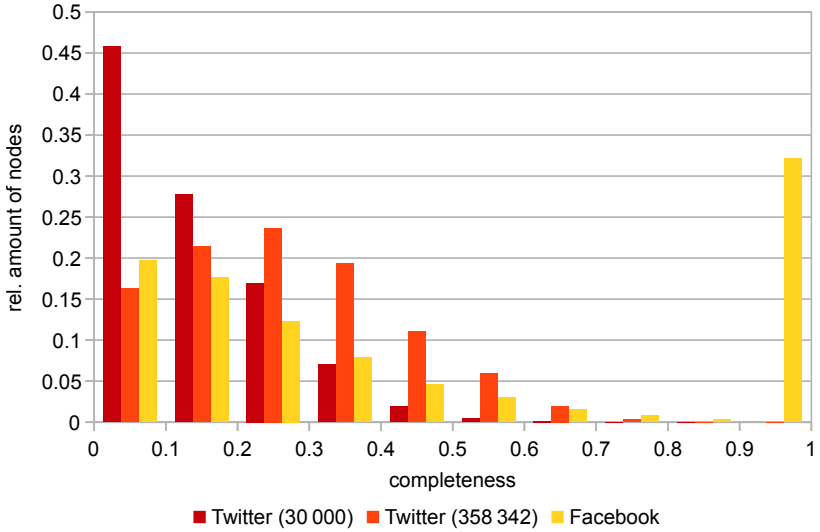


Figure 3.16: Distribution of local completeness in samples of explicit social network graphs

one neighbor, a value of $d_{\min} = 1$ for directed, 2 for undirected graphs is subtracted from numerator and denominator. Equation 3.2 defines the local completeness $c(v)$ of node v in graph G, where G is the graph that contains all nodes and edges that are known to the crawler, G' is the subgraph induced by the actually visited nodes, and $d_{G,\text{in}}(v)$ and $d_{G,\text{out}}(v)$ are the in- and out-degree of node v in graph G, respectively. The local completeness is undefined if $d_{G,\text{in}}(v) + d_{G,\text{out}}(v) \leq d_{\min}$, i.e., if a node has no known edges except the one through which it was discovered.

$$c(v) = \frac{d_{G',\text{in}}(v) + d_{G',\text{out}}(v) - d_{\min}}{d_{G,\text{in}}(v) + d_{G,\text{out}}(v) - d_{\min}} \tag{3.2}$$

Upcoming experiments (chapter 5 onwards) operate under the assumption that information about a node can be recovered from its neighborhood in the graph, so the degree of preservation of the neighborhood in the sample is an important measure of quality. In the full Twitter sample graph, the average local completeness is 26.4%; in the 30 000 user subgraph, it is 13.8%. In the Facebook sample graph, the average local completeness is 47.2%. Figure 3.16 compares the distribution of local completeness in the sample graphs. Although the average completeness in the two Twitter samples rises with increasing sample size, nodes of high completeness are rare in both. The distribution of completeness in the Facebook sample exhibits a large number of nodes with maximal completeness. This is an artifact of Facebook's privacy settings: any user with a hidden list of friends, who is connected to at least two already crawled users via an undirected edge, has maximal completeness by definition. Without these users, the distribution is similar to that of Twitter. In a setting where a high local completeness takes priority over preserving the global graph structure, the efficiency of sampling could be improved by moving from blind to informed BFS: insert each newly discovered node into a priority queue; when encountering an edge to an already

Table 3.6: Metrics of the implicit social networks of Twitter and Facebook (communication graphs)

	Twitter	Facebook	Facebook (largest WCC)
vertices	15 614	5 235	3 333
edges	109 743	6 928	5 371
type	directed	directed	directed
mean degree	7.0	1.3	1.6
in-degree std.dev.	19.3	2.5	3.0
out-degree std.dev.	11.6	1.6	1.8
density	4.50×10^{-4}	2.53×10^{-4}	4.84×10^{-4}
norm. in-degree variance	3.41×10^{-3}	8.86×10^{-4}	1.63×10^{-3}
norm. out-degree variance	1.23×10^{-3}	3.56×10^{-4}	6.03×10^{-4}
WCC count	175	594	
largest WCC	97.46%	63.67%	
clustering coefficient	5.52×10^{-2}	7.02×10^{-2}	6.68×10^{-2}
assortativity	0.035	0.073	0.003

discovered node, increment its priority.

Tables 3.6 and 3.7 show the metrics of the implicit network graphs formed by communication in the observation period. By definition, the graph does not contain any isolated nodes, or, equivalently, any components of size one. With the exception of the Facebook communication graph, all communication graphs have a giant WCC that contains more than 97% of nodes. In the case of the Facebook communication graph, the largest WCC covers less than two thirds of the graph and has noticeably different metrics: higher density, higher degree variance, and an assortativity that is much closer to neutral. In terms of these metrics, the largest WCC is similar to the Twitter communication graph.

The Facebook communication graph decomposes into a high number of WCC of non-negligible size. A possible explanation is that compared to other social media, communication on Facebook is more equally distributed among the different modes of communication: While on Twitter, tweeting is the dominant mode of communication, Facebook is undergoing a shift, where posting messages on each others' timelines is falling out of favor, and private direct messages are becoming more popular (Lorenz, 2017). If a mode of communication is, like direct messaging, principally invisible to the crawler, certain pairs of users may erroneously end up being disconnected in the communication graph.

The three e-mail communication graphs are similar to each other, with the Enron graph being somewhat less dense and less variable in node degree than the HackingTeam graphs. Compared to the social media communication graphs, the e-mail graphs are smaller, more dense, and exhibit a higher normalized degree variance. The e-mail graphs are close to neutral in assortativity, with the exception of the Italian language HackingTeam graph. Since

Table 3.7: Metrics of the implicit e-mail social networks (communication graphs)

	Enron	HackingTeam (en)	HackingTeam (it)
vertices	126	47	47
edges	450	737	924
type	directed	directed	directed
mean degree	3.6	15.7	19.7
in-degree std.dev.	2.9	9.1	10.2
out-degree std.dev.	3.3	9.2	11.0
density	2.86×10^{-2}	3.41×10^{-1}	4.27×10^{-1}
norm. in-degree variance	1.85×10^{-2}	1.69×10^{-1}	1.92×10^{-1}
norm. out-degree variance	2.42×10^{-2}	1.70×10^{-1}	2.22×10^{-1}
WCC count	2	1	1
largest WCC	98.41%	100.00%	100.00%
clustering coefficient	3.58×10^{-1}	6.49×10^{-1}	6.86×10^{-1}
assortativity	0.027	-0.030	-0.265

Italian is the native language of the actors, it stands to reason that this graph more accurately reflects the characteristics of internal communication. One possible explanation is that in a company with a strong hierarchy, the bulk of communication may happen between senior employees and their subordinates rather than among peers. A deeper analysis of the communication patterns would be required to conclusively explain this observation.

Comparing the communication graphs to the follower graphs, one mainly notices a lower density and degree variance, which is an immediate manifestation of the temporal sparsity discussed earlier (see section 3.5.3). The lower absolute mean node degree can be explained by Dunbar's number, considering that communicative ties require active maintenance, while accumulating friendship and following ties does not directly increase the cognitive load. These observations are generally consistent with the literature: According to a review of Olteanu et al. (2019), "Wilson et al. showed that the network constructed based on explicit links among users was significantly denser than the one based on user interactions, while Viswanath et al. showed that in the interaction-defined network, nodes tends to have a smaller, bounded degree." (citing Wilson et al., 2009; Viswanath et al., 2009)

Finally, we directly compare the sets of edges of the explicit and implicit network graphs. The nodes of each graph are subsets of a common set of actors, so an edge in one graph is considered equal to an edge in the other if their endpoints are equal. Since the communication graphs are generally substantially smaller, a symmetric measure of similarity like the Jaccard coefficient would not be informative. Instead, we compute the overlap coefficient, which, in this case, can be interpreted as the percentage of edges in the communication graph that are also present in the explicit network graph. The Twitter communication graph has 81.6% overlap with the follower graph, while the Facebook communication graph has 87.9%

overlap with the friendship graph (treating each undirected edge as two directed edges). Both platforms have a small amount of users who communicate with others, despite not being explicitly connected.

3.6.2 Community Structure

An explorative analysis of graph data typically starts with a visualization of the graph structure, designed to make salient structural properties of the graph stand out visually. Social groups are central to higher-level social organization (see section 2.1.3), so any sufficiently large social network graph is expected to have a *community structure*, where communities are the graph-structural manifestations of social groups. A community is usually defined as a subset of nodes that is densely connected internally, but sparsely connected to outside nodes (Newman and Girvan, 2004). Accordingly, a visualization algorithm that is suitable for social network graphs should minimize the spatial distance between nodes in the same community, while keeping nodes that belong to separate communities at a distance.

Huang (2014) and Schneider (2014) independently attempt to visualize the 30 000 user sample of the Twitter follower network and conclude that this is a difficult task, not only due to the size of the graph: Generic visualization algorithms fail to discover and amplify any underlying structure, so that even a smaller subgraph just appears as an amorphous blob of nodes (figure 3.17). The same phenomenon has been earlier studied by Ahn et al. (2010), who introduce the term "pervasive overlap". Observation of social structures clearly shows that people can be — and usually are — members of multiple social groups at the same time. Ahn et al. hypothesize that the resulting structure of highly overlapping communities does not lend itself well to visualization. To gain a better understanding of the problem, we apply community detection algorithms that can identify overlapping communities to the various social network graphs and compare the results.

Among the various methods of overlapping community detection (Coscia, 2019), two are of particular interest, because they each address the problem of pervasive overlap in a different way. These two methods are clique percolation (Palla et al., 2005) and edge clustering (Ahn et al., 2010). Both build upon what Barabási (2016a) calls the "structure hypothesis": the assumption that the graph structure contains sufficient information for recovery of the true community membership of each node.

Non-overlapping communities are usually characterized by a high ratio of inter- to intra-community density. The most simple structure that (trivially) has this property is the maximal clique: First, its members are fully connected, which is equivalent to maximal inter-community density. Second, no node outside of the clique can be connected to all nodes inside the clique, so the intra-community density must be lower. Reid et al. (2012) conclude that "a clique is a good conservative lower bound estimate of community structure, in so far as an observed clique more than likely is wholly contained inside some real-world community". Communities are always connected, so a similar argument can be made about the community size being bounded above by the size of the largest WCC. Among the social network graphs analyzed here, this upper bound is only relevant for the Facebook communication graph, which decomposes into multiple large components, whereas all other graphs have a single "giant component".

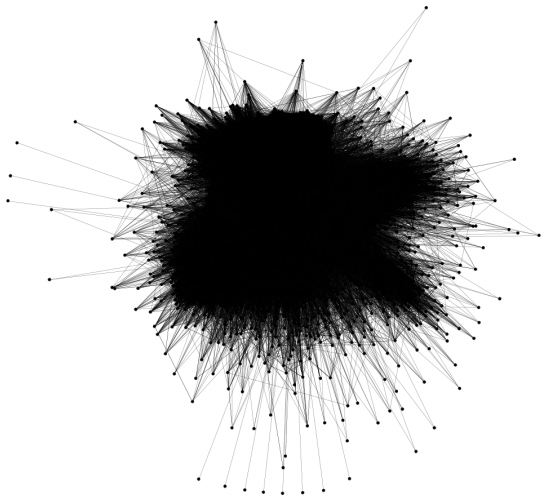


Figure 3.17: An attempt to visualize a 2 500 user subset of the Twitter follower graph with a force-directed layout algorithm

Maximal cliques impose stricter requirements on local density than one would expect from real-world social groups, where, with increasing group size, it is more and more unlikely that each member is in some kind of social relationship with each other. Even though one would intuitively expect maximal cliques to be rare, Reid et al. (2012) observe that network graphs of online SNS tend to contain a large number of highly overlapping cliques, which is inconsistent with our informal understanding of social groups as distinct entities. Relaxing the requirement of full connectedness leads to a family of substructures (Pattillo et al., 2013) like s-plexes, where each node is allowed to be lacking $s - 1$ (or fewer) edges to other members (a clique is a 1-plex), but such relaxed substructures are expected to occur even more frequently in a graph than cliques.

The clique percolation method (Palla et al., 2005) works by merging overlapping cliques, which permits communities of less than maximal density while also reducing the overall number of communities. A community is defined as the union of cliques of size k (or, equivalently, maximal cliques), which are chosen so that each clique shares $k - 1$ nodes with at least one other clique. The higher the value of k, the more dense are the resulting communities. A suitable value for parameter k has to be determined manually. In an ER random graph, there is a threshold for edge probability p that depends on k, above which clique percolation will almost surely produce a giant community (Derényi et al., 2005). In other words, for graphs of higher density and lower values of k, a large community is likely to emerge that is not supported by ground truth and may subsume a number of legitimate communities. Nodes that are not member of at least one clique cannot be assigned to a community, so the higher k, the more nodes are potentially excluded (Lee, 2013). It follows that there is an optimal value of k, but Lee observes that for some real-world graphs, no choice of k yields satisfactory results. As a heuristic, Palla et al. (2005) propose to "simply select the smallest

value of k for which no giant community appears". While clique percolation can be trivially adapted to work with other types of dense subgraphs, e.g., s-plexes, Palla et al. note that "in most cases [this] is practically equivalent to lowering the value of k".

Clique percolation is computationally expensive. The enumeration of all maximal cliques takes exponential time in the number of nodes (in the worst case), since a graph can have up to $O(3^{\frac{|V|}{3}})$ maximal cliques (Moon and Moser, 1965). In practice, the maximal cliques of a moderately sized graph like the 30 000 user subset of the Twitter crawl can be enumerated within a reasonable time frame using the Bron-Kerbosch (BK) algorithm (Bron and Kerbosch, 1973). The BK algorithm is formulated as a recursive procedure and can therefore be parallelized in a straightforward way by deferring each recursive call to a task that is appended to a shared work queue. Computing the overlap between the cliques is less computationally complex, but the actual bottleneck is memory: The algorithm of Palla et al. (2005) keeps the entire clique-clique overlap matrix in memory, which requires quadratic space in the number of cliques. Reid et al. (2012) propose a more space-efficient algorithm (we use "algorithm 1"), but show that any algorithm for clique percolation requires random access to all maximal cliques of the graph. They conclude that "current algorithms still scale worse than some other overlapping community finding algorithms" (Reid et al., 2012).

While the communication graphs are sufficiently small and sparse, and pose no problem with regards to memory consumption, the cliques of the explicit network graphs of Twitter and Facebook exceed the computational resources available to us. Clique percolation operates on undirected graphs, so directed graphs have to be transformed. The amount of cliques can be reduced by choosing a transformation that promotes sparsity, for example, inserting an undirected edge between two nodes if and only if they are connected by reciprocal directed edges. This heuristic ensures mutual awareness between adjacent nodes, and thus reduces the number of cliques in the Twitter follower graph to a manageable level, but cannot be applied to the Facebook graph, where edges already are undirected. After transformation, the 30 000 user subset of the Twitter graph contains 386 million maximal cliques of size three and above (maximum clique size 36), while the undirected Facebook graph contains 108 billion maximal cliques (maximum clique size 59). If memory consumption was not an issue, a variant of clique percolation that is applicable to directed graphs (Palla et al., 2007) could be used in order to make full use of the available information.

Edge clustering (Ahn et al., 2010) is an alternative, less computationally complex way of finding overlapping communities. It is based on the assumption that a person can be a member of multiple social groups simultaneously, but a relational tie, represented by an edge in the social network graph, has the predominant purpose of associating a person with one particular group. By clustering the edges according to structural similarity, these groups can be recovered in terms of the edges between their members ("link communities"), and can be converted back to overlapping sets of nodes if so desired. The hypothesized relationship between link communities and pervasive overlap is as follows: Usually, people are members of multiple social groups and participate in social contexts created by these groups, but prefer to keep them separate, because each context imposes different norms of behavior, including self-presentation, on the participants. Therefore, one person's egocentric social network can be neatly partitioned into that person's distinct group memberships, but the global structure

that results from the aggregated group membership of many people obscures the group boundaries: "highly overlapping communities can have many more external than internal connections" (Ahn et al., 2010).

Ahn et al. (2010) propose a hierarchical clustering procedure based on a pairwise similarity measure. The similarity of two edges is defined as zero if the edges do not have a common endpoint. Otherwise, for each of the two distinct endpoints, a set is constructed that contains the endpoint and its neighbors, and the similarity is defined as the Jaccard index of the two sets. A hierarchy of clusters is constructed via single-linkage clustering, an agglomerative procedure that, starting from one cluster for each data point, iteratively identifies the two most similar data points that are assigned to different clusters, and merges these clusters, until all data points have been subsumed by a single cluster. Each merge creates a new level in the hierarchy of clusters. The level that maximizes the "partition density", the weighted average of a metric derived from each cluster's density, is accepted as the solution. The SLINK algorithm requires $O(n^2)$ time and $O(n)$ space for single-linkage clustering of n data points (Sibson, 1973). There is no obvious way of parallelizing the clustering itself, but the computation of the similarity matrix and the identification of the maximal partition density can be trivially distributed across multiple threads.

Edge clustering can be adapted to directed graphs (Ahn et al., 2010, supplement S4.1), but in the following experiments we use the undirected variant to enable a comparison with clique percolation. We further note that the line graph framework of Evans and Lambiotte (2009) includes the edge clustering procedure of Ahn et al. as a special case. After transformation to a line graph, any algorithm for non-overlapping community finding can be applied, and appropriate back-transformation yields overlapping communities of the original graph.

By applying the two methods of community detection to the explicit and implicit social network graphs of the various social platforms, we can empirically compare their performance and, at the same time, learn about the latent structure of the network graphs. For clique percolation, parameter k is set to a value of three to minimize the number of nodes not assigned to any community. Edge clustering does not have any tunable parameters. We compare two heuristics for transforming directed to undirected graphs: inserting an undirected edge between two nodes if they are connected by at least one directed edge, and only inserting an undirected edge if there is a bidirectional connection.

The quality of the detected communities is usually evaluated by comparison to ground truth. In some cases, explicit information about group membership is (partially) available, e.g., some SNS let users form topically focused discussion groups with public membership rosters. Otherwise, reference data must be obtained by manual annotation, which is usually not feasible for large social networks, or derived from node attributes external to the graph, under the assumption that, due to homophily, the similarity of two nodes in terms of these attributes is correlated with the likelihood of having one or more communities in common. Hric et al. (2014) test a variety of graphs for alignment of their communities with prominent, observable node attributes, and find that, in many cases, no such alignment exists. Yang and Leskovec (2012) obtain similar results for networks with known group membership. However, Newman and Clauset (2016) point out that "there are often multiple meaningful community divisions of a network [...], and the fact that one division is uncorrelated with a

given metadata variable does not rule out the possibility that another could be."

Instead of evaluating the detected communities against external reference data, we rely on intrinsic metrics inspired by known failure modes of clique percolation. Without reference data, one cannot check if communities correspond to actual social groups, but it is possible to rule out communities that violate intuitive notions about the nature of social groups. The results are collected in table 3.8 for clique percolation and table 3.9 for edge clustering. In both tables, "bd." refers to the transformation of directed graphs that creates an undirected edge for each bidirectional connection.

The ratio of communities to nodes provides a way to compare the number of detected communities across graphs of different size. *Coverage* refers to the percentage of nodes that are member of at least one community. While it is certainly possible that a node is not member of any community, or the graph structure does not provide sufficient evidence for community membership, one would expect this case to be rare. A set of related metrics summarizes the size distribution of the communities: The size of the largest community, relative to the number of covered nodes, indicates whether the graph contains one giant community or decomposes into many smaller ones. The larger the network, the less plausible is the existence of a social group that includes the majority of actors. The ratio of the size of the second largest to the size of the largest community indicates whether or not there is a group of similarly sized, large communities. Furthermore, communities that contain the minimum number of nodes (two for edge clustering, k for clique percolation) do not contain information beyond the already known structural elements of the graph (edges, cliques), so a high percentage of such communities is a negative indicator of quality.

Both methods of community detection do not perform satisfactorily on the social network graphs. Clique percolation shows a variety of failure modes: The comparatively large and dense explicit Facebook graph cannot be processed due to a lack of memory, while the small implicit e-mail network graphs either do not contain any cliques of size k or above, or all cliques percolate into a single community. The remaining graphs either have low coverage, a giant community that includes the majority of nodes, or both. Edge clustering is more robust in that it works equally well for small and large graphs, and, with respect to our metrics, produces more reasonable communities for all graphs. However, smaller graphs still tend to form a giant community, while a high number of communities is identified in large graphs. For most graphs, the median community size is equal to the minimum size or slightly above, which implies a long tail of tiny communities, and is consistent with the observed percentage of minimum size communities. Although edge clustering is able to uncover some of the latent community structure, no progress is made towards the original goal of producing a meaningful visual representation of the graph. When transforming directed to undirected graphs, the requirement of bidirectional connectivity has a strong sparsifying effect and therefore severely reduces the number of communities that are found. When applied to the Twitter follower graph, this appears to be beneficial, as the graph structure is simplified to the point that clique percolation becomes tractable. Applied to implicit network graphs, however, the transformation often substantially reduces the coverage of the resulting communities.

Further investigation is required to find out why state-of-the-art community finding algorithms perform badly on real-world social network graphs. Up to this point, or study has

Table 3.8: Metrics of the communities identified by clique percolation

platform	network	comm.	comm. per node	coverage	largest comm.	2nd/1st largest	min.size comm.
Twitter	explicit (bd.)	197	0.007	79.0%	98.7%	0.1%	79.7%
	implicit	528	0.034	48.1%	87.7%	0.6%	75.4%
	implicit (bd.)	384	0.025	13.9%	55.3%	4.0%	72.4%
Facebook	explicit	-[1]					
	implicit	197	0.038	14.3%	5.6%	61.9%	59.9%
	implicit (bd.)	13	0.002	0.8%	10.0%	75.0%	92.3%
Enron	implicit	-[2]					
	implicit (bd.)	-[2]					
HT (en)	implicit	1	0.021	44.7%			
	implicit (bd.)	-[2]					
HT (it)	implicit	1	0.021	74.5%			
	implicit (bd.)	1	0.021	89.4%			

[1] Not enough memory.
[2] Graph does not contain any cliques of minimum size or above.

Table 3.9: Metrics of the communities identified by edge clustering

platform	network	comm.	comm. per node	coverage	largest comm.	2nd/1st largest	min.size comm.
Twitter	explicit (bd.)	441 130	14.704	93.2%	5.7%	47.7%	92.9%
	implicit	55 116	3.530	99.2%	9.4%	41.0%	88.3%
	implicit (bd.)	13 798	0.884	54.0%	0.6%	76.5%	87.3%
Facebook	explicit	192 915	11.460	100.0%	4.2%	71.1%	86.1%
	implicit	3 867	0.739	93.1%	0.3%	100.0%	79.6%
	implicit (bd.)	395	0.075	16.1%	0.8%	85.7%	73.2%
Enron	implicit	96	0.762	100.0%	14.3%	88.9%	69.8%
	implicit (bd.)	64	0.508	61.9%	7.7%	83.3%	78.1%
HT (en)	implicit	26	0.553	100.0%	70.2%	39.4%	46.2%
	implicit (bd.)	40	0.851	97.9%	52.2%	54.2%	62.5%
HT (it)	implicit	8	0.170	100.0%	85.1%	30.0%	12.5%
	implicit (bd.)	12	0.255	95.7%	80.0%	30.6%	33.3%

been guided by the assumption that all relevant information about community structure is encoded into the graph itself, and pervasive overlap is to blame for the difficulty of uncovering community structure. Some alternative hypotheses might be worth investigating: If communities tend to form around a celebrity (Lim and Datta, 2012) or some other kind of social object (see section 2.1.3), would external information about such entities be helpful? Is there a negative impact of our Twitter crawling scheme, which includes heuristics for avoiding well-connected users, on the preservation of community structure? Automatically generated accounts of spammers that form densely connected subgraphs without any natural community structure may further contribute to the difficulty of clustering the explicit graph. Finally, the user base of Twitter is often claimed to be partitioned into producers and consumers of content (Wu et al., 2011), so community detection algorithms that are specifically designed for bipartite graphs may yield better results.

3.7 Ethical Considerations

Considering the ethical implications of ones actions is an important part of any scientific endeavor; particularly so if an experiment or observational study involves humans, whether as active participants or as passive subjects of an analysis of trace data. Any kind of research that involves the collection of data from a social platform requires a discussion of the associated ethical questions. Ethical standards for research at the intersection of social science, Internet research, and machine learning are still emerging (Zwitter, 2014), so it is not possible to build upon existing, codified norms of the research community. Some guidance can be derived from previously published studies, as they represent a weak consensus on what kind of research is ethically justifiable. Beyond that, the researcher is called upon to identify the relevant ethical questions and to propose answers. This process necessarily involves some subjectivity on the part of the researcher, who is embedded in a personal framework of norms and values, so the results of the following discussion should not be seen as a final statement on the ethical legitimacy of the experiments described in this work, but as a contribution to the ongoing dialogue. Legal aspects are deliberately excluded from this discussion, considering that a meaningful treatment of the arising issues would require collaboration with legal professionals, which is out of scope for this work.

Focus is placed on three moral responsibilities of the researcher, a subset of the "accountabilities" identified by boyd and Crawford (2012): First, towards ones peers, a responsibility to ensure correctness and reproducibility of experiments and their results. This entails publishing, or making available upon request, all material that is required to successfully repeat the experiment, including even personal data collected from the participants. It naturally follows that the second responsibility is towards the participants, who must not incur harm from the experiment. This entails a specific responsibility to protect their privacy and respect their individual rights. Third, and finally, there is a responsibility towards society as a whole, which arises from the notion that scientific research should work towards the benefit of mankind. Applied to the domain of social media research, this maxim implies a responsibility to study online platforms that are relevant, in the sense that they act as hosts to the discussion of current social, cultural, or political movements, or otherwise contribute

to their enactment. In selecting these responsibilities, we make no claim to generality or exhaustiveness, but intend to set up a frame for the discussion.

No single one of these moral responsibilities may overrule the others. Neither the desire for reproducibility nor the desire for relevancy can be allowed to subvert the individual right to privacy, but conversely, the desire for protecting individual privacy cannot be allowed to completely inhibit a researcher's ability to conduct reproducible and relevant research (Salganik, 2017, ch. 6.6.1). In the following, we investigate how an appropriate compromise may be found. The desire for relevant research provides the overall motivation for this discussion: How can research on user behavior on online social platforms be conducted in an ethically responsible way? The desire to protect individual privacy prompts us to ask which methods of data collection and data use are acceptable, while our responsibility as researchers makes us look for responsible ways of sharing data.

Each of the datasets that were introduced in this chapter poses its own set of ethical problems. Twitter has been the subject of a large body of research (Rogers, 2014). Its popularity among researchers can be attributed mainly to the public nature of tweets, which is deeply rooted in the design of Twitter: Tweets are visible to the public by default, and mechanisms for sharing tweets and embedding them into websites are provided. The phenomenon of a tweet reaching a disproportionally large audience by "going viral" is well-known even to non-technically-minded users, and tweets frequently cross medium boundaries by being featured in traditional news media. Twitter makes tweets freely accessible through a rate-limited API (Twitter, Inc., n.d.d), sells high-volume API access to enterprise customers (Twitter, Inc., n.d.c), and has donated data to the US Library of Congress (2013). In sum, the public nature of Twitter communication can be assumed to be generally known and accepted. However, Zimmer and Proferes (2014) cite negative reactions to the archival of tweets by the Library of Congress as evidence that parts of Twitter's user base had up to that point misunderstood tweets to be ephemeral, and considered the large-scale archival of historic tweets, together with public accessibility of that archive, to be a violation of their privacy.

Despite these objections, current consensus among researchers appears to be that the use of public Twitter data is ethically permissible. Facebook is substantially different in that respect. While there are precedents for the study of textual communication on Facebook (e.g., Gilbert and Karahalios, 2009), its dual nature as a medium for communication among personal acquaintances and information broadcast to an unspecified and potentially large audience complicates the assessment of the conditions under which data collected from Facebook can be used in an ethically legitimate way. Individual posts can be private, visible to specific users, or visible to any other user, i.e., to the entire public space formed by the user base of Facebook. Is a publicly accessible post more akin to a statement that is explicitly addressed to the public, or to a conversation that happens to be held in a public space, but may possibly be private in nature? This question has sparked a broader discussion of whether the public accessibility of social media data — on its own — entails permission for researchers to collect and use it (Zimmer, 2010; boyd and Crawford, 2012).

Both e-mail datasets, Enron and HT, are being freely distributed online and have been previously used in published research. The Enron corpus, the first large-scale dataset of organizational e-mail communication available to researchers, has been analyzed in a wide

range of studies (Diesner et al., 2005), while the lesser-known HT dataset has mainly been the subject of conference workshops' "shared tasks" (e.g., El Aassal et al., 2018). The Enron corpus was published by an organ of the US executive in the course of legal proceedings, while the HT data was leaked to the public by "hackers". In both cases, neither the sender nor the recipients of the messages consented to their publication. Furthermore, it cannot be assumed that the senders had any reason to anticipate that their messages would ever be accessible outside of their own and the recipients' corporate environment. The Enron corpus is known to contain sensitive personal information, ranging from institutional data like social security and credit card numbers to details of the intimate life of Enron employees (Cassidy and Westwood-Hill, 2013). It is reasonable to believe that the same is true for the HT dataset. It follows that the use of such material in research is ethically questionable (Poor and Davidson, 2016). In the cases of these two particular datasets, some justification for their use can be derived from the fact that both are primary sources pertaining to incidents of interest to the worldwide general public, and the messages that make up their content were written and sent in the context of work life, where expectations of individual privacy are lower, compared to communication in a domestic setting.

In summary, there are two distinct cases: The public availability of the two e-mail datasets releases us from the obligation to address the ethical questions associated with collection and redistribution of the data. However, it is important to acknowledge that private communication was publicized without the consent of the senders and recipients, potentially causing them harm. The ethical legitimacy of research using these e-mail datasets hinges on the question of whether the possibility of causing additional harm by conducting experiments and publishing the results can be ruled out with sufficient certainty. In the case of Facebook, and, to a lesser extent, Twitter, we first have to make an argument for the legitimacy of data collection. Ethical concerns arise from the tension between the accessibility of a message to a large potential audience, and the preconceptions and expectations of its author on the composition of the actual audience.

3.7.1 Privacy in Online Public Spaces

By definition, any kind of publicly accessible communication on a social platform is non-addressive to some degree. Can there be an expectation of privacy if communication happens in public and does not have a clearly delineated circle of recipients? When approaching this problem from the theoretical side, one encounters several competing points of view.

In the field of medical research, ethics and legal regulations are closely tied. In the US legal system, research that involves *human subjects* is "subject to federal regulation and oversight" (Solberg, 2012). If a researcher obtains information about a living, natural person by means of intervention or interaction, or observes the behavior of said person in a context in which observation is not the norm, that person is considered a human subject. Solberg (2012) argues that "collect[ing] data for a study solely by mining social networking sites" is not sufficient grounds for classifying a study as human subjects research, because "individuals have a very limited expectation of privacy when it comes to information they post online". Ballantyne and Schaefer (2018) go one step further and suggest that individuals may even have a civic duty to share personal data with researchers, if the research in question

is directly beneficial to others. On the other side of the debate, Internet scholars call into question the dichotomy of private and public data: "People may operate in public spaces but maintain strong perceptions or expectations of privacy." (Markham and Buchanan, 2012)

Contextual Integrity and the Imagined Audience

According to Nissenbaum's theory of contextual integrity (2004), the production, dissemination and consumption of information in social spaces is governed by highly context-dependent norms, adherence to which is socially expected even if the space is public in the sense that access is not explicitly restricted. Nissenbaum distinguishes norms of appropriateness, which define the kind of information that can be disclosed in a given context, from norms of information flow, which define how information, after its disclosure in a given context, can be further distributed. Norms of information flow may give rise to expectations of privacy that are incongruent with the accessibility of the information. A researcher that receives information within a certain context is called upon to abide by its norms and thus to preserve the contextual integrity.

Nissenbaum's norms of information flow apply to dyadic relationships between actors in a social network. In the case of non-addressive communication, information flows from an actor to an unspecified group of others, including subscribers, neighbors in the social graph, and accidental listeners. Arguably, the motives for sharing information with a group of unknown composition can be better understood by framing this kind of social media usage as a public performance. Marwick and boyd (2011) introduce the concept of an *imagined audience*, which reflects the user's expectations at a point in time where no information about the actual audience is available. The imagined audiences of social media users include "both those [...] whom they are actively thinking of when sharing content as well as those they may consider as a potential viewer of that same piece of content later on" (Vitak et al., 2015). Public communication is the product of dialectic tension between this imagined reality and the speaker's communicative goals.

The imagined audience conflates the *target audience* that the speaker actively seeks out and the *potential audience*, a broader group of people that the speaker expects to be able to receive the message. For the purpose of this discussion, it proves beneficial to distinguish the two groups. First, the speaker conceives of a message with a target audience in mind, and chooses a public space under the assumption that the potential audience afforded by that space is sufficiently similar to the target audience. Tension results from the awareness that the potential audience is likely to include people who are not in the target audience, possibly at a disproportionally large ratio. If the chosen space is an online social platform, the speaker may decide to resolve this tension by technical means of privacy management, for example, by choosing a more private channel of communication or limiting the visibility of the message to select individuals (Vitak et al., 2015). Litt and Hargittai (2016) point out that this frequently does not happen, and list a number of possible reasons: the speaker may be unaware of the privacy management functions that are offered by the platform, may lack the technical skills to use them, may see no other way of reaching a particular target audience, or might even appreciate the larger potential audience.

If, at some point, it is infeasible to further adapt the audience to the message, the speaker

has to satisfy any remaining privacy needs by adapting the message to the audience. Researchers have observed several strategies to that end, which are referred to in the literature as privacy management, audience management, or boundary regulation. One such strategy is for individuals to "only post things they believe their broadest group of acquaintances will find non-offensive" (Marwick and boyd, 2011), potentially up to the point of deciding not to post at all, which Vitak et al. (2015) refer to as "self-censorship". The practice of hiding information in plain sight, for example by alluding to shared experiences of the speaker and the target audience, is known as "social steganography". It allows the speaker to send messages that are "technologically accessible by anyone, but socially accessible by the targeted audience only" (Litt and Hargittai, 2016). Finally, a strategy best described as passive privacy management relies on the cooperation of others: Content is published in acknowledgement of its potential visibility to a large audience, but in the hope or expectation that it is not actively sought out. For example, a tweet briefly appears in the timeline of followers before it is drowned out by newly incoming tweets, so its author may reasonably expect that only a subset of followers will actually take notice. The speaker might even expect accidental readers to realize that they are not part of the target audience and exercise "civil inattention" (boyd and Marwick, 2011).

In summary, social media users have access to a rich repertoire of technical and social strategies for privacy management, which they use to address the tension between target audience and potential audience of a message. The imagined audience can be understood as the result of this process, which takes place before the message is sent. Litt and Hargittai (2016) note that social media users are "more likely to engage in audience-reaching strategies to reach their imagined audiences, rather than strategies to limit or exclude people in the potential audience who were not part of the target imagined audience." This paints a picture of the imagined audience as a heterogenous group of people that the speaker is explicitly aware of and either welcomes or at least grudgingly tolerates as listeners. The imagined audience is not only an important factor in the identification of the context and the norms it imposes on the speaker, but it also represents the speaker's expectations towards information flow and can therefore be understood as an implicit norm of its own.

A second point of tension exists between the imagined audience and the actual audience. After "performing" a message in public, the speaker receives feedback about the composition of the *actual audience* from the reactions of audience members to the message. If the potential audience, as conceived by the speaker, does not match the actual group of recipients, the message was exposed to people whom the speaker was not aware of when constructing the imagined audience by means of privacy management. As soon as people are exposed to the message, the speaker can no longer meaningfully recant it or modify its content, so it is fundamentally impossible for the speaker to address a mismatch between the imagined and actual audience. If the imagined audience, and the norms it implicitly represents, have to be considered as fixed at the time of communication, then any unexpected recipient will be in violation of these norms. It follows that any use of communication data by a researcher who is not part of the imagined audience is a violation of contextual integrity, even if the communication happened in public.

A commonly held sentiment is that researchers cannot generally presume to be part of the target audience of public communication. For example, an appendix to the AoIR Ethi-

cal Guidelines 3.0 contains the claim that "researchers are rarely the intended audience of user-generated content" (franzke, 2020). Hoser and Nitschke (2010) elaborate: "Researchers are probably not the audience an average user intends to reach by his or her postings and serving as a research object is normally not the purpose an average user has in mind when posting on a social network site [...]" In an interview study conducted by Fiesler and Proferes (2018), 61% of respondents were initially unaware that their tweets were possibly used for research purposes. It follows that a researcher cannot even reasonably assume to be part of the imagined audience, which encompasses the target audience and a broader periphery.

In the absence of norms that generally permit the research use of social media data, the consent of each participant has to be sought out individually. The practice of obtaining *informed consent* originates from the US legal framework for medical research, but has been adopted by other disciplines, including the social sciences and Internet research (Ess and AoIR ethics working committee, 2002), as a de-facto requirement if an experiment poses a risk of harm to the participants. The process of informed consent typically involves explaining the experiment and the nature of the collected data to prospective participants, documenting the consent of those who choose to participate, providing a point of contact for follow-up questions, and allowing participants to withdraw their consent at any time. Obtaining informed consent before collecting a user's social media data ensures that norms of information flow are not violated. While being desirable from an ethical point of view, informed consent has a number of drawbacks which render it impracticable for large-scale data collection from social media.

First and foremost, the cost of obtaining consent grows linearly with the number of participants, which limits the amount of data that can realistically be obtained by an individual researcher or a small team. Furthermore, an informed consent procedure may be a source of bias in the collected data: selection bias may arise from the fact that privacy-minded users are more likely to refuse consent, and the remaining participants may be influenced in their behavior by their awareness of being observed (Olteanu et al., 2019). In particular, the refusal to participate manifests itself in the social network graph as an arbitrary pattern of missing nodes, which conflicts with our goal of obtaining a dense sample of a platform's social network graph. Finally, one has to consider that the act of asking for consent in itself may violate norms of appropriateness. For example, users might feel bothered by repeated requests for their consent from different research teams, and platform operators might view bulk messaging of a large amount of users as an abuse of their resources ("spamming").

If, at a certain scale, obtaining informed consent is effectively impossible, can large-scale observational social media research be ethically justified at all? The AoIR Ethical Guidelines 3.0 acknowledge this issue (franzke et al., 2020, section 3.1.2) as an ethical dilemma that is inherent to all "big data" studies. In chapter 2, we argue that large-scale observational studies of online social behavior potentially yield insights that cannot be obtained otherwise. The ethical problems associated with the collection of large amounts of observational data cannot be completely avoided, but, under the assumption that the research use of that data is ultimately beneficial to society, an experiment may still be ethically permissible if its overall benefit outweighs the violation of the individual rights of the participants. From this point of view, the researcher's responsibility amounts to minimizing, rather than ruling out entirely, violations of contextual integrity. To understand how this can be done in practice, we

propose a refined model of the imagined audience.

Refining the Imagined Audience

A number of studies have, directly or indirectly, advanced the concept of the *unintended audience*, a group of people that the speaker specifically does not address, and would prefer they remain ignorant of the communication (Vitak et al., 2015). Since the term itself leaves some room for ambiguity, we place emphasis on defining the unintended audience, in parallel to the target audience, as a conceptualization of the desires and expectations of the speaker. This implies that a person can be member of the unintended audience of a message without actually being exposed to that message. In boyd and Marwick's case studies of the online behavior of teenagers (2011), the unintended audience is usually found to be composed of people who have a direct influence on the life of the speaker, such as parents, teachers, or classmates. Marwick and boyd (2011) identify conceptually similar groups of "parents, partners, and bosses" that frequently appear in the unintended audiences of adult social media users. However, it is also conceivable that the speaker wants to exclude a broader and more socially distant group from the conversation, for example, due to ideological differences. Some empirical evidence for the duality of targeted and unintended audience is provided by Johnson et al. (2012), who interviewed Facebook users about their privacy practices. Facebook lets users organize their friends in lists. Asking participants about the purpose of the lists they created, Johnson et al. found that 75% were created to prevent specific friends from seeing non-public posts, while only 17% were created to grant a group of friends exclusive access to certain posts. They further note that throughout the interviews, "in some cases the participants described their [privacy] concern in terms of their intended audience, while others described their concern by the people that should be excluded."

Public communication can now be redefined as the product of a tension between the target audience and unintended audience on one side and the potential audience on the other. The speaker chooses a public space that offers maximal potential exposure to the target audience and at the same time minimizes the potential exposure to the unintended audience. The speaker may then use strategies of privacy management to further improve the expected ratio of target audience to unintended audience. One would expect typical target audiences and unintended audiences to be groups with clearly defined boundaries. For example, boyd and Marwick (2011) remark that "[w]hile teens generally do not account for invisible third parties, they do account for eavesdroppers and gossipmongers." While the imagined audience serves as an implicit representation of "positive" norms of permissible information flow, the unintended audience constitutes an explicit specification of "negative" norms that characterize certain information flows as violations of contextual integrity.

The target audience is conceived together with the message and does not change over time. The imagined audience, being a construct that arises from the speaker's efforts to adapt the message to the potential audience, is effectively fixed: once the message is public, the speaker may learn about the actual audience and regret some of the decisions he or she made in the process of privacy management, but the message cannot be recalled from those who have already received it. In contrast to these two groups, the unintended audience is mutable. The speaker may decide, upon learning about members of the actual audience

whom he or she was not previously aware of, that these are not welcome. Similarly, the speaker may at any time reevaluate members of the unintended audience.

We note that the positive and negative norms that are represented by the imagined and the unintended audience, respectively, are not necessarily complementary. It is possible for an unregulated space to exist between the imagined and unintended audiences. Whenever the speaker learns about unexpected audience members, they are either added to the unintended audience, or remain in this unregulated space. Consequently, the actual audience can be partitioned into four groups of people: first, those who are part of the target audience, second, those who are accepted into the imagined audience by means of privacy management, third, those towards whom the speaker feels indifferent, and fourth, those who are part of the unintended audience. Researchers cannot generally expect to be part of the first two groups. However, efforts can be made to avoid being part of the group that is actively rejected by the speaker. To that end, we try to identify what causes social media users to react negatively to the use of their public data for research purposes.

Identifying General Norms for Social Media Research

The "Facebook emotional contagion experiment" of Kramer et al. (2014) is an example for social media research that has been negatively received not only by scientific peers (Jouhki et al., 2016), but also by news media and the general public (Hallinan et al., 2020). For this experiment, a team of researchers, in collaboration with Facebook, changed the behavior of the news feeds of more than 500 000 randomly selected English-speaking users without their knowledge. Participants were assigned to one of two experiment groups or a control group. Depending on their group membership, posts carrying positive, negative, or arbitrary sentiment were randomly withheld from their news feeds. The authors claim to have the participants' informed consent, since the experiment was "consistent with Facebook's Data Use Policy", which all users agree to at the time of account creation (Kramer et al., 2014).

Hallinan et al. (2020) analyze readers' comments on online news articles on the emotional contagion experiment, and observe a wide range of opinions. Some commenters disagree with "experimentation as a general practice", others specifically complain about a "lack of transparency and consent", or voice concerns about the possible effects of the experiment on users' emotional state, their relationships, and their overall mental health. Commenters in a third group express various forms of acceptance and lack of concern. While Hallinan et al. note that commenters tend to "conflate all forms of platform-based research", which they attribute to "a lack of awareness about research practices", what the commenters object to is not a passive audience of researchers, but specific practices of research: in particular, subjecting people to an interventional study with insufficient procedures for informed consent. This is consistent with the results of an earlier survey of Twitter users by Fiesler and Proferes (2018), who report that "the majority felt that researchers should not be able to use tweets without consent", but add that "these attitudes are highly contextual, depending on factors such as how the research is conducted [...]".

We conduct a brief meta-analysis of three studies published between 2016 and 2018 (Mikal et al., 2016; Kennedy et al., 2017; Fiesler and Proferes, 2018) to get an impression of what kind of research involving public social media data is considered inappropriate by the users

themselves if researchers do not explicitly obtain consent. The studies are referred to by the initials of their authors: Mikal et al. (2016) as MHC, Kennedy et al. (2017) as KEM, and Fiesler and Proferes (2018) as FP. In all three studies, social media users are questioned about their attitudes towards a range of research practices. Two studies are specific to Twitter (MHC, FP), while the third (KEM) examines attitudes towards social media research in general. The three studies vary in the number of participants (26–268), their demographics, and the fictional settings discussed by the participants, which range from general (FP: "use of tweets in research") to specific (MHC: "utilizing public domain Twitter data for population-level mental health monitoring"). Free-form discussion, either with individual participants (FP) or in focus groups (MHC, KEM), is subjected to qualitative evaluation. Two studies (KEM, FP) also provide a quantitative evaluation of interviews or questionnaires with answers on a Likert scale. These quantitative results inform our search for research practices that are considered inappropriate by a large group of users. The qualitative results allow us to verify that all privacy concerns are considered, not just those that have been anticipated by the researchers and incorporated into the study design. By grouping topically related concerns of the participants, we are able to identify a number of specific research practices, pertaining to the collection, use, and dissemination of data and the dissemination of research results, which the participants object to:

1. *Circumventing technical means of privacy management:* In FP, 75.4% of participants are uncomfortable with the use of tweets from protected accounts, which are only visible to manually confirmed followers of the account owner. 63.8% express discomfort with the continued use of public tweets that have been deleted after their acquisition by researchers. Participants in the focus group discussions of KEM strongly oppose the collection of private (direct) messages and posts that have been explicitly restricted in visibility.

2. *Violating expectations of data (in-)accessibility:* When people publish content on social media, they have expectations about who can access that content and under which conditions it can be accessed. Usually, these expectations are created by their own experiences with the platform. For example, the personal news feeds of Twitter and Facebook show the most recent posts first, and place limits on how far one can "scroll back" to access older posts. This might lead users to the mistaken belief that posts become inaccessible or are deleted after a certain period of time (Zimmer and Proferes, 2014). The expectation of ephemerality of tweets is reflected in the survey results of FP, where 48.5% of participants object to research use of the entire tweet history, as opposed to single, recent tweets. The focus group discussion of MHC uncovers further misconceptions about the accessibility of older tweets and the visibility of tweets to non-followers.

3. *Dissemination of non-anonymized data:* In FP, 55.9% of participants are uncomfortable with tweets being quoted verbatim in published research papers if these quotations are attributed to the authors' Twitter handles. Only 25.8% object to the lack of attribution when tweets are being quoted anonymously. In the focus groups of MHC, participants

are more broadly concerned about the use of data if no measures are taken to "protect the identity of the people represented."

4. *Dissemination of non-aggregated data and individual-level results:* In FP, 46.8% of participants are uncomfortable with research that involves small units of analysis ("a few dozen tweets"), while only 21.3% object to the study of larger groups of tweets, where a tweet is "analyzed along with millions of other[s]". Similarly, the focus groups of MHC are worried about a "potential for disaggregation of data to identify individuals", while those of KEM are concerned about the use of data to make predictions about individual behavior.

5. *Combining multiple data sources:* In FP, 55.4% of participants are uncomfortable with research that uses public profile information of Twitter users in addition to the content of their tweets. Only 20.3% object to an analysis of the tweets on their own. This may be indicative of a deeper concern about joining different kinds of data that were intentionally kept separate, and the associated potential for inference of information that was intentionally kept hidden.

6. *Collecting data from vulnerable groups or content of sensitive nature:* In KEM, 34% of participants are concerned about a hypothetical study of suicidal teenagers, while only 10% express the same level of concern about a proposal to collect blog posts in order to study the spread of information. Participants of both KEM and FP are concerned about research use of certain kinds of social media posts, for example, those that are personal, offensive, or embarrassing in nature. KEM further notes that some platforms are generally considered to be more personal than others, even intimate, despite their content being publicly accessible.

7. *Manual exploration and analysis of the collected data:* In FP, 37.3% of participants are uncomfortable with a research setting where human researchers read the collected tweets, while only 16.9% are uncomfortable with entirely algorithmic processing.

8. *Research goals not aligned with values of the participants:* This point of concern appears in all three studies, but takes a different shape in each. Participants in the focus groups of KEM have a more positive attitude towards research that is "for the social good" rather than driven by commercial interests. 26% of participants are concerned about "data mining companies collect[ing] information from [social media] to see how brands [...] are talked about", while only 10% are concerned about academic researchers studying blog posts. FP observes that participants tend to view their social media data as their property, and expect some kind of individual (possibly non-material) benefit in compensation if it is used by others. Both FP and MHC find that participants are concerned about unintended negative consequences of sharing their data with researchers.

We identify eight points of concern with social media research practices, six of which are attested in two or more independent studies. In terms of Nissenbaum's theory, they can be understood as negative norms of behavior, i.e., research practices that constitute violations

of contextual integrity and therefore should be avoided. These norms apply to the space of social media in general. Working with general norms glosses over individual differences, which particularly affects members of vulnerable groups, e.g., children or members of the LGBT community (McDonald and Forte, 2020), which often form a minority group inside of a larger context. These groups may have their own norms, expectations, and strategies of privacy management, so their perception of research practices may differ as well. However, we believe that, if inquiry about individual norms (or equivalently: obtaining informed consent) is impracticable due to the number of participants, general norms are a workable approximation. In conclusion, the data collected from Twitter and Facebook poses similar problems to the e-mail datasets discussed earlier: legitimate use of the data is possible if the contextual integrity is protected as described above. However, special care must be taken to directly protect the privacy of vulnerable individuals.

3.7.2 Consequences for Research

Concluding the discussion of social media research ethics, we describe the actual measures taken to protect the privacy of the people whose data was collected in the context of this work. The eight points of concern identified earlier are addressed roughly in order of the research process. First off, we believe that the research conducted in this work carries a potential benefit to society. In chapter 1, we present an argument to that end: understanding social relationships and how they serve as conduits for social influence may contribute to understanding higher-level social phenomena like politically motivated manipulation of public discourse.

Preparation for our experiments begins with the collection of real-world data from online social platforms. Only data that is publicly accessible was considered for use in the experiments. This includes communication that happens in an online public space (Twitter), is accessible to any registered user (Facebook), or has been published by a third party (the two e-mail collections). No deception or circumvention of technical access restrictions was necessary to obtain the data. Where possible, the scope of collection was limited to the type and specific amount of data required for the experiments. In the case of Twitter, where data can be retrieved via an API, the active collection was limited to user profile data and tweets posted within a window of two months, directly preceding the start of crawling. In the case of Facebook, profile and post data was acquired by means of screen scraping, and individual posts and their meta-data were extracted from the raw data afterwards. Since post dates were not accessible at the time of data acquisition, the timelines of the visited users had to be retrieved in full. Similarly, there are no technical means to reliably identify sensitive content or members of vulnerable groups, neither during the crawling process nor afterwards, so data was collected indiscriminately from all visited users. Measures were taken to minimize the network traffic and computational load on the side of the platform operator, such as delays between requests for data, loading only the required resources (e.g., excluding images when loading Facebook profile pages), and taking care not to request the same content twice by means of bookkeeping and caching.

The experiments do not require any interaction with the participants nor any interventions that would affect the participants in any way. Manual exploration, annotation, or pro-

cessing of raw, unaggregated data is minimized. As one of the earliest processing steps, the textual communication artifacts are transformed into a vector space representation via topic modeling, while vertices and edges of the social graph are identified by synthetic numeric identifiers, so that any accidental identification of individual participants by the researchers is ruled out. However, for each of the datasets, a manual review of a small sample of messages was necessary in order to identify medium-specific noise, e.g., e-mail signatures or systematically corrupted meta-data, and testing the efficiency of technical counter-measures. The micro-level social influence experiments described in chapter 6 require manually annotated reference data. In acknowledgement of the more sensitive nature of Facebook and the e-mail datasets, the annotation effort was limited to the posts of 150 Twitter users. In the course of the experiments, some information extracted from user profiles is analyzed jointly with information extracted from communication artifacts. The social influence experiments (chapter 6 and 7) use the explicitly declared social network, i.e, the graph induced by friend or follower relationships, in conjunction with message content. Demographic data from Facebook user profiles is used to assess the dataset's representativity of the general population of Facebook users, but is never directly linked to post content.

Despite public accessibility of their data, social media users expect low individual visibility due to the large amount of information and the high volume of information flow. These assumption is preserved in "big data" studies, where the individual user contributes little to the result, yet the sum of individual contributions leads to relevant findings. When we report experimental results, the unit of analysis is usually the dataset as a whole, with the expectation that the results generalize to the entire population from which the data was sampled. In some experiments, the dataset is subdivided according to certain attributes of users or their messages. In all of these cases, we believe that the results are sufficiently aggregated so that it is neither possible to derive information about individual users, nor to identify these users with the aid of external data. In addition, we do not directly reproduce any of the collected data, which also rules out verbatim quotations of textual content. The exemplary social media posts that were shown earlier to illustrate the unique nature of each social platform originate from people of public interest. The tweet quoted in chapter 1 is well-known, in the sense that it has been frequently retweeted and has crossed medium boundaries to become a "viral image". We feel that, in this case, crediting the author for his or her original thought is more important than protecting the author's (pseudonymous) identity.

In order to ensure the reproducibility of the experiments, other researchers may have to be given access to the collected data. In the absence of a formal vetting process for researchers, granting access to anyone who asks is equivalent to releasing the data to the public. Appropriate measures need to be taken to protect the privacy of the people whose data is being published. One way of doing this would be to release an anonymized version of each dataset from which all personally identifying information has been removed. However, the multimodal nature of the data renders an effective anonymization difficult: the data consists of communication artifacts generated by social media users and their declared social relationship, so a de-anonymization effort may target the communication meta-data, the textual content, the social network graph, or any combination of these modes (Beigi et al., 2018).

Textual data poses a challenge to anonymization. Fung et al. (2010) point out that even if

all personal identifiers are removed from a piece of text, it may still be possible to deduce the identity of the author by aggregating pieces of individually non-identifying information into a "quasi-identifier" that "often singles out a unique or a small number of record owners." In the special case of publicly available communication, anonymization can be circumvented by retrieving the original text via full-text search. Even converting the text to a vector of word frequencies, which despite the loss of information would be sufficient for all experiments conducted in this work, can not reliably prevent de-anonymization. A single, sufficiently unique word may allow identification of the author. This problem can be addressed by not releasing the dictionary, that is, the mapping of word index to the actual word, because the less frequent a word, the less likely is its re-identification by histogram matching or the use of co-occurrence statistics.

The utility of meta-data for author identification can be reduced by an appropriate k-anonymity scheme (Fung et al., 2010), which leaves the problem of anonymizing the social network graph. Using random, synthetic identifiers as node labels instead of the true user IDs effectively conceals the identity of individual nodes, but is susceptible to structure-based attacks (Ji et al., 2017). Any anonymization scheme that involves perturbation of the graph structure has the potential to negatively affect the experiments. Ji et al. (2017) show that such anonymization schemes are good at preserving global structural characteristics of the graph (e.g., degree distribution, centrality measures), but in our experiments, we expect the neighborhood of a node to be an important source of information. It is unclear if an anonymization scheme that is sufficient to prevent re-identification will also preserve the utility of the data for the experiments.

The goal of ensuring reproducibility can also be met by enabling others to repeat the experiment using new data captured under the same conditions as the original dataset. A commonly seen compromise between the two extremes is to only publish the platform-assigned unique IDs of users and posts, along with full documentation and possibly source code of the crawling process, and leaving the actual retrieval ("hydration") to the reproducers. This approach has the added benefit of giving users more control over research use of their data, as unpublishing data from a social platform renders it inaccessible to future researchers, but a large amount of inaccessible data poses an obvious problem for replication. Blodgett et al. (2017) attempt to hydrate a uniformly random sample of tweet IDs and find that only 62% could be successfully retrieved. How many of the unavailable tweets were actively withdrawn by their authors is unknown, as the removal of spam may also be a major contributing factor. Furthermore, in the absence of a stable API, as in the case of Facebook, there can be no out-of-the-box solution for repeating the crawling process. Even if source code is provided, any system for screen scraping is likely to require major updates to keep up with changes in design and functionality of the target website, which raises a large hurdle for reproducibility.

During the experimental phase, the collected data has to be stored on a computer system that is directly accessible to the researcher. Good research practice demands that all data analyzed and produced in the course of the experiments is moved to an archive after completion. This non-anonymized version of the dataset must be kept around for access only by the research team, enabling them to address external criticism of methods and practices, as well as allowing them to answer questions about aspects of the data not discussed in the orig-

inal research work. This creates a certain risk for the subjects of research, who are unable to withdraw their data from the researchers' archive. The negative effects of archival can be mitigated by ensuring that proper IT security measures are in place to prevent unauthorized access to the data, and that access to the data after conclusion of the original experiments is only granted for the purposes outlined above.

In conclusion, these measures, especially when seen in the context of other studies conducted on similar datasets, make us consider the experiments presented in this work to be ethically acceptable. In the medical field, researchers who are operating in ethically controversial settings benefit from an infrastructure of institutional review boards (IRB), which can already provide meaningful guidance in the early design phase of a research project. Not having access to a local IRB that is familiar with social media research, we would welcome further public discussion on the ethics of conducting large-scale experiments on user data from online social network services.

4 Probabilistic Topic Models for Online Communication

Parts of this chapter are based on work that was done in the context of the master's theses of Benjamin Koster (2013) and Florian Hartl (2013), as well as the bachelor's thesis of Felix Sonntag (2015).

Probabilistic topic models, as defined by Blei (2012), are mixed-membership models for categorical data. Mixed-membership models are a generalization of mixture models (Airoldi et al., 2014): In a mixture model, the probability of an observation x is expressed as a weighted sum over a finite number of mixture components $p_i(x)$, so that $p(x) = \sum_i w_i p_i(x)$. In a mixed-membership model, observations are grouped, and the group membership g is an observed variable. Each group has its own mixing probabilities $w_{g,i}$, while the mixture components $p_i(x)$ are shared across all groups. A mixed-membership model where observations and their groups correspond to words and documents is known as a topic model, and its mixture components are called *topics*. As a first approximation, a probabilistic topic model can be seen as a black box that takes a *corpus*, i.e., a collection of documents, as input and represents each document as a probability distribution over a fixed number of topics, which in turn are probability distributions over the set of unique words in the corpus.

This well-defined probabilistic interpretation is what sets probabilistic topic models apart from other approaches to uncover the latent semantics of document collections, such as LSI, which maps documents to arbitrary vectors within a low-dimensional space. LSI is functionally equivalent to a probabilistic topic model in that it aggregates semantically related words and represents documents in terms of the resulting semantic classes, but lacks probabilistic interpretability. Another benefit of probabilistic topic models is the strong correspondence between the mathematical formulation of topics and the human concept of "conversation topics": In an experimental setting, the generated topics and document-topic distributions have been assessed as semantically coherent and appropriate descriptions of documents by human annotators (Chang et al., 2009b).

Latent Dirichlet allocation (LDA; Blei, 2012) is the most basic implementation of probabilistic topic modeling. Its simplicity and direct applicability to all kinds of textual data has caused LDA to be generally accepted as the reference topic model, from which more specific models are derived, and against which such models are evaluated. In this chapter, we first describe how LDA works from a theoretical point of view and how to implement it, and then examine the particular challenges of applying LDA to social media data. Finally, we discuss how some of the identified shortcomings of LDA can be overcome and novel insights can be gained by adapting the model itself to the structure of the data.

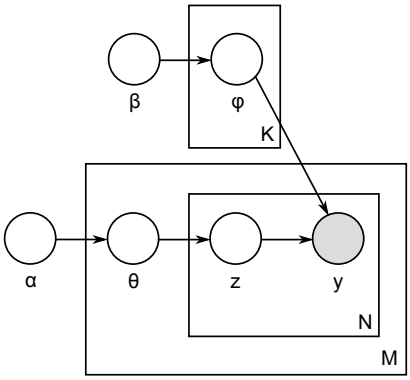


Figure 4.1: Graphical structure of latent Dirichlet allocation (adapted from Blei et al., 2003)

4.1 Latent Dirichlet Allocation

We pick up the discussion of LDA where we left off in section 2.2.3. Figure 4.1 visualizes the assumptions of independence among the variables of the LDA model as a directed acyclic graph in plate notation (Buntine, 1994). By convention, upper case Latin letters refer to cardinalities, lower case letters identify observations and latent variables, and Greek letters identify model parameters. LDA is usually understood as a Bayesian model: For each unobserved variable in the model, we specify our *prior belief* about its distribution. If this specification depends on other unobserved variables, their prior distribution is specified as well, until we arrive at a distribution that can be fully specified, i.e., a leaf node in the model graph. Parameters of the prior distributions are called *hyper-parameters*. In machine learning terms, topic modeling via LDA is an instance of unsupervised learning: we estimate the distribution of data under certain model assumptions.

The model contains one observed variable $y_{m,n}$ for each of N words in each of M documents. The distribution of observation $y_{m,n}$ is determined by its topic assignment: the associated latent variable $z_{m,n}$ selects one of K topics, which are categorical distributions (multinomial distributions with a single trial) over the set of W unique words. The parameters of the selected topic-word distribution are represented by vector $\varphi_{z_{m,n}}$ ($\varphi \in \mathbb{R}^{K \times W}$), which in turn has a Dirichlet distribution with concentration parameter $\beta \in \mathbb{R}^{W}$. The topic assignment $z_{m,n}$ follows a per-document categorical distribution over the set of topics. Parameter vector θ_m ($\theta \in \mathbb{R}^{M \times K}$) of this document-topic distribution has a Dirichlet distribution with parameter $\alpha \in \mathbb{R}^{K}$. The hyper-parameters α and β are treated as known and fixed quantities for now. Note that the document-topic and topic-word distributions are commonly referred to, *pars pro toto*, by their respective parameters θ and φ. The same is true for the prior distributions and the hyper-parameters.

By combining the independence assumptions and the assumptions about the distribution of the variables, one can formulate the joint probability of latent and observed variables, given the hyper-parameters, as in equation 4.1 (Carpenter, 2010). Here and in all following equations, *Dir* refers to the probability density function of the Dirichlet distribution and *Disc*

to the probability mass function of the categorical distribution.

$$\begin{aligned} &p(\theta, \varphi, z, y|\alpha, \beta) = p(\theta|\alpha) \cdot p(\varphi|\beta) \cdot p(z|\theta) \cdot p(y|\varphi, z) \\ &= \prod_{m=1}^{M} \mathrm{Dir}(\theta_m|\alpha) \cdot \prod_{k=1}^{K} \mathrm{Dir}(\varphi_k|\beta) \cdot \prod_{m=1}^{M} \prod_{n=1}^{N_m} \mathrm{Disc}(z_{m,n}|\theta_m) \cdot \prod_{m=1}^{M} \prod_{n=1}^{N_m} \mathrm{Disc}(y_{m,n}|\varphi_{z_{m,n}}) \end{aligned} \tag{4.1}$$

Considering that LDA is a generative model, that is, a model of the joint probability of observations and parameters, it can be equivalently expressed as a data-generating process ("generative story"):

1. For each topic index $1 \leq k \leq K$:
   a) Sample the parameters φ_k of the topic-word distribution from $\mathrm{Dir}(\beta)$.
2. For each document index $1 \leq m \leq M$:
   a) Sample the parameters θ_m of the document-topic distribution from $\mathrm{Dir}(\alpha)$.
   b) For each word index $1 \leq n \leq N_m$:
      i. Sample the topic $z_{m,n}$ from $\mathrm{Disc}(\theta_m)$.
      ii. Sample the word $y_{m,n}$ from $\mathrm{Disc}(\varphi_{z_{m,n}})$.

Some assumptions of the model are not directly obvious from the specification: The per-word variables are grouped by document, but are exchangeable within the document, so changing the word order does not affect the likelihood. This corresponds to the bag-of-words assumption of vector space models. In section 2.2.3, we show that, under certain conditions, LDA can be interpreted as an approximative factorization of the matrix of per-document word counts. It follows that topics reflect the co-occurrence of words in documents, with two words being more likely to be assigned to the same topic if they frequently co-occur. Finally, the hyper-parameters α and β allow us to incorporate prior knowledge about the sparsity of the document-topic and topic-word distributions. Positive values less than one make the model favor sparse distributions, which are easier to interpret as distinct concepts. Sparsity also reduces the computational complexity of parameter estimation.

Before moving on to the details of parameter estimation and comparative evaluation, we note that none of the methods discussed in this section are specific to LDA. We present them in the context of LDA, so that a concrete operationalization can be provided, but each method can be applied to a broad class of LDA-like probabilistic topic models, requiring no more than minor adjustments.

4.1.1 Parameter Estimation

Given a collection of documents, we are interested in estimating the document-topic probabilities θ and the topic-word probabilities φ. The Bayesian approach is to obtain these estimates using the *posterior distribution* (equation 4.2), which defines the probability of the latent variables and parameters θ, φ, z given the observations y, and can be derived from the likelihood of the observations by application of Bayes' rule. The numerator of equation 4.2 is equivalent to the joint probability (equation 4.1), and the denominator can be

obtained by marginalizing the joint probability over the unobserved variables θ, φ, z. This probability $p(y|\alpha, \beta)$, defined in equation 4.3, can be understood as the probability of corpus y being generated by all possible instantiations of the LDA model with arbitrary parameters θ and φ. In section 4.1.2, "Estimating the Likelihood", we give a closed-form solution of the integral in equation 4.3, which contains a sum over all K^{N_*} possible values of z, where $N_* = \sum_{m=1}^{M} N_m$ is the overall number of words in the corpus. Its analytical evaluation is therefore computationally intractable (Griffiths and Steyvers, 2004; Blei, 2012).

$$p(\theta, \varphi, z|y, \alpha, \beta) = \frac{p(y|\theta, \varphi, z, \alpha, \beta) \cdot p(\theta, \varphi, z|\alpha, \beta)}{p(y|\alpha, \beta)} \tag{4.2}$$

$$\begin{aligned} p(y|\alpha, \beta) = \iint \prod_{m=1}^{M} p(\theta_m|\alpha) \cdot \prod_{k=1}^{K} p(\varphi_k|\beta) \cdot \\ \prod_{m=1}^{M} \prod_{n=1}^{N_m} \sum_{z_{m,n}} \left(p(z_{m,n}|\theta_m) \cdot p(y_{m,n}|\varphi_{z_{m,n}}) \right) d\theta d\varphi \end{aligned} \tag{4.3}$$

Although the exact estimation of the model parameters via the posterior is intractable, a number of approximative methods exist. Asuncion et al. (2009) compare the two most popular methods, Gibbs sampling and variational inference, and find no substantial difference in the predictive performance of the fitted models. We therefore approximate the posterior via Gibbs sampling, a Markov chain Monte Carlo (MCMC) method. Since we aim to implement and compare different LDA-like models, an important advantage of Gibbs sampling is that a sampling procedure for the posterior of most directed graphical models can be derived in an almost mechanical way (Buntine, 1994).

Approximation of the Posterior via Gibbs Sampling

Given a vector of model parameters Θ and observations O, Gibbs sampling works by constructing an ergodic Markov chain with a stationary distribution $\pi(\Theta)$ that matches the posterior $p(\Theta|O)$. Starting from a random initialization $\Theta^{(0)}$ of the parameter vector, a random walk on this Markov chain will, after a certain number of steps, reach a convergent state, in which $\Theta^{(i)}$ can be interpreted as samples from a distribution that approximates the posterior (Liu, 2001, ch. 6). The transition matrix of the Markov chain is not defined explicitly. Instead, for each of the n components of Θ, the univariate *full conditional distribution* $p(\Theta_i|\Theta_1, \ldots, \Theta_{i-1}, \Theta_{i+1}, \ldots, \Theta_n)$ of that component Θ_i given all other components[1] Θ_{-i} is derived from the multivariate joint distribution $p(\Theta_1, \ldots, \Theta_n)$. Informally speaking, Gibbs sampling is applicable whenever these univariate distributions are easy to sample from. One iteration of Gibbs sampling performs one step of the random walk as follows: given the previous sample $\Theta^{(t-1)}$, each component $\Theta_i^{(t)}$ of the new sample is drawn from its full conditional distribution, conditioned on the current value of all components $\Theta_{j<i}^{(t)}$ that have already been determined, and the previous iteration's value of the remaining components $\Theta_{j>i}^{(t-1)}$.

[1] From here on, we use x_{-i} as a shorthand for all variables from a sequence x, excluding x_i.

When applying Gibbs sampling to LDA, instead of sampling the space spanned by all unobserved variables, it is common to "collapse" the joint distribution by marginalizing over θ and φ, and only sample the latent variable z. This strategy is known as collapsed Gibbs sampling (CGS). Collapsing the joint distribution is counter-intuitive, considering that estimating θ and φ is what motivates our application of Gibbs sampling in the first place. However, estimates of θ and φ can be easily recovered from samples of z, and CGS has several desirable properties, compared to regular Gibbs sampling. A direct effect of collapsing is that fewer variables have to be sampled per iteration, and these variables follow a categorical distribution, which is less computationally expensive to sample than the Dirichlet distribution. From a theoretical point of view, collapsing accelerates convergence of the sampler (Liu, 2001, ch. 6.7). There are few systematic studies on the practical utility of collapsing in the context of LDA parameter estimation. Newman et al. (2009a, ch. 2.1) briefly examine the effect of collapsing on the likelihood of held-out data, and confirm that a model that was fit using a collapsed Gibbs sampler attains a higher likelihood after fewer iterations of sampling.

The full conditional distribution of $z_{m,n}$ in a collapsed Gibbs sampler, first published by Griffiths and Steyvers (2004), is given in equation 4.4. A more detailed derivation is provided by Carpenter (2010). The probabilities only have to be known up to proportionality, as a normalization factor can be trivially computed by summing the probabilities of all K topics. Expressions of the form $c_{k,m,w}$ are functions of y and z: they represent the number of words that are assigned to topic k, contained in document m, and have index w in the vocabulary, with the asterisk acting as a wildcard (i.e., a summation over all possible values of the corresponding variable). The superscript $c^{-(m,n)}$ indicates that the word at position (m, n) is excluded from the count. Computing the word counts for each evaluation of the probability mass function can be avoided by maintaining a $M \times K$ matrix of document-topic counts, a $K \times W$ matrix of topic-word counts, and a K-dimensional vector that counts the overall number of assignments to each topic, and updating them whenever a new topic assignment is sampled. Due to the exchangeability of words within documents, these data structures contain sufficient information for parameter estimation, and there is no need to explicitly store the individual topic assignments $z_{m,n}$ (Griffiths and Steyvers, 2004).

$$p(z_{m,n}|z_{-(m,n)}, y, \alpha, \beta) \propto (c^{-(m,n)}_{z_{m,n},m,*} + \alpha_{z_{m,n}}) \cdot \frac{c^{-(m,n)}_{z_{m,n},*,y_{m,n}} + \beta_{y_{m,n}}}{c^{-(m,n)}_{z_{m,n},*,*} + \sum_{w=1}^{W} \beta_w} \tag{4.4}$$

When starting the Gibbs sampling procedure from a uniformly random initialization of topic assignments z, the first t samples are unlikely to be a good approximation of random samples from the posterior, as the Gibbs sampler has not reached a convergent state yet. An important question, therefore, is how to determine the number of initial "burn-in" iterations, after which the samples are usable. Gibbs sampling is Markovian in that the probability of drawing a particular sample at iteration t only depends on the sample drawn at $t - 1$. If it is possible to initialize the sampler to an arbitrary point in a high-probability region of the posterior, there is no need for burn-in. However, the LDA posterior is what Geyer (2009) calls a "black box" distribution: we only know its unnormalized probability density, which is insufficient for identifying a suitable starting point for Gibbs sampling, and therefore also

insufficient for testing for convergence.

Distribution-independent convergence diagnostics exist, but cannot detect convergence in general, only certain manifestations of its absence (Cowles and Carlin, 1996). Furthermore, these diagnostics usually require additional computational effort, e.g., for running multiple instances of Gibbs sampling in parallel. Heinrich (2009, pp. 29) proposes the more pragmatical approach of testing for "convergence of some measure of model quality", in particular the likelihood of held-out data, as a proxy for convergence of the sampler. However, estimating this predictive likelihood requires non-negligible computational effort (see section 4.1.2, "Estimating the Likelihood") and estimates are noisy. In preliminary experiments, we found it difficult to determine if the "elbow point", after which the increase in test-set likelihood slows down, had been reached. Geyer (2009) concludes: "In the black box situation, the best diagnostic is to run the chain for a very long time [...]" In line with this recommendation, we perform 2 000 iterations as a compromise between computational effort and likelihood of reaching convergence.

Estimating the Parameters of the Document-Topic and Topic-Word Distributions

Given a sequence of samples from the posterior distribution of z, how do we obtain estimates of the parameters of interest? Griffiths and Steyvers (2004) express φ and θ as the posterior predictive probabilities associated with a hypothetical, still unobserved word $y_{m,+}$ and its topic assignment $z_{m,+}$ at position $(m, N_m + 1)$ in the corpus. The probability of observing a particular topic assignment $z_{m,+}$ is an estimate of θ_m, and the probability of observing word $y_{m,+}$ assigned to topic k is an estimate of φ_k. The posterior predictive density of $y_{m,+}$ (equation 4.5) includes the (collapsed) posterior probability as a factor, so, like the posterior itself, it is not analytically tractable. However, since it can be expressed as the expected value of a function of z, it can be approximated from a sequence $z^{(s)}$ of Gibbs samples from the posterior using the standard Monte Carlo sampler (Papanikolaou et al., 2017). The right-hand side of equation 4.5 is the expectation of $P(z) := p(y_{m,+}|z, y, \alpha, \beta)$ with respect to the posterior. It follows from the Ergodic Theorem for Markov chains that the arithmetic mean $\frac{1}{S}\sum_{s=1}^{S} P(z^{(s)})$ will converge to that expectation as $S \to \infty$. The predictive distribution of $z_{m,+}$ can be approximated in the same way.

$$p(y_{m,+}|y, \alpha, \beta) = \sum_{z} p(y_{m,+}|z, y, \alpha, \beta) \cdot p(z|y, \alpha, \beta) \tag{4.5}$$

Griffiths and Steyvers (2004) provide closed-form estimators $P(z)$ for θ and φ from a single Gibbs sample z (equations 4.6 and 4.7). The count $c_{*,m,*}$ is simply N_m, the number of words in document m, so these estimates can be computed directly from the count matrices of the Gibbs sampler.

$$\hat{\theta}_{m,k} := p(z_{m,+} = k \mid z_m, \alpha) = \frac{c_{k,m,*} + \alpha_k}{c_{*,m,*} + \sum_{k'=1}^{K} \alpha_{k'}} \tag{4.6}$$

$$\hat{\varphi}_{k,w} := p(y_{m,+} = w \mid z, z_{m,+} = k, y, \beta) = \frac{c_{k,*,w} + \beta_w}{c_{k,*,*} + \sum_{w'=1}^{W} \beta_{w'}} \tag{4.7}$$

From these equations, it becomes apparent in what way LDA is similar to the methods for latent semantic analysis discussed in section 2.2.3: Sampling the topic assignments z from the posterior is an implicit factorization of the document-term co-occurrence matrix into matrices of document-topic and topic-word counts. The probabilities θ and φ are derived from these matrices by normalization after smoothing, the strength of which is specified by the parameters α and β of the Dirichlet priors.

While θ and φ can be estimated from a single Gibbs sample, this practice does not properly account for the variability of the posterior. Nguyen et al. (2014) and Papanikolaou et al. (2017) independently provide empirical evidence for the benefit of averaging estimates from multiple samples. Another problem with estimation from a single sample stems from the fact that we are trying to estimate continuous quantities θ and φ from discrete variables y, z. In an estimate $\hat{\theta}_m$ according to equation 4.6, each component can only take on one of $\frac{1}{N_m}$ possible values. In consequence, the granularity or resolution of $\hat{\theta}$ is limited by the length of documents in the corpus, and social media text in particular tends to be short (see section 4.2.1). In an actual implementation, the resolution may be even lower, as not all possible values necessarily have a unique floating point representation. Estimates of φ are affected to a lesser degree, because in equation 4.7 the variable part of the denominator is the number of words assigned to a particular topic, which is bounded above by the size of the corpus. Some practical consequences of the limited resolution are discussed in section 6.2.2, "Robustness to Violation of Modeling Assumptions". Increasing the number of samples will increase the resolution proportionally.

In general, a Monte Carlo estimate should be formed from as many samples as computationally feasible. However, LDA suffers from the *label switching* problem, rooted in the non-identifiability of the topics φ: In a model with symmetric Dirichlet priors α, β (see 4.1.1, "Choice of Hyper-Parameters"), the posterior density has as many modes as there are permutations of topics (Dietz, 2011, ch. 2.4.4). In other words, all permutations of the topic indices (e.g., swapping φ_i and φ_j, then swapping $\theta_{m,i}$ and $\theta_{m,j}$ for all documents m) lead to models of equal likelihood. If at least one prior is asymmetric, this is only true for the limit of the likelihood as the size of the corpus approaches infinity and the evidence overrules the prior (Omar, 2016, ch. 5.3.1).

Even after convergence, the Gibbs sampler is not guaranteed to stay in the vicinity of one particular mode, but may freely move between them. In that case, averaging over samples from two or more modes severely reduces the interpretability of the resulting topic-word distributions. Their parameters φ effectively turn into linear combinations of multiple, likely unrelated topics. As a compromise, we estimate φ from a single sample, directly after burn-in. Then, Gibbs sampling continues with φ held fixed at its estimated value. The corresponding full conditional distribution of $z_{m,n}$ is given in equation 4.8. If the topics φ are constant, label switching cannot occur, and it is safe to estimate θ from an arbitrarily high number of samples (Papanikolaou et al., 2017).

$$p(z_{m,n}|z_{-(m,n)}, y, \varphi, \alpha, \beta) \propto (c^{-(m,n)}_{z_{m,n},m,*} + \alpha_{z_{m,n}}) \cdot \varphi_{z_{m,n},y_{m,n}} \tag{4.8}$$

The samples $z^{(s)}$ produced by the Gibbs sampler are not independent, which directly follows from the Markov-dependency of subsequent samples. This dependency results in autocorrelation of the sequence $z^{(s)}$. Independent samples could be obtained by starting a new

run of Gibbs sampling for each sample, but in the case of LDA this is prohibitively expensive due to the required burn-in. In the topic modeling literature (Griffiths and Steyvers, 2004; Heinrich, 2009), *thinning* is routinely recommended as a method of producing approximately independent samples. The correlation of two samples quickly drops off with increasing distance in the sequence, so only every n-th sample is used for parameter estimation, and the intermediate samples are discarded. It is unclear why this procedure is necessary: MacEachern and Berliner (1994) show that the variance of a Monte Carlo estimate from Gibbs samples is always lowest when using all samples. A discussion of the effects of autocorrelation on the quality of topic models, either grounded in theory or empirical results, is notably missing from the literature. Since conducting our own investigation would be outside of the scope of this thesis, we do perform thinning, if only to comply with the best practices of the field.

For the estimation of θ and φ, we use 20 samples with a thinning interval of 5. Similar to the number of burn-in iterations, these values are not the result of theoretical considerations or an attempt to push the observed variability or autocorrelation below a predefined threshold, but merely driven by the desire to maximize the accuracy of estimation while staying within a given computational budget.

Estimating Topic Distributions of Unseen Documents

After a topic model has been fit to a collection of documents, and the topics φ have been determined, a common task is to estimate the topic distributions θ of new, previously unseen documents. It is assumed that the new documents and the collection used for fitting the topic model come from the same population, so that both have the same topics φ. Heinrich (2009) refers to this process as *querying* the topic model.

Querying means estimating the parameters θ of the topic distributions of one or more previously unseen documents, given the model parameters φ and hyper-parameters α and β. Gibbs sampling is performed exclusively on the new documents. Since φ is known and fixed, the simplified form of the full conditional distribution given in equation 4.8 can be used. Because of the meaningful initialization of topics, and the reduced number of parameters to be fit, one can reasonably expect the sampler to take fewer iterations to converge. We perform 200 iterations for each query, one tenth of the number of iterations for the initial parameter fitting. After sampling, estimates of θ specific to the new documents can be obtained via the estimation procedure for the multinomial parameters.

Choice of Hyper-Parameters

The hyper-parameters α and β, as well as the number of topics K, which can be seen as an implicit hyper-parameter, are not estimated by the Gibbs sampling procedure, and have to be specified beforehand. Finding the optimal value of K is particularly difficult, because there are two conflicting notions of optimality: Increasing the number of topics also increases the overall number of model parameters, which improves the model's ability to fit to the data. Until the point of overfitting is reached, increasing K will improve the predictive performance of the model. However, the interpretability of the model parameters rests on the correspondence between topics as probability distributions over words and topics as mean-

ingful clusters of semantically related words. Steyvers and Griffiths (2007) observe that "[a] solution with too few topics will generally result in very broad topics whereas a solution with too many topics will result in uninterpretable topics that pick out idiosyncratic word combinations."

Choosing a value of K according to some notion of topic quality or interpretability usually involves manual inspection of the topic-word distributions. A common quantitative approach is to perform cross-validation on a small development set and choose the value of K that maximizes some measure of model quality (see section 4.1.2). A principled method of finding an optimal K for a given dataset with respect to likelihood can be found in the Hierarchical Dirichlet Process (HDP), a non-parametric extension of LDA that is able to infer the number of topics from the data (Teh et al., 2006). This ability comes at the cost of higher model complexity and computational effort compared to LDA.

The hyper-parameters α and β are the parameter vectors of the Dirichlet prior of the document-topic distributions and the topic-word distributions, respectively. A K-dimensional Dirichlet distribution has the $(K-1)$-dimensional unit simplex Δ^{K-1} as its support, so it can be interpreted as a distribution over the space of parameter vectors of K-dimensional categorical distributions. A Dirichlet distribution with a parameter vector α is called *symmetric* if all components α_i have the same value. If $0 < \alpha_i < 1$, a symmetric distribution is sparse, that is, it assigns higher probability to a vector if most of its components are zero. The parameter vector of an asymmetric Dirichlet distribution can be expressed as $\alpha = \alpha_0 \cdot m$, which is the product of a scalar concentration parameter α_0 and a vectorial base measure $m \in \Delta^{K-1}$. The base measure m can be understood as a preference towards particular components over others, which, in the context of LDA, translates to higher prior probability of observing certain topics in a document or words in a topic. If α_0 is large, samples from the distribution are likely to be close to m (Minka, 2012).

Steyvers and Griffiths (2007) suggest using symmetric priors with $\alpha_i = \frac{50}{K}$ and $\beta_i = 0.01$. A fully Bayesian approach would be to integrate out the hyper-parameter, but Wallach et al. (2009a) show that similar results can be achieved at a lower computational cost with an empirical Bayes strategy. Applying the Expectation Maximization algorithm, one alternates between estimating the hyper-parameter given the data and current model parameters, and performing one or more iterations of Gibbs sampling with a fixed value of the hyper-parameter. An efficient algorithm for obtaining a maximum-likelihood estimate of the Dirichlet hyper-parameter via fixed-point iteration is due to Wallach (2008, algorithm 2.2).

Wallach et al. (2009a) test the utility of applying this optimization strategy to the two Dirichlet priors. They find that a data-driven choice of β offers no benefit over a fixed, symmetric value. However, a data-driven, asymmetric α substantially improves performance, which reflects the intuitive notion that some topics are more general, and therefore more likely to appear in a document than more specialized topics. They further observe that optimizing α makes LDA more robust to specifying a number of topics K that is higher than the (unknown) true number of topics. Excess topics receive a small prior probability and end up being rarely used, which mitigates the negative effect of a misspecification of K on the likelihood of the model as well as on the interpretability of the topics. We rely on the beneficial effect of a data-driven α instead of attempting to find a number of topics that is

optimal in some sense.

We briefly investigate the relationship between the value of α and the interpretability of a topic. To that end, we manually annotate the 150 topics of a topic model fit to the 30 000 user Twitter dataset described in section 3.1. A topic is characterized by the set of words it assigns high probability to. An attempt is made to label each topic by summarizing these words with a single word or short phrase. Looking at the results of this labeling, one can broadly distinguish three classes of topics: In 64% of topics, the high probability words are clearly associated with one concept, or a small number of concepts (class 1). 16% of topics cannot be easily associated with a concept, because they assign high probability to common words with a very broad or general meaning, or words from a non-English language (class 2). The remaining 20% cannot be labeled, because they assign high probability to words that have a specific meaning, but are associated with many different concepts (class 3).

Hyper-parameter α was optimized, and the values of its components are in the interval $[1.47 \times 10^{-4}, 2.15 \times 10^{-1}]$ with a mean of 7.75×10^{-3}. The classification of a topic i is visibly correlated with the value of α_i. Among the 15 topics with highest α_i (mean $\mu = 5.20 \times 10^{-2}$), ten belong to class 2, the others to class 1. Among the 15 topics with lowest α_i ($\mu = 4.56 \times 10^{-4}$), nine belong to class 3, the others to class 1. In other words, a high α_i indicates overly broad or general topics, while a low α_i indicates a mixture of different concepts. Despite this association, α_i alone is not sufficient to reliably distinguish overly general topics from topics that occur frequently in the corpus, or to distinguish rare topics from topics that mix different concepts.

Updating the Model

In a setting where data arrives as a continuous stream, the process of updating a model and its parameter estimates with each incoming unit of data is called *online learning*. If continuous updates are too expensive, or finely spaced parameter estimates are not required, the model can be updated in larger increments. Here, we are mainly interested in the retrospective application of incremental updating: All data points are available, and each point bears a time-stamp. The time line is subdivided into a sequence t of arbitrary intervals, and we wish to fit a model to the data in each interval t_i, so that the model only contains information from intervals t_j where $j \leq i$.

An efficient procedure for incrementally updating a topic model must meet a number of requirements:

1. The computational effort and memory that is required for performing an update must not grow with the number of previous updates. This implies that the procedure cannot revisit data from earlier increments, which would have to be kept in memory.

2. It must be able to fully use the information provided by the data in the current and all previous increments.

3. In addition to assigning topic distributions to all new documents, the procedure must be able to update the topics themselves.

4. The procedure must either preserve the topic indices (φ_i^t must be more similar to φ_i^{t+1} than to any φ_j^{t+1} with $i \neq j$), or provide explicit information about their mapping. This is to ensure that the document-topic distributions θ are comparable across time.

These requirements rule out various naive approaches. If the amount of data in each increment is sufficiently large, one can fit a separate topic model to each, but this approach completely ignores information from earlier increments and does not guarantee any kind of consistency between the topics of two successive models. Alternatively, one could fit a model to the first increment, and assign topic distributions to documents in subsequent increments by querying. This approach ensures that document-topic distributions are comparable across all increments, but the topics remain fixed after the initial fitting, so the data of the first increment has to be sufficiently representative for all future increments (Canini et al., 2009). The resulting model only contains information from the first and the current increment. Finally, one could fit a model to the union of the current and all earlier increments (Yao et al., 2009, algorithm "Gibbs1"). Consistency of topics can be achieved by using the sampling state of the previous model as a starting point. This procedure satisfies all requirements except computational and memory efficiency. All increments have to be kept in memory indefinitely, and each update has to iterate over all previous data.

According to the principle of Bayesian updating, an existing model can be updated to incorporate new evidence by using the posterior of the existing model as the prior of a model that is fit to the new data. Once again, the intractability of the LDA posterior makes it impossible to derive an exact procedure, but various approximations have been proposed. We only consider those that are based on Gibbs sampling. One can distinguish two classes of algorithms: those that implicitly incorporate prior knowledge by reusing the sampling state of an earlier model, and those that explicitly update the parameters of the prior distributions. Yao et al. (2009, algorithm "Gibbs2") keep the topic assignments z of old documents fixed, and only perform sampling for the new documents. Unlike querying, the topics φ are not held fixed, and an updated estimate can be obtained after sampling via equation 4.7. This procedure is memory-efficient, since only a part of the previous sampling state, the matrix of topic-word counts, has to be saved in preparation for the next update. Starting from meaningful topics and not resampling old documents makes label switching unlikely to occur.

Canini et al. (2009) extend this procedure as follows: after sampling a topic assignment for a word from a new document, the topic assignments for a subset of earlier words, which may include words from previous increments, are re-sampled as well. Speed of convergence can be improved by the adoption of a Sequential Monte Carlo (SMC) sampling strategy (see the work of Scott and Baldridge, 2013 for details on the application of SMC to LDA parameter estimation), but the re-sampling of old topic assignments requires data from all earlier increments to be kept in memory. The algorithm is still notable, because unlike all other algorithms discussed here, it is able to revise old topic assignments in the light of new data, and thus correctly approximates the posterior of the entire dataset at the time of the update.

AlSumait et al. (2008) fit a separate model to each increment, but use the matrix of topic-word counts of the previous model as a parameter for the prior distribution of the topic-word

distributions φ of the current model. When fitting a model to data from the first increment, β can have an arbitrary value. In all following models, instead of a single parameter vector β for all φ_k, there is one vector β_k for each φ_k ($\beta \in \mathbb{R}^{K \times W}$), where $\beta_{k,i}$ is the number of times word i was assigned to topic k by the previous model. Each component of the parameter vector of a Dirichlet distribution must be strictly positive, so we deviate from the algorithm of AlSumait et al. by adding a small value (0.01) to each word count to deal with words that do not occur in the current increment. Label switching is unlikely due to the strength of the prior. Since the algorithm of AlSumait et al. is efficient and easy to implement, we use it for certain experiments in chapter 7 that require incremental updating.

All of the above algorithms assume that the size W of the vocabulary is known in advance, which only is the case when working retrospectively. A solution for the true online setting is provided by Zhai and Boyd-Graber (2013), who, however, operate in a framework of variational inference instead of Gibbs sampling. Much like the HDP turns the fixed number of topics K into a parameter that is estimated from the data, their non-parametric model treats the vocabulary size W as a parameter. This comes at the cost of higher model complexity and computational complexity.

Optimizing Runtime and Memory Usage

The size of a corpus that can be processed within a reasonable time frame is bounded by the available computational resources. Even though the asymptotic complexity of LDA parameter estimation via Gibbs sampling is linear in the size of the corpus N_* (if the number of iterations and the number of topics K are fixed), performing Gibbs sampling on a large corpus can be highly time consuming in practice (Newman et al., 2009a; Smola and Narayanamurthy, 2010). Corpus size is also limited by the available memory: The amount of memory required for parameter estimation depends on the overall number of words (N_*), and the size of the matrices of document-topic counts ($M \times K$) and topic-word counts ($K \times W$). All computations described in this work were carried out on a 6-core multiprocessor workstation with 64 GB of RAM, so we limit our discussion to optimizations that are viable for this class of machine.

Via the Dirichlet priors placed on the document-topic distributions and the topic-word distributions, it is possible to induce sparsity, which manifests as a high amount of zero entries in the corresponding count matrices. This can be exploited by storing only the non-zero elements of these matrices in an associative array (usually implemented as a hash table), using a key that encodes the row and column index (Griffiths and Steyvers, 2004; Lu et al., 2013). The vector of word frequencies in a document is naturally sparse. Due to the bag-of-words assumption of LDA, a document is fully characterized by its non-zero word frequencies, which can also be stored in an associative array, using the word's index in the vocabulary as the key. If the corpus is still too big to fit into memory as a whole, documents can be stored in a circular buffer of bounded size that is replenished with data read from persistent storage by a worker thread.

The Gibbs sampling procedure offers many opportunities for speed optimization. We start with specific improvements to the inner loop of the algorithm, which, in the case of LDA, repeatedly takes samples from a categorical distribution. Given a pseudo-random number

generator (PRNG) for floating-point numbers that are uniformly distributed over the interval $[0, 1)$, samples from arbitrary distributions can be obtained by inversion of their cumulative density function (CDF; Devroye, 1986, ch. II.2.1). Therefore, the performance of the Gibbs sampler directly benefits from a fast uniform PRNG. The `xorshift` family of generators is known to be fast and suitable for simulation purposes, but its generators do not pass BigCrush, the largest and most stringent set of tests offered by the TestU01 PRNG test suite, without systematic failures (Vigna, 2016). To address this issue, Vigna proposes an improvement of the method, called `xorshift*`, and claims that the `xorshift1024*` variant in particular passes BigCrush. We decided to adopt this variant as the PRNG for our implementation of the Gibbs sampling procedure based on the strength of this claim and the substantial speed improvement over comparable generators like the Mersenne Twister (The Apache Software Foundation, 2019, ch. 4.1).

However, subsequently to the conclusion of our topic modeling experiments, Lemire and O'Neill (2019) reported that `xorshift1024*` systematically fails some of the BigCrush tests if its output is transformed in a specific way. Each invocation of the generator yields a random 64-bit integer, while the test suite expects 32-bit values. Test failures appear if, instead of successively feeding the upper and lower 32 bits to the test suite as separate values, the 64-bit output is reduced to 32 bits by discarding the most significant bits, followed by a reversal of the bit order. This indicates weaker randomness of the least significant bits. While Lemire and O'Neill are correct in noting that their test results point towards a flaw in the design of the PRNG, we do not expect this to be a problem in our use case: Our chosen implementation, which is part of the Apache Commons RNG library, version 1.0, follows the reference implementation of Vigna in constructing IEEE 754 double precision floating-point numbers from the upper 52 bits of the 64-bit output of the PRNG. Since we exclusively generate floating-point numbers, we never end up using the least significant 32 bits on their own, and the most unreliable bits are always discarded. Still, judging from the results of tests conducted by the Commons RNG authors (The Apache Software Foundation, 2019, ch. 4-5), the SpitMix64 generator (Steele Jr. et al., 2014), among others, appears to be a more suitable choice for future experiments.

Sampling from an arbitrary discrete distribution $p(x)$ by inversion of its CDF $F(x)$ means generating a uniform random value $u \in [0, 1)$ and finding the smallest x for which $u < F(x)$. In the case of Gibbs sampling, $p(x)$ is usually only known up to proportionality, so the sum of $q(x) \propto p(x)$ over all possible values of x needs to be computed first, in order to obtain the normalization constant. At that point, the x for a given u can be efficiently found using binary search (Devroye, 1986, ch. III.2.3). When sampling a topic assignment z within an LDA model, further optimization is possible. The FastLDA algorithm (Porteous et al., 2008) operates under the assumption that, when sampling the topic assignment of a word, the probability mass is mostly concentrated on a small subset of topics due to sparsity. Before sampling the topic assignments of a particular document, the topics are heuristically sorted, so that the topic probabilities that are likely to be highest in that document are computed first. Along with the unnormalized conditional probability of the k-th topic $q(k) \propto p(z_{m,n} = k \mid z_{-(m,n)}, y, \alpha, \beta)$, an upper bound Z_k on the true normalization constant Z is computed, so that $Z_{k+1} \leq Z_k$ and, ultimately, $Z_K = Z$. If $u \leq \frac{1}{Z_k} \sum_{1 \leq i \leq k} q(i)$, it is possible to select one of the topics $i \leq k$ without having to compute $q(i)$ for the remaining topics $i > k$. In the

course of our experimental work, verifying that $Z_K = Z$ after computing $q(K)$ turned out to be a valuable diagnostic for an inconsistent sampling state caused by implementation errors elsewhere in the parameter estimation process. SparseLDA (Yao et al., 2009) improves upon FastLDA by providing an efficient way of computing the normalization constant Z directly. Yao et al. claim that SparseLDA is approximately twice as fast as FastLDA, but we do not follow up on this claim.

The Approximate Distributed LDA (AD-LDA) algorithm (Newman et al., 2009a) is a distributed variant of the LDA parameter estimation procedure that is suitable for use on a shared-memory multiprocessor system. The M documents of the corpus are distributed approximately uniformly across P threads of execution. Each thread only keeps track of its own document-topic counts, but keeps a local copy of the entire topic-word count matrix. Using only its own count matrices, each thread performs one iteration of Gibbs sampling on its assigned documents. The threads wait for each other to finish, and then one designated thread merges the P local topic-word matrices into a single matrix that is again distributed to all other threads. This process is repeated until the requested number of Gibbs iterations have been performed. Considering that each thread is only aware of its own updates to the topic-word counts until the global synchronization happens, this can only be an approximation of the original Gibbs sampling process, where an update to the topic-word counts has an immediate effect on the probability of all following topic assignments. Newman et al. provide no formal convergence guarantees, but demonstrate on different corpora that there is no substantial difference in predictive performance between models produced by AD-LDA parameter estimation and the original, non-distributed procedure.

Since the matrix of topic-word counts is the largest part of the Gibbs sampling state, the need to keep P copies of this matrix implies a substantially increased memory consumption. This is particularly important during the initial iterations of Gibbs sampling, when the sampler has not yet fully converged to the posterior distribution. The multinomial parameters, and therefore also the associated count matrices, have not yet reached their final degree of sparsity, and their effective size is proportionally larger. If the available memory is not sufficient for the worst case of P fully dense topic-word matrices, we recommend to start with a single thread, monitor the density of the matrices, and successively redistribute work to additional threads as the density decreases. This is particularly effective if each thread does not receive a full copy of the topic-word matrix, but a reference to a read-only version that is shared among all threads, and updates to that matrix are performed in a copy-on-write manner.

AD-LDA is easy to implement, but its approximative nature, the need for synchronization after each iteration, and its high memory requirements present potential for improvement. Smola and Narayanamurthy (2010) introduce a more sophisticated alternative that addresses most of these weaknesses and scales to compute clusters where memory is no longer shared by all processors.

Our optimization efforts can be summarized as follows: To reduce the memory footprint of parameter estimation, the sparsity of the count matrices is exploited by using a data structure that does not explicitly store zero elements, and corpus data is loaded into a circular buffer of fixed size on demand. Runtime is improved by using the FastLDA algorithm to sample the topic assignments, CDF inversion by binary search for sampling from all other discrete

distributions, and the `xorshift1024*` PRNG for generating uniform random values in all sampling tasks. The resulting algorithm is distributed across multiple processors using a parallelization scheme based on AD-LDA.

4.1.2 Evaluation

When applying topic models to the analysis of text documents, we are faced with a number of tasks that involve assessing the quality of the model: Before any data analysis can take place, the parameter estimation process and its implementation have to be tested for internal consistency, that is, the ability to recover the true parameters from data. After fitting a topic model to a corpus of documents, one is usually interested in evaluating the quality of the topics, or comparing two or more distinct models.

Topic models are generative in nature, so testing an implementation for internal consistency is conceptually simple. Parameter estimation can be viewed as reversing the model's data-generating process to obtain parameters from observational data. Conversely, one can follow the generative process to obtain synthetic data from a model with known parameters: start from a set of reasonable hyper-parameters α, β, generate parameters θ, φ by random sampling from their prior distributions, and then randomly generate topic assignments z and finally observations y. Fitting a model to that synthetic dataset should yield parameter estimates that are close to the known, true parameters. In practice, choosing an appropriate threshold for the estimation error is not always trivial, particularly if the computational budget for the test is low, so that the amount of synthetic data and the number of Gibbs sampling iterations are limited. Due to label switching, the indices of the estimated topics are highly likely to be a permutation of the reference topic indices, so it is necessary to use a similarity measure that is invariant to permutation or otherwise mitigate the effect of label switching, e.g., by greedy matching of topics.

LDA can be understood as an implicit factorization of the document-word co-occurrence matrix (Steyvers and Griffiths, 2007), and a factorization with a lower reconstruction error manifests as a model with higher likelihood. The ratio of two likelihoods tells us about goodness-of-fit, that is, which one of two models is more likely to have generated a particular corpus. Topics are meaningful to the human analyst, because they uncover latent semantic information that is present in the co-occurrence statistics. One would expect that a better fit to the data implies more interpretable topics, but testing this hypothesis requires a way to measure interpretability.

Human evaluation of topic quality can be done *ad hoc* by inspecting the top-n most probable words, but this is time consuming and entirely based on subjective judgement. Chang et al. (2009b) propose the word intrusion test, which is designed to formalize this process and reduce the effect of individual bias. A human judge is presented with a fixed number of highly likely words from a particular topic, and a single highly unlikely word (the "intruder"), in random order. The task is to identify the intruder, which is easy if the topic exhibits a high degree of semantic coherence. Chang et al. compare models with different numbers of topics, fit to the same data, and find that with an increasing number of topics, "models are often trading improved likelihood for lower interpretability." The non-linear nature of the relationship between model likelihood and topic interpretability suggests that

there are two independent aspects of topic model quality: The model's fit and generalization ability are represented by the likelihood of training data and held-out test data, respectively. The interpretability of a topic can be assessed manually or by means of metrics that have been shown to be good proxies for the human assessment. For example, the pointwise mutual information (PMI) score (Newman et al., 2009b) can be viewed as an attempt to eliminate human judgement from the word intrusion test by drawing upon co-occurrence statistics from a large external corpus like Wikipedia. For each possible word pair, the probability of the two words being generated independently by the topic under consideration (i.e., the product of the topic-word probabilities) is compared to the relative frequency of co-occurrence in the reference corpus.

Goodness-of-fit, generalization, and topic interpretability are intrinsic qualities of a topic model. One could also carry out an extrinsic evaluation, which means testing the utility of the model as a part of a larger system that is designed to perform a particular task.

All implementations of topic models that are used in experiments in this work are tested for internal consistency as described earlier. Model comparisons are based on the likelihood of held-out data. From chapter 5 onwards, topic models are employed as a central component of two larger systems aimed at the detection of social influence in online communication. The experiments conducted in these chapters extrinsically confirm the utility of topic modeling, but cannot be counted as a true extrinsic evaluation, since there is no comparison to a system that performs the same task without making use of topic models.

Estimating the Likelihood

The likelihood function of the LDA model is given in equation 4.9 (Heinrich, 2009). However, one is usually not interested in the likelihood of a model as an absolute value, but wants to compare two or more models that differ in some aspects. For example, one may want to compare the likelihood of the model, given the training data, before and after I iterations of Gibbs sampling. Although Gibbs sampling does not explicitly maximize the likelihood, one would expect the likelihood to increase as the sampler moves from a random initial assignment to a high-probability region of the parameter space, so comparing the likelihood at different iterations of sampling can serve as a diagnostic. To make a comparison meaningful, it may be necessary to compare marginal likelihoods: If the effect of a variable on the likelihood is considered irrelevant to the comparison, it is marginalized out. In the case of the training set likelihood of LDA, the vector of latent topic assignments z is commonly thought of as a nuisance parameter and marginalized out, as in equation 4.10 (Heinrich, 2009).

Evaluating the model's ability to generalize involves the computation of its likelihood given unseen test data, which is usually a held-out portion of the training data. Given a topic model, represented by its parameters φ learned from a training corpus y', and its fixed hyper-parameters α, β, one wants to determine the likelihood of a test corpus y being generated by that model. This requires marginalizing out z and the document-topic distributions θ, as in equation 4.11 (adapted from Wallach et al., 2009b; Griffiths and Steyvers, 2004). Finally, marginalizing out all variables (equation 4.12; Griffiths and Steyvers, 2004) yields an expression that is equivalent to the denominator of the posterior (compare equations 4.2 and 4.3). This "fully marginalized" likelihood can be used for model selection, i.e., the compari-

son of models with different hyper-parameters and structure, such as different values of K (Griffiths and Steyvers, 2004).

$$p(y|\theta, \varphi, z, \alpha, \beta) = p(y|\varphi, z) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} \varphi_{z_{m,n}, y_{m,n}} \tag{4.9}$$

$$p(y|\theta, \varphi, \alpha, \beta) = p(y|\theta, \varphi) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} \sum_{k=1}^{K} \theta_{m,k} \cdot \varphi_{k, y_{m,n}} \tag{4.10}$$

$$\begin{aligned} p(y|\varphi, \alpha, \beta) = p(y|\varphi, \alpha) = &\left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \right)^M \cdot \\ &\prod_{m=1}^{M} \sum_{z_m} \left(\frac{\prod_{k=1}^{K} \Gamma(c_{k,m,*} + \alpha_k)}{\Gamma(N_m + \sum_{k=1}^{K} \alpha_k)} \cdot \prod_{n=1}^{N_m} \varphi_{z_{m,n}, y_{m,n}} \right) \end{aligned} \tag{4.11}$$

$$\begin{aligned} p(y|\alpha, \beta) = &\left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \right)^M \cdot \left(\frac{\Gamma(\sum_{w=1}^{W} \beta_w)}{\prod_{w=1}^{W} \Gamma(\beta_w)} \right)^K \cdot \\ &\sum_{z} \left(\prod_{m=1}^{M} \frac{\prod_{k=1}^{K} \Gamma(c_{k,m,*} + \alpha_k)}{\Gamma(N_m + \sum_{k=1}^{K} \alpha_k)} \cdot \prod_{k=1}^{K} \frac{\prod_{w=1}^{W} \Gamma(c_{k,*,w} + \beta_w)}{\Gamma(c_{k,*,*} + \sum_{w=1}^{W} \beta_w)} \right) \end{aligned} \tag{4.12}$$

All of these marginal likelihoods have a closed-form expression, but with each variable that is marginalized out, fewer conditional independence assumptions (see figure 4.1) apply, and the computational complexity of its evaluation increases. Recall that N_m is the number of words in the m-th document, and $N_* = \sum_{m=1}^{M} N_m$ the overall number of words in the corpus. The non-marginal likelihood (equation 4.9) can be computed in $O(N_*)$ time. Marginalizing out the latent topic assignments z (equation 4.10) increases the complexity to $O(N_* K)$. If the document-topic distributions θ are marginalized out as well, computation of the marginal likelihood (equation 4.11) requires iterating over all possible combinations of topic assignments for each document, separately, which has exponential complexity $O(K^{\max_{m=1}^{M} N_m})$. If all variables are marginalized out (equation 4.12), it is necessary to iterate over all combinations of topic assignments for the entire corpus, with a complexity of $O(K^{N_*})$.

Given these marginal likelihood functions, there are two possible approaches to model comparison: Comparing the full marginal likelihood (equation 4.12) of two models is what Nicenboim et al. (2021) call the prior predictive perspective on model comparison. The full marginal likelihood represents the "support that the data give to the model", and particularly the assumptions that are encoded into the prior distributions (Nicenboim et al., 2021). Conversely, the posterior predictive perspective is comparing the held-out likelihood, a measure of generalization. Since we are more interested in selecting the model with the best predictive performance than in validating our choice of priors, we choose the posterior predictive approach. The remainder of this section is concerned with the efficient computation of the held-out likelihood.

Since the evaluation of equation 4.11 is intractable, a suitable approximation needs to be

found. Due to the mutual conditional independence of the topic assignments z_m of documents $m \in \{1 \dots M\}$ given the topics φ, the likelihood can be factorized: $p(y|\varphi, \alpha, \beta) = \prod_{m=1}^{M} p(y_m|\varphi, \alpha)$. A number of methods for estimating this per-document held-out likelihood can be found in Wallach et al.'s comprehensive review (2009b). The most basic method is due to Newton and Raftery (1994, sect. 7): After expressing the marginal likelihood as an expectation (equation 4.13), an MC estimate (equation 4.14) can be obtained by sampling θ from its prior. Since sampling from a Dirichlet distribution is computationally expensive, this method is too slow for practical use. By application of Importance Sampling (IS), one can use samples from the (collapsed) posterior instead, which can be more cheaply obtained via Gibbs sampling. Using the posterior as the proposal distribution in an importance sampling scheme yields equation 4.15 (Newton and Raftery, 1994), where HM refers to the harmonic mean. Consequently, this way of estimating the held-out likelihood is known as the harmonic mean method. While its individual iterations are fast, it has been found to require an impracticably high overall number of iterations for convergence (Wallach et al., 2009b).

$$p(y_m|\varphi, \alpha) = \int p(\theta_m|\alpha) \cdot \sum_{z_m} p(y_m|z_m, \varphi) \cdot p(z_m|\theta_m)\, d\theta_m \tag{4.13}$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} p(y_m|\theta_m^{(s)}, \varphi) = \frac{1}{S} \sum_{s=1}^{S} \prod_{n=1}^{N_m} \sum_{k=1}^{K} \theta_{m,k}^{(s)} \cdot \varphi_{k,y_{m,n}} \tag{4.14}$$

$$\text{with } \theta_m^{(s)} \sim \mathrm{Dir}(\alpha)$$

or

$$\approx \mathrm{HM}\left(\left\{p(y_m|z_m^{(s)}, \varphi)\right\}_{s=1}^{S}\right) = \mathrm{HM}\left(\left\{\prod_{n=1}^{N_m} \varphi_{z_{m,n}^{(s)}, y_{m,n}}\right\}_{s=1}^{S}\right) \tag{4.15}$$

$$\text{with } z^{(s)} \sim p(z|y, \varphi, \alpha, \beta)$$

Since these two naive methods are not usable in practice, Wallach et al. (2009b) propose the "left-to-right" (LR) algorithm, which is situated in the framework of Sequential Monte Carlo (SMC). Its mathematical properties are discussed in-depth by Scott and Baldridge (2013). Here, we only address certain technical details that pertain to its efficient implementation. The basic principle of operation is that the per-document marginal likelihood is further factorized into the conditional probabilities of individual words $y_{m,n}$ given all previous words $y_{m,<n}$ (equation 4.16). This is equivalent to an online learning setting where the document arrives one word at a time. Evaluating this probability involves a sum over all topic assignments $z_{m,\leq n}$ (equation 4.17), which is approximated using a Monte Carlo algorithm (Wallach et al., 2009b).

$$p(y_m|\varphi, \alpha) = \prod_{n=1}^{N_m} p(y_{m,n}|y_{m,<n}, \varphi, \alpha) \tag{4.16}$$

$$= \prod_{n=1}^{N_m} \sum_{z_{m,\leq n}} p(y_{m,n}, z_{m,\leq n}|y_{m,<n}, \varphi, \alpha) \tag{4.17}$$

$$\approx \prod_{n=1}^{N_m} \frac{1}{R} \sum_{r=1}^{R} \sum_{k=1}^{K} p(y_{m,n}, z_{m,n}^{(r)} = k | z_{m,<n}^{(r)}, \varphi, \alpha) \tag{4.18}$$

(sampling scheme for $z_{m,<n}^{(r)}$ described in text)

The process of approximation is similar to Gibbs sampling, but instead of iterating over the document multiple times, only a single pass is performed. For each word index n, a topic assignment $z_{m,n}$ is sampled from $p(z_{m,n}|z_{m,<n}, y_{m,n}, \varphi, \alpha)$. This is equivalent to sampling from the full conditional with fixed φ (equation 4.8), if the sampling state (i.e., the count matrices derived from the topic assignments) is constructed so that it only covers the words $n' < n$ that have already been assigned a topic. Before updating the sampling state with the new topic assignment, θ_m is estimated from the current state, and the likelihood with respect to the n-th word (innermost sum of equation 4.18) is computed as $\sum_{k=1}^{K} \theta_{m,k} \cdot \varphi_{k,y_{m,n}}$. To account for the variability of the posterior, this sampling process is performed R times in parallel, with each sampler ("particle") operating on its own set of topic assignments $z^{(r)}$. The final likelihood estimate is the product of the arithmetic means of the per-word likelihood estimates of the particles (equation 4.18).

As it stands, this process does not produce an accurate approximation of the likelihood. The topic assignments are drawn from a distribution that differs from the posterior, so the estimate of θ_m is biased. If, however, before sampling the topic assignment of the n-th word, all previous assignments are resampled using the current sampling state, the distribution of topic assignments converges to the posterior (Wallach et al., 2009b; Canini et al., 2009). Due to resampling, the overall number of samples that have to be drawn is quadratic in the length of the document. Scott and Baldridge (2013) caution against simply omitting the resampling, pointing out that doing so will seriously impair the accuracy of estimation. The error that is introduced by skipping the resampling step is proportional to the document length, which is a particular problem for our experiment that tests the effect of document length on held-out likelihood (section 4.2.2).

Depending on the application, resampling may be unacceptably expensive. Canini et al. (2009) apply SMC techniques to the online training of LDA in a way that is conceptually similar to the LR algorithm. They reduce the complexity of the resampling step by limiting the resampling to a random subset of words. If the size of this subset is proportional to the logarithm of the document length, the runtime is log-linear instead of quadratic, and the correlation between error and document length is reduced. We therefore adopt the following strategy: Before sampling the topic assignment of the n-th word, $\frac{n}{N_m} \cdot \ln(N_m)$ earlier words (at least one) are resampled. These words are chosen by sampling from the set of word indices uniformly and with replacement. Sampling without replacement would ensure that no word is unnecessarily resampled more than once, but is more computationally expensive than sampling with replacement, and was therefore found to reduce overall efficiency.

The perplexity of the model, defined as $2^{-\frac{1}{N_*} \log_2(\ell)}$, normalizes the likelihood ℓ with respect to the corpus size in words N_*. This is useful for comparing the likelihood of a model with respect to corpora of different sizes, or simply as a convention to indicate that the size of the corpus is irrelevant.

Comparing Topic Distributions

Any application of topic modeling is likely to involve, at some point, the comparison of topic distributions. Viewing a topic distribution (either document-topic or topic-word) as the parameter vector of a categorical distribution, a number of similarity measures apply: In the context of vector space models of text, the similarity of two vectors is commonly expressed in terms of their cosine similarity (equation 4.19). However, since cosine similarity is based on the concept of Euclidean distance, it is not appropriate for vectors on the unit simplex (Manning and Schütze, 1999, pp. 303). Note that this is a property of probabilistic topic models, not topic models in general, which stems from the modeling decision to express documents and topics as categorical distributions. LSI, and also certain Bayesian models like the spherical topic model (Reisinger et al., 2010), embed documents and topics in a vector space where the cosine distance is applicable.

In information theory, the Kullback-Leibler divergence (KL, equation 4.21) is the canonical measure of dissimilarity of probability distributions. The divergence $\mathrm{KL}(p \parallel q)$ can be interpreted as the amount of information that is lost when using q to approximate the "true" distribution p. It has two properties that render it unsuitable for some use cases: First, it takes on values in the range $[0, \infty)$, with $\mathrm{KL}(p \parallel p) = 0$. Since it is not bounded above, it cannot be easily transformed into a measure of similarity. Second, it is asymmetric, i.e., there are p, q for which $\mathrm{KL}(p \parallel q) \neq \mathrm{KL}(q \parallel p)$. This makes the KL divergence hard to interpret in cases where one is not comparing a prediction or other kind of approximation to a known-good reference.

Therefore, in the context of topic modeling, symmetrized variants of the KL divergence are preferred (Steyvers and Griffiths, 2007). The Jensen-Shannon divergence (JSD, equation 4.20) has a convenient range of $[0, 1]$ when using the base 2 logarithm. The divergence $\mathrm{JSD}(p \parallel q)$ has an interpretation in terms of mutual information, i.e., the information gained about one variable by observing the other. If m is interpreted as a mixture of p and q with equal weight given to each component, then $\mathrm{JSD}(p \parallel q)$ is equivalent to the mutual information of a random variable distributed according to m and the binary indicator variable that selects the mixture component. Informally speaking, if $p = q$, random draws from m contain no information about which component was chosen, while in the case of dissimilar p and q, each draw is highly informative. It should be noted that, according to equation 4.21, $\mathrm{KL}(p \parallel q)$ is only defined if all components of p, q are non-zero. Pairs p, q, where $q_i = 0 \implies p_i = 0$ for all i, are by convention handled by summing only over components i for which $p_i \neq 0$. With this extension to the definition of $\mathrm{KL}(p \parallel q)$, $\mathrm{JSD}(p \parallel q)$ is defined for arbitrary non-negative vectors on the unit simplex.

$$sim_{\cos}(p, q) = \frac{\sum_i p_i q_i}{\sqrt{\sum_i p_i^2}\sqrt{\sum_i q_i^2}} \tag{4.19}$$

$$\mathrm{JSD}(p \parallel q) = \frac{1}{2}\mathrm{KL}(p \parallel m) + \frac{1}{2}\mathrm{KL}(q \parallel m) \text{ with } m = \frac{1}{2}(p + q) \tag{4.20}$$

$$\mathrm{KL}(p \parallel q) = -\sum_i p_i \log\left(\frac{p_i}{q_i}\right) \tag{4.21}$$

$$\mathrm{BC}(p, q) = \sum_i \sqrt{p_i q_i} \tag{4.22}$$

The Bhattacharyya coefficient $\mathrm{BC}(p, q)$, given in equation 4.22, provides an alternative representation of the similarity of two probability distributions. Neither the JSD nor the Bhattacharyya coefficient is a metric, as both violate the triangle inequality, but a metric can be derived from either by taking the square root: $\sqrt{1 - \mathrm{BC}(p, q)}$ is equivalent to the Hellinger distance, while $\sqrt{\mathrm{JSD}(p \parallel q)}$ is known as the Jensen-Shannon distance or metric (Endres and Schindelin, 2003). The Hellinger distance has a geometric interpretation as the Euclidean distance of two points after projection from the unit simplex Δ^n onto the surface of an n-sphere. Nielsen et al. (2010) identify a relationship between the Bhattacharyya distance $-\ln(\mathrm{BC}(p, q))$ and the JSD when comparing distributions that belong to the same exponential family. Own preliminary experiments, as well as experiments in the literature (Cha, 2007), show a strong (but not perfect) linear correlation between the two similarity measures, which makes us assume that the choice of measure is unlikely to have a noticeable impact on experimental results. For use in the following experiments we prefer the JSD, because of its prevalence in the literature on topic modeling and topical social influence, see for example TwitterRank (Weng et al., 2010). Since we do not require a true metric, we use its regular, non-metric form.

4.2 Application to Online Communication Data

The distinct nature of online communication, compared to traditional modes of written communication, presents several challenges to the application of topic modeling. We focus on what we consider to be the two main issues: First, online communication is characterized by use of informal language, slang, spelling errors, language mixing and code switching, and medium-specific social conventions (see, e.g., section 3.1.2). The resulting increase in the number and variety of words that are associated with a particular concept reduces the data efficiency of topic modeling, which can be counteracted by preprocessing steps that filter out irrelevant words or normalize spelling variants. Second, online communication typically takes the shape of short fragments of text, which are often close to meaningless on their own, and have to be understood in the context of an ongoing conversation. As a general rule, the shorter the average document, the lower the amount of co-occurrence information that can be obtained from the corpus. We investigate how this problem can be addressed by aggregation of messages into larger units.

4.2.1 Linguistic Preprocessing

A series of surface-level linguistic preprocessing steps is performed with the goal of making the textual content of the collected social media messages more amenable to topic modeling. Denny and Spirling (2017) point out that the choice of preprocessing steps can have a strong effect on the parameters of the topic model, so care must be taken that the preprocessing strategy is aligned with the goals of the subsequent analysis.

The first step is detecting which language is being used. This happens at the time of crawling or data cleaning, depending on the source of the data. Language detection is a prerequisite for all following processing steps, which are either language-specific or operate under the assumption that all input text is written in a single language. Topic modeling is not language-aware, but a typical multi-lingual corpus is unlikely to contain sufficient information to enable the model to gather semantically related terms from different languages in one topic: Usually, each document is written in a single language, so that co-occurrence of words from different languages is limited to rare exceptions like loan words and direct quotations. In consequence, topics will be separated by language, even if they are highly similar semantically. This makes it difficult to choose a suitable overall number of topics in advance. In social media text, language mixing is more common, but still unlikely to produce truly multi-lingual topics.

Crawling can be effectively limited to English-speaking users of a platform, while the e-mail datasets are known to contain mainly English language messages (Enron) or a mixture of English and Italian (HackingTeam). Therefore, we provide the necessary linguistic resources for processing text in these two languages, and attempt to filter out text written in any other language. Language detection is performed using the Java library `language-detection` (Nakatani, 2010b), which is built around a Naive Bayes classifier trained on character n-grams. Nakatani (2010a) evaluates the classifier on two separate corpora of Wikipedia pages and news articles and reports an average classification accuracy of 99.9%. Since the classification is based on character frequency statistics, it is reasonable to assume that a certain minimum amount of text is required to attain the reported accuracy.

Blodgett et al. (2017) evaluate `langid.py`, which, like `language-detection`, uses a Naive Bayes n-gram classifier internally. They apply the classifier to tweets of varying length and find that accuracy is positively correlated with tweet length, but even the best-case accuracy remains below the figures reported by Nakatani. Individual messages on social platforms generally tend to be short, so in order to increase classification accuracy, we aggregate messages by author. In consequence, it is no longer possible to detect people who use different languages to address different audiences, or even use multiple languages in a single sentence ("code-switching") (Blodgett et al., 2017). The most likely failure mode is that the aggregated messages of people who use a non-English language, but do so rarely, will be detected as English, and traces of other languages will remain in the filtered dataset. In consequence, subsequent processing needs to be robust towards the presence of small amounts of non-English language text. The remaining foreign language words mostly co-occur with each other and are therefore likely to end up in a common topic.

Further linguistic analysis requires dividing the textual content of a message into a sequence of words and word-like entities, a process known as *tokenization*. This process transforms a sequence of characters into a sequence of tokens, which are members of a finite vocabulary. Tokenization is easy for languages where word boundaries are indicated by whitespace and punctuation, and therefore tokenization can be performed by linearly scanning the text. For example, according to the Penn Treebank tokenization rules for English (The Penn Treebank Project, 1999), words are delimited by whitespace and each punctuation character is a separate token. These rules were originally formulated for the processing of error-free standard English text, an assumption that is frequently violated on social plat-

forms, where the conventions of language use are much closer to those of spoken language than to those of more traditional forms of writing, such as newspaper articles or personal letters. The register of a typical social media post is informal, spelling and grammar errors are tolerated, and some aspects of non-verbal communication are recreated by textual means (Gimpel et al., 2011; Thelwall, 2009).

To handle the non-standard style of writing and punctuation found in social media text, we use a modified version of the TweetMotif tokenizer (O'Connor et al., 2010). The tokenizer assumes that a word is delimited by at least one whitespace or punctuation character on both sides. First, the tokenizer identifies all substrings that should be treated as a single token despite violating this assumption: numeric and time values (may contain decimal points, thousands separators and colons), URLs, abbreviations, emoticons, emoji, and words that contain apostrophes or dashes. The remaining text is split into tokens at the boundaries indicated by whitespace and punctuation, including symbol characters used for decorative purposes. The boundary characters are swallowed, i.e., they are neither part of the word token, nor do they appear as a separate token. Numbers (but not time or date values) and single-character tokens are removed. Finally, all non-URL tokens are normalized by transformation to lower case.

Textual messages on the Internet have long contained pictorial expressions of emotion, which can be understood as a side-channel, embedded into the text, that enables a limited form of non-verbal communication. An *emoticon* is a sequence of letters, numbers, and/or punctuation characters that looks like a facial expression. For example, the emoticon `:)` resembles a sideways smiling face and is commonly used as a marker for statements with humorous intent, which should not be taken at face value. The most frequently used types of emoticon are recognized by the tokenizer and output as a single token. *Emoji*, a set of small pictographs that includes expressions of emotion, became part of the Unicode standard in 2009 (Davis and Edberg, 2019), and quickly became popular among users of instant messenger applications and social media. Since its introduction, the set of emoji has been extended several times. The HackingTeam and Facebook datasets are sufficiently recent to contain emoji, but only the latter is expected to contain a substantial amount. The Twitter crawler does not support the complete range of Unicode characters, and therefore does not preserve any emoji that may have originally been present in the collected tweets. Facebook supports emoji, but also has a separate, proprietary system for graphical expressions of emotion that predates the standardization of emoji. For most of these pictographs, an equivalent emoji exists (Päper, 2017). We represent the remaining ones by unique textual tokens.

Proper tokenization of emoji is difficult. The goal is to treat emoji like emoticons, and output each pictograph as a separate token. The core of the Unicode standard is a list of characters, each of which is assigned a unique numeric identifier, the *code point*. Usually, one code point corresponds to one visible character. For certain typographic elements, such as diacritics, which can be combined with many different characters, this approach would lead to a combinatorial explosion of the number of required code points. For that reason, most typographical elements that can be said to augment or modify the appearance of other characters receive their own code points. A sequence of code points that consists of one base character and one or more modifiers is displayed as a single visible character. This principle has been applied extensively in the mapping of emoji to code points. For example,

emoji that depict human beings can have modifiers that specify gender or skin tone. Each sequence of code points that results in a single visible emoji should be a single token.

UTS #51, an extension to the Unicode standard, provides a list of recommended code point sequences to produce each emoji (Davis and Edberg, 2019). To avoid having to reproduce the complex state machine that is necessary to recognize any valid code point sequence that encodes a given emoji, we only accept the recommended sequences. These sequences are loaded from two files that are distributed as supplementary material to UTS #51, `emoji-sequences.txt` and `emoji-zwj-sequences.txt`, and stored in a prefix trie. The input text is scanned for continuous sequences of maximal size that match complete trie entries, which are then output as tokens.

Words with a predominantly grammatical role, for example, the determiners "the" and "a", are known as *stop words* or *function words* in an NLP context (Manning and Schütze, 1999, pp. 20). While the presence or absence of a function word in a particular place can have a strong effect on the semantics of a sentence as a whole (negators like "not" being the most obvious example), taken individually and out of context, a function word carries little to no semantic information. Topic models operate under the "bag-of-words" assumption: the meaning of a unit of text (e.g., a document or sentence) is composed of the meaning of its constituent words, and the order of words within that unit is irrelevant and can be ignored. Under this assumption, function words lose their semantic value, so removing them from the text reduces the memory and computational requirements of topic modeling and is expected to improve the interpretability of topics. Due to its high frequency, a function word is likely to occur in almost every document, and therefore co-occurs with almost every word, but its most frequent co-occurrences will be with other high-frequency words. Therefore, all function words tend to be concentrated in a small number of topics, and removing such words will reduce the amount of non-interpretable topics.

We use the stop word list distributed with the KEA system (Jones and Paynter, 2001) for English language text, and a list from the NLTK Stopword Corpus (NLTK, 2019) for Italian. For each dataset, an additional list of domain-specific stop words is manually created as follows: The words that occur in the corpus are enumerated in order of decreasing frequency. A human annotator reviews each word, starting from the most frequent, until a word is encountered that evokes a specific conversational context in which it is likely to be used. All more frequent words are treated as stop words. Table 4.1 shows the number of manually identified stop words for each dataset. We hypothesize that it reflects the diversity of language use and subject matter.

Finally, a set of preprocessing steps addresses peculiarities of the different online social platforms. The use of URL shortening services is popular on social media, particularly on Twitter, where the number of characters per message is limited. Since 2010, Twitter automatically shortens URLs using its own service `t.co` (@SG, 2010). If the same URL is tweeted multiple times, a unique short URL is generated for each instance, so the shortened URLs contribute little information to a topic model. Therefore, all URLs are removed from tweets after tokenization, but are preserved in the other datasets. User names that are mentioned in a message only have indirect semantic value as references to the characteristics and behavior of these users, and are opaque to the human analyst. From the Twitter dataset, all @-mentions are removed, as well as the retweet indicators "RT" and "via", if they occur be-

Table 4.1: Number of manually identified stop words

Twitter	39
Twitter + websites	37
Facebook	100
Enron	10
HT-en	7
HT-it	7

fore an @-mention. Similarly, if users are mentioned in a Facebook post, their names are removed from the text. Hashtags are duplicated to give them extra weight in later processing steps, in lieu of a more principled term weighting scheme (Wilson and Chew, 2010). The leading '#' sign is removed.

In summary, we filter out non-English (or Italian, depending on the dataset) messages, perform tokenization, remove punctuation and numbers, convert all text to lower case, and remove function words. Emoticons and emoji are normalized, and some platform-specific processing is done. We deliberately skip some common preprocessing steps (cf. Denny and Spirling, 2017). Notably, we do not perform stemming and do not remove infrequent words. We consider stemming to be an overly aggressive normalization for our use case, since reducing a word to its stem obscures the semantic differences between different forms of a word, and common stemming algorithms would require adaptation to social media language use. Defining an appropriate threshold for the removal of rare words is hard, because unique spelling errors and variants cannot be easily distinguished from relevant, but rare vocabulary. Furthermore, we do not aggregate n-grams or specific classes of multi-word expressions into tokens, because we have no reason to believe that our datasets contain a substantial amount of specialized vocabulary that would benefit from such treatment.

The distribution of message length in the various datasets is summarized in table 4.2. Despite being subject to a much less restrictive size limit, the average Facebook post is shorter than the average tweet, and even the longest post stays well below the size limit imposed by the platform. The average e-mail is longer than the average social media post by a factor of 2.5 to 12, and e-mail is highly variable in the number of tokens per message. While social media posts are generally short, the e-mail datasets contain a mixture of short-form and longer communication. Message length appears to be a point of concern, but primarily when working with social media data.

4.2.2 Data Aggregation and Augmentation

It is generally accepted that the low word count of a typical social media post is detrimental to the quality of the topics produced by LDA. An intuitive explanation is that in short documents, each word only co-occurs with a few others. Therefore, the co-occurrence statistics do not contain enough information about the "fat tail" of rare, but semantically rich words, predicted by Zipf's law, which prevents them from being assigned to appropriate topics.

Table 4.2: Distribution of message length

	messages	tokens	tokens per message			
			median	mean	std.dev.	max.
Twitter	17 332 536	125 465 639	7.0	7.2	3.8	62
Facebook	1 962 836	8 044 232	2.0	4.1	6.4	511
Enron	37 398	661 132	8.0	17.7	48.3	3 233
HT-en	72 785	3 646 300	17.0	50.1	135.3	5 931
HT-it	124 017	2 941 347	17.0	23.7	40.2	10 151

Tang et al. (2014) provide a theoretical justification for this notion by expressing the statistical efficiency of the topic model, i.e., its ability to identify the true topics (under the assumption that the data was generated according to the LDA "generative story"), as a function of the volume of input data. They show that the statistical efficiency only benefits from additional documents if the average document length increases as well. One way of increasing the document length without drawing on external data sources is to aggregate the observed documents into larger *virtual documents*. Mehrotra et al. (2013) empirically demonstrate that systematic aggregation of tweets provides an advantage over treating each tweet as a single document. Aggregation by author is shown to outperform non-aggregated topic modeling in terms of cluster purity, while aggregation by hashtag also performs better according to information theoretic measures (PMI).

We define a virtual document as the concatenation of the content of its constituent documents, and call an *aggregation strategy* any surjective mapping of observed documents to virtual documents. The absence of aggregation corresponds to a bijective mapping. We hypothesize that not all effects of aggregation are beneficial: On one hand, aggregation creates longer documents that provide more co-occurrence information, which potentially enables a more accurate estimation of the topic-word distributions. On the other hand, merging topically unrelated documents introduces bias. To test the effect of aggregation on the quality of topics extracted from a corpus with a low average per-document word count, we conduct an experiment on the 30 000 user Twitter dataset.

To narrow down the search space of aggregation strategies, two assumptions are made about the nature of the data. Each tweet is associated with a sender-recipient pair (for modeling convenience, the sender of a non-addressive tweet acts as the recipient). First, we assume that two tweets are more likely to have a similar topic distribution if they are associated with the same sender-recipient pair. Second, we assume that temporally close tweets of a sender-recipient pair are likely to be part of a burst of activity, e.g., a conversation, and are therefore likely to have similar topics. Three families of aggregation strategies are derived from these assumptions. As a general rule, the set of tweets is first partitioned by sender-recipient pair, and then virtual documents are formed by applying the strategy to each subset, so that each virtual document only contains tweets of one user pair.

Fixed size (n) Each virtual document is made up from n temporally consecutive tweets,

with $n \in \{2, 5, 10\}$. An additional virtual document aggregates the remaining tweets, if any. Setting $n = 1$ is equivalent to no aggregation.

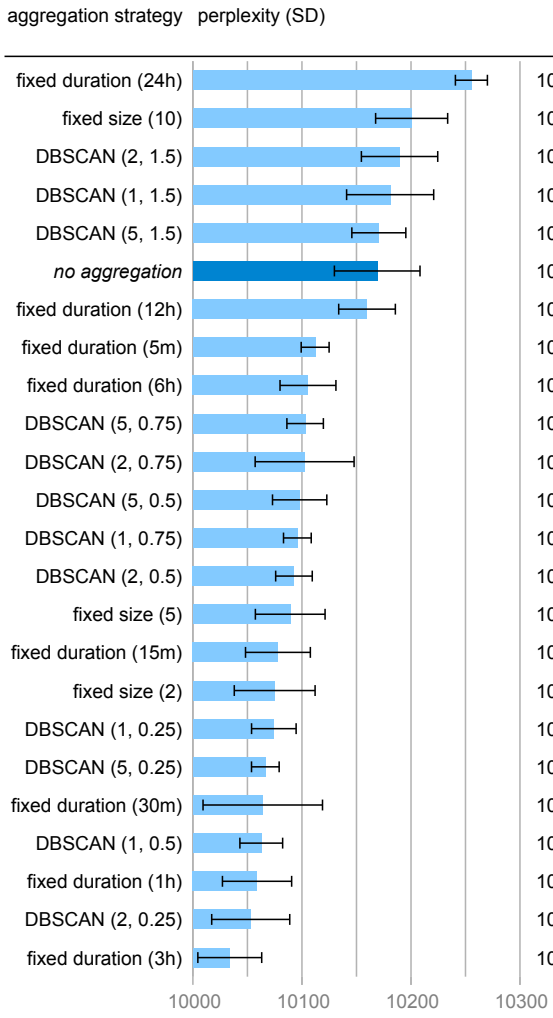
Fixed duration (l) The time line is partitioned into intervals of fixed length $l \in \{$ 5m, 15m, 30m, 1h, 3h, 6h, 12h, 24h $\}$. If an interval contains at least one tweet, its tweets are aggregated in a virtual document. A very short l is effectively equivalent to no aggregation.

DBSCAN ($MinPts$, f) DBSCAN is an algorithm for non-parametric, density-based clustering of data points in an arbitrary metric space (Ester et al., 1996). A data point is classified as a core point of a cluster if there are at least $MinPts - 1$ other points within a distance of Eps, as a boundary point if it is within distance Eps of at least one core point, and as an outlier, otherwise. A cluster is defined as the union of its core and boundary points. Each cluster contains at least $MinPts$ points. We apply DBSCAN to the timestamps of the tweets, using the ℓ_1-distance as the metric, and setting $Eps = f \cdot d$, where d is the average temporal distance between two successive tweets. The points of each cluster are aggregated in a virtual document, and each outlier forms a separate virtual document. To explore the parameter space, we test all combinations of $MinPts \in \{1, 2, 5\}$ and $f \in \{0.25, 0.5, 0.75, 1.5\}$.

Our evaluation of these aggregation strategies complements the similarly motivated study of Mehrotra et al. (2013), who find a positive effect of aggregation on various measures of topic quality that are related to human interpretability. We side-step the question of how to measure topic quality, and instead ask whether or not the statistical efficiency of the topic model, i.e., its ability to recover the topics from the data (Tang et al., 2014), benefits from aggregation. If aggregation leads to a more accurate estimate $\hat{\varphi}^{\text{agg}}$ of the topic-word probabilities φ, compared to the estimate $\hat{\varphi}$ from the same corpus without aggregation, then a model with topics $\hat{\varphi}^{\text{agg}}$ should also have a higher predictive likelihood (and lower perplexity) with respect to a corpus of non-aggregated, held-out data. Analogous to the theoretical analysis of Tang et al. (2014), the underlying assumption is that a true φ exists, because the held-out data was generated according to the generative process of LDA, which is a bold assumption for real-world observational data. In effect, the validity of the results is contingent on how well an LDA topic model fits the data.

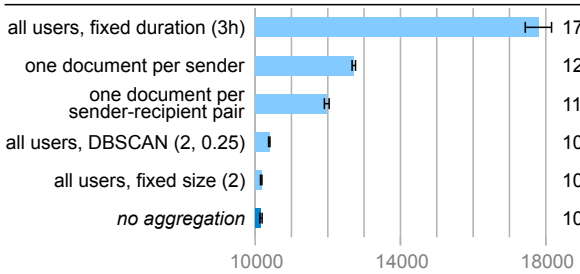
The tweets of a randomly selected 10% of users are set aside as test data. The hyper-parameters of LDA are chosen to match, as close as possible, the hyper-parameters used in later applications of topic modeling in this thesis (chapter 5 onwards): $K = 150, \alpha = \frac{50}{K}, \beta = 0.01$. However, only 200 iterations of Gibbs sampling are performed for the initial parameter fitting, to keep the overall computational effort manageable. The held-out likelihood is computed using the LR algorithm with 100 particles and the logarithmic resampling strategy. For each aggregation strategy, the process of parameter fitting and computation of the held-out likelihood is repeated five times, and the mean and standard deviation of the resulting perplexities are reported.

The results of this experiment are shown in figure 4.2. The variability of the perplexity estimates can be mainly attributed to the low number of Gibbs sampling iterations. Still, it is possible to establish an approximate ranking of the different strategies and their parameters.



aggregation strategy	perplexity (SD)	documents	tokens per document median	mean	std.dev.	max.
fixed duration (24h)	10 255.4 (14.7)	96 223	15	35.4	59.1	1 408
fixed size (10)	10 200.6 (33.1)	57 141	68	59.6	36.7	216
DBSCAN (2, 1.5)	10 189.5 (35.0)	94 501	18	36.1	54.9	2 193
DBSCAN (1, 1.5)	10 180.9 (39.9)	98 554	16	34.6	54.2	2 193
DBSCAN (5, 1.5)	10 170.5 (24.8)	157 215	9	21.7	43.0	1 525
no aggregation	10 169.0 (39.3)	430 478	8	7.9	3.4	37
fixed duration (12h)	10 159.7 (26.0)	127 539	13	26.7	41.1	1 159
fixed duration (5m)	10 112.1 (12.8)	379 969	8	9.0	5.3	295
fixed duration (6h)	10 105.5 (25.6)	159 618	12	21.4	30.2	1 078
DBSCAN (5, 0.75)	10 102.9 (16.7)	237 879	8	14.3	28.5	1 525
DBSCAN (2, 0.75)	10 102.4 (45.4)	141 242	13	24.1	37.0	1 525
DBSCAN (5, 0.5)	10 097.9 (24.9)	276 201	8	12.3	23.5	1 411
DBSCAN (1, 0.75)	10 095.8 (12.7)	145 642	12	23.4	36.7	1 525
DBSCAN (2, 0.5)	10 092.6 (16.8)	172 484	11	19.8	30.2	1 411
fixed size (5)	10 089.2 (32.0)	97 843	37	34.8	16.0	114
fixed duration (15m)	10 078.0 (29.7)	337 048	9	10.1	8.0	420
fixed size (2)	10 075.0 (37.1)	221 751	15	15.4	6.0	54
DBSCAN (1, 0.25)	10 074.2 (20.4)	229 277	10	14.9	22.1	1 196
DBSCAN (5, 0.25)	10 066.4 (12.7)	322 812	8	10.6	17.8	1 170
fixed duration (30m)	10 064.0 (54.8)	302 930	9	11.3	10.4	427
DBSCAN (1, 0.5)	10 062.7 (19.6)	177 013	11	19.3	30.0	1 411
fixed duration (1h)	10 058.8 (31.7)	263 119	10	13.0	14.0	602
DBSCAN (2, 0.25)	10 053.0 (35.8)	224 635	10	15.2	22.2	1 196
fixed duration (3h)	10 033.8 (29.3)	196 085	11	17.4	22.2	895

Figure 4.2: Comparison of aggregation strategies in terms of held-out perplexity and their effect on document length



aggregation strategy	perplexity (SD)	documents	tokens per document median	mean	std.dev.	max.
all users, fixed duration (3h)	17 795.6 (361.3)	480	7 070	7 100.9	4 457.9	23 708
one document per sender	12 716.3 (47.3)	2 116	884	1 610.8	2 279.4	29 833
one document per sender-recipient pair	11 971.3 (65.4)	18 183	8	187.5	906.2	29 833
all users, DBSCAN (2, 0.25)	10 389.8 (18.4)	76 807	38	44.4	32.7	368
all users, fixed size (2)	10 168.6 (20.2)	215 239	16	15.8	4.9	46
no aggregation	10 163.0 (29.5)	430 478	8	7.9	3.4	37

Figure 4.3: Ablation study, comparing the effects of the two components of an aggregation strategy in isolation

With the right choice of parameters, all three families of strategies lead to topics that explain the held-out data better than those generated by LDA without any aggregation. The "fixed size" strategy works best if a low number of tweets (2–5) is aggregated in a virtual document. Quantization into intervals of fixed duration performs best with intervals between 15 minutes and 3 hours, and distinctly worse with shorter (5 minutes) or longer (6 hours and above) intervals. The performance of clustering with DBSCAN appears to be largely insensitive to the choice of $MinPts$, but improves as f gets closer to zero.

Strong aggregation is characterized by a high number of virtual documents and a low number of tokens per document. Comparing the strength of aggregation to the improvement in perplexity, we find the results to be consistent with our initial hypothesis. The combinations of strategy and parameter values that produce the strongest aggregation also perform the worst. A possible explanation is that Twitter users communicate in short bursts, as shown in section 3.5.2, and these bursts are characterized by high topically consistency. To exploit this consistency, an aggregation strategy would need to create many small virtual documents. Fixed-size aggregation of five tweets per document is a notable outlier, performing well, despite being a simple strategy that aggregates strongly. This may be a hint at the existence of other sources of topical consistency that could be exploited to greater effect by a more sophisticated strategy.

The aggregation strategies that were tested in this experiment consist of two parts: first, the tweets are partitioned according to their sender-recipient pair, then a specific strategy is applied to the tweets of each pair. To better understand the contribution of each part to the improvement in perplexity, we perform an ablation study, where we test each part in isolation. Five partial aggregation strategies are compared. The strategy "one document per sender-recipient pair" aggregates the tweets of each sender-recipient pair into a single virtual document. The strategy "one document per sender" goes one step further and aggregates all tweets of a particular sender into one virtual document. By comparing these two strategies, we can test the assumption that the topics are negotiated between sender and recipient, rather than being mostly determined by the sender. These two partial strategies isolate the effect of partitioning, i.e., keeping the tweets of different senders or sender-recipient pairs separate. Conversely, one can isolate the effect of a specific aggregation strategy by skipping the partitioning and applying the strategy to the temporal sequence of tweets of all users. For each of the thee families of strategies defined earlier, we use the parameter values that performed best in the previous experiment.

The results of the ablation study are shown in figure 4.3. In isolation, neither of the two parts is able to outperform topic modeling without aggregation, which is strong evidence for a joint effect. There is a visible benefit of partitioning by sender-recipient pair rather than by sender alone, but we suspect that this effect is obscured to some extent by the high volume of non-addressive communication in the Twitter dataset. The disparity between addressive and non-addressive communication is reflected by the document length statistics of the "one document per sender-recipient pair" strategy: the low median indicates that a high number of sender-recipient pairs contributes little to the overall volume of communication, while the much higher mean points towards a small group of highly active non-addressive tweeters.

Augmentation with External Data

If the goal is to improve the quality of topics, an alternative to aggregation is the augmentation of individual messages with information from an external source. Across all the social platforms examined in chapter 3, it is common for messages to contain references to external information: e-mails may have attachments, while Tweets and Facebook posts often contain URLs. Hashtags implicitly link topically related messages. In the absence of references to directly related content, it is possible to retrieve text that is similar to the message content from an external corpus, e.g., by extracting keywords and feeding them to a web search engine. Here, we discuss two augmentation strategies that make use of URLs and the referenced web content. They are subjected to a brief extrinsic evaluation on the Twitter dataset in section 7.4.1, as part of a study on meso-level social influence.

We assume that, in a previous processing step, all URLs that occur in the dataset have been resolved to their final location, which is the case when a HTTP `GET` request for that URL yields either the desired content or an error, but not a redirection to a different URL. URLs that could not be resolved are removed from the source document. We further assume that the text body of the website has been extracted, i.e., navigational elements and other non-content have been discarded. This process is described in detail in section 3.1.1.

The purpose of a URL is to uniquely identify a resource and a location from which it can be retrieved, e.g., an HTML document on a web server. The general syntax of an absolute URL, as specified by RFC 3986 (Internet Engineering Task Force, 2005), can be expressed in extended Backus-Naur form as follows:

```
URL = scheme "://" authority path ["?" query] ["#" fragment]
```

The scheme specifies the method of access and determines the format of the other components. Here, we are only interested in URLs of websites, which use a scheme of “http” or “https”. In that case, the authority is the name or address of the server that hosts the site. Path and query are (mostly) free-form strings that are interpreted by the server, and jointly identify the desired HTML document (“web page”). By convention, the path contains hierarchical data, and is structured like a UNIX path name: a sequence of components separated by the ‘/’ character, with earlier components corresponding to higher levels of the hierarchy. The query contains non-hierarchical data, usually in the form of key-value pairs. The fragment identifies a logical subdivision of the resource, e.g., a particular paragraph of text in a document. In contrast to path and query, the fragment identifier is not sent to the server, but interpreted entirely by the client. Non-ASCII characters and characters with a special function are escaped (“percent-encoding”).

The scheme and the following separator characters are frequently omitted if it is sufficiently clear from the remaining parts or the context that the URL refers to a website. RFC 3986 calls this practice “suffix reference” and notes that such deliberately incomplete URLs cannot be reliably distinguished from non-URLs, so they are “primarily intended for human interpretation” (Internet Engineering Task Force, 2005, section 4.5). If an authority is specified, which is mandatory for absolute HTTP(S) URLs, the path may be empty. Query and fragment are generally optional. It follows that the minimum amount of information one can obtain from an URL is the identity of the server, but the amount of information in a typical

URL is higher. We distinguish three archetypes: In URLs of the first type, the path reflects the hierarchical organization of data on the server. Sometimes, information about the target document is deliberately encoded into the path for the benefit of the visitor or to make it easier for web search-engines to discover the document ("search engine optimization"). URLs of the second type resemble a remote procedure call (RPC): the path identifies the procedure and the query contains the parameters. The third type of URL is mostly opaque. Examples for each type are given below:

```
https://www.nytimes.com/2021/12/15/world/asia/china-russia-summit-xi-putin.html
https://www.urbandictionary.com/define.php?term=Thoughts%20and%20prayers&page=2
https://lore.kernel.org/lkml/a19f492d-6027-3e9b-9816-63b2f24c759a@oracle.com/T/#t
```

As one can see, the URL itself may already provide a useful summary of the target document. The URL-based augmentation strategy is to tokenize a URL and append the tokens to the source document instead of eliminating the URL from the source document entirely. Splitting URLs into meaningful tokens requires a domain-specific process. To detect both full URLs and suffix references, we use a regular expression that matches the scheme identifiers of the HTTP(S) protocol, but also any string that looks like a proper host name. Any found URL is split into its components. Host name and query are treated as individual tokens, while the fragment is ignored. Any percent-encoding in the path is reversed. The path is then split into tokens, with any character that is neither letter nor digit acting as a token boundary. Words in "CamelCase" are split further, then all characters are transformed to lower case. Tokens that consist entirely of digits are discarded. Finally, stop words are filtered out according to an URL-specific list. Like the medium-specific stop word lists, it is manually curated by inspecting the list of unique tokens in order of decreasing frequency and stopping at the first meaningful token.

Under the assumption that a tweet references a website in order to disseminate or engage in discussion with its content, it makes sense to consider the content of the website to be part of the message the tweet intends to convey. The content of the referenced website provides necessary context for understanding the tweet. This motivates the content-based augmentation strategy: the URL is removed from the source document, then the textual content is extracted from the referenced website and appended to the document. Both strategies, URL-based and content-based, simply append text from external sources to the original tweet, thus giving equal weight to internal and external information.

Post-Hoc Aggregation

Up to this point, we have looked at aggregation as a means of improving the quality of topics. Another use case for aggregation is to obtain topic distributions θ for units of analysis that are larger than single documents, e.g., the topic distribution of all documents written by one author, or published within a specific time period. The most simple way to achieve this is to aggregate the observed documents into virtual documents of the desired granularity, and fit a new topic model to the aggregated corpus. This approach has several drawbacks: fitting a topic model from scratch is computationally expensive, large units of analysis imply strong aggregation, which may impair the quality of topics, and document-topic distributions of

different granularity are not comparable, since each model has its own topics φ. A better option is to fit a topic model to the non-aggregated corpus, and query that model with the aggregated corpus (or corpora). Querying a common base model requires fewer iterations of Gibbs sampling, and yields document-topic distributions that are comparable across all levels of granularity. We refer to this procedure as *post-hoc aggregation*.

If post-hoc aggregation by querying is too expensive for a particular application, a fast approximation is possible. As before, a base model is fit to the non-aggregated corpus. The Gibbs sampler's state is then reused for approximating the aggregated distributions. Since aggregation changes the grouping of words into documents, but does not affect the words themselves, the formula for estimation of θ (equation 4.6) can be adapted to use the sum of word counts over a set of documents $D \subset \{1 \dots M\}$. This yields equation 4.23, an approximation of the document-topic distribution θ_D of virtual document D. If $|D| = 1$, this is equivalent to the original estimator of θ for a single document. An application of approximate post-hoc aggregation to the analysis of e-mail is described by Geng et al. (2008).

$$\theta_D \in \mathbb{R}^K \text{ with } \theta_{D,k} \approx \frac{\left(\sum_{m \in D} c_{k,m,*}\right) + \alpha_k}{\left(\sum_{m \in D} c_{*,m,*}\right) + \sum_{k'=1}^{K} \alpha_{k'}} \tag{4.23}$$

4.2.3 The Author-Recipient-Topic Model

Instead of using a general-purpose topic model like LDA and addressing the issues that arise from the distinct nature of online communication data by means of heuristic-driven pre- and post-processing, one may ask how the model can be adapted to accommodate the characteristics of the data. The Author-Recipient-Topic model (ART; McCallum et al., 2007) is a variant of LDA that was originally designed for e-mail messages, but can be applied to any kind of addressive, textual communication. It extends LDA by introducing two observed variables, a for the sender of a message and r for the set of its recipients, and yields a distribution over topics for each sender-recipient pair, the *relationship-topic distribution*. In other words, ART assigns topic distributions to the edges of the implicit social network graph induced by communication. The graphical model of ART is shown in figure 4.4. Each person that appears in the corpus as a sender a_m or recipient $r \in r_m$ of a message m is identified by a unique index $1 \leq i \leq A$. Let *Unif* refer to the probability mass function of the discrete uniform distribution. The generative process of ART can then be stated as follows:

1. For each topic index $1 \leq k \leq K$:
   a) Sample the parameters φ_k of the topic-word distribution from $\text{Dir}(\beta)$.
2. For each sender-recipient pair (i, j) with $1 \leq i, j \leq A$:
   a) Sample the parameters $\theta_{(i,j)}$ of the relationship-topic distribution from $\text{Dir}(\alpha)$.
3. For each message index $1 \leq m \leq M$:
   a) For each word index $1 \leq n \leq N_m$:
      i. Sample a recipient $x_{m,n}$ from $\text{Unif}(1, |r_m|)$.

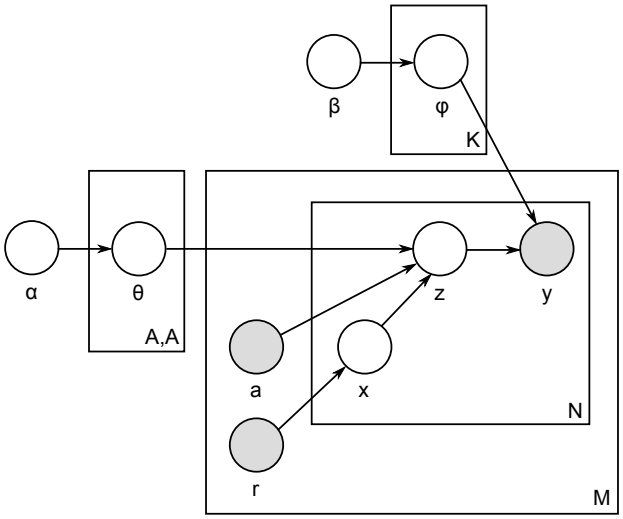


Figure 4.4: Graphical structure of the Author-Recipient-Topic model (adapted from McCallum et al., 2007)

ii. Sample the topic $z_{m,n}$ from $\text{Disc}(\theta_{(a_m, x_{m,n})})$.

iii. Sample the word $y_{m,n}$ from $\text{Disc}(\varphi_{z_{m,n}})$.

The equivalent joint probability is given in equation 4.24 (McCallum et al., 2007).

$$\begin{aligned} &p(\theta, \varphi, x, z, y | a, r, \alpha, \beta) = p(\theta|\alpha) \cdot p(\varphi|\beta) \cdot p(x|r) \cdot p(z|\theta, a, x) \cdot p(y|\varphi, z) \\ &= \prod_{i=1}^{A} \prod_{j=1}^{A} \text{Dir}(\theta_{(i,j)}|\alpha) \cdot \prod_{k=1}^{K} \text{Dir}(\varphi_k|\beta) \cdot \\ &\prod_{m=1}^{M} \prod_{n=1}^{N_m} \left(\text{Unif}(x_{m,n}|r_m) \cdot \text{Disc}(z_{m,n}|\theta_{(a_m, x_{m,n})}) \cdot \text{Disc}(y_{m,n}|\varphi_{z_{m,n}}) \right) \end{aligned} \tag{4.24}$$

The relationship-topic distribution of a sender-recipient pair (i, j) is informed by all messages from i to j. If every message has exactly one recipient, ART is equivalent to LDA in conjunction with aggregation of messages by their sender-recipient pair. ART deals with multiple recipients by assuming that the sender has written each word of the message with one particular recipient in mind, and equal parts of the message are addressed to each recipient. This modeling assumption is consistent with the concept of addressivity, which requires the sender to be individually aware of each recipient, and appropriate for messages long enough to plausibly be associated with multiple topics. For short messages this assumption is counter-intuitive; one would rather associate the message as a whole with a particular recipient. However, as long as a sufficient amount of messages of a particular sender-recipient pair is provided, the implicit aggregation of messages by sender-recipient pair ensures that a meaningful relationship-topic distribution can be obtained, even when the individual messages are short.

Compared to LDA, ART yields topic distributions of suitable granularity for social network analysis without the need for post-hoc aggregation, mitigates the effect of low average message length to some extent, and handles messages with more than one recipient in a principled way. In a setting where a message can be associated with multiple sender-recipient pairs, plain LDA is particularly difficult to apply: a message with multiple recipients would have to be processed once for each sender-recipient pair, which introduces bias towards the content of messages with many recipients. One disadvantage of ART is that it is specifically designed for addressive communication and requires each message to have a sender and a recipient, while many online social platforms support both addressive and non-addressive modes of communication. To obtain representations for a mixture of addressive and non-addressive messages from an ART model, we introduce the convention of assigning any non-addressive message to a loop edge of its sender, i.e., treating the sender of the message as its recipient for modeling purposes.

McCallum et al. (2007) derive a collapsed Gibbs sampler for ART. The full conditional (equation 4.25) is a joint distribution over $x_{m,n}$ and $z_{m,n}$. One can either sample from this joint distribution directly (blocked Gibbs sampling), or split it into separate full conditionals for x and z. The associated estimators for θ and φ are given in equations 4.26 and 4.27, and the marginal likelihood, with the latent variables x and z marginalized out, in equation 4.28. The word counts c are defined as for LDA, but a sender-recipient pair (i, j) takes the place of the document index.

$$\begin{aligned} &p(x_{m,n}, z_{m,n} | x_{-(m,n)}, z_{-(m,n)}, y, a, r, \alpha, \beta) \\ &\propto \frac{c^{-(m,n)}_{z_{m,n},(a_m,x_{m,n}),*} + \alpha_{z_{m,n}}}{c^{-(m,n)}_{*,(a_m,x_{m,n}),*} + \sum_{k=1}^{K} \alpha_k} \cdot \frac{c^{-(m,n)}_{z_{m,n},(*,*),y_{m,n}} + \beta_{y_{m,n}}}{c^{-(m,n)}_{z_{m,n},(*,*),*} + \sum_{w=1}^{W} \beta_w} \end{aligned} \tag{4.25}$$

$$\hat{\theta}_{(i,j),k} = \frac{c_{k,(i,j),*} + \alpha_k}{c_{*,(i,j),*} + \sum_{k'=1}^{K} \alpha_{k'}} \tag{4.26}$$

$$\hat{\varphi}_{k,w} = \frac{c_{k,(*,*),w} + \beta_w}{c_{k,(*,*),*} + \sum_{w'=1}^{W} \beta_{w'}} \tag{4.27}$$

$$p(y | a, r, \theta, \varphi, \alpha, \beta) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} \left(\frac{1}{|r_m|} \sum_{x \in r_m} \sum_{k=1}^{K} \theta_{(a_m,x),k} \cdot \varphi_{k,y_{m,n}} \right) \tag{4.28}$$

The techniques for hyper-parameter optimization, querying and updating the model, accelerated Gibbs sampling (FastLDA, AD-LDA), and estimation of the likelihood, as discussed previously, can all be applied to ART, mostly without modifications. One caveat is that for LDA, the units of sampling (words, grouped into documents) match the conditional independence assumptions of the model: documents are independent given the topic-word distributions φ. This is not the case for ART, where two messages may not be conditionally independent if they have the same sender and at least one common recipient. The corpus

can be partitioned into conditionally independent subsets $P_1 \dots P_n$ accordingly. When distributing the processing of messages across multiple threads of execution via AD-LDA, two messages may only be processed by separate threads if they are known to be conditionally independent, so all messages of a subset P_i have to be processed by the same thread. Similarly, when computing the predictive likelihood of ART by marginalizing out θ, the resulting term is not factorized by document, as in equation 4.11, but by subset P_i. This needs to be taken into account in the implementation of the LR algorithm.

Post-hoc aggregation over a set of senders and a set of recipients $S, R \subseteq \{1 \dots A\}$ works analogously to the aggregation over documents in the case of LDA. Given a corpus of messages, an aggregated relationship-topic distribution $\theta_{S,R}$ can be computed by replacing all senders $i \in S$ with a virtual sender i', all recipients $j \in R$ with a virtual recipient j', and querying an existing model with a corpus that consists of all messages that have i' as a sender and j' as one of the recipients. This aggregation can also be approximated via equation 4.29. Setting $S = \{i\}$ and $R = \{j\}$ yields the original estimator for the relationship-topic distribution of a single sender-recipient pair (i, j).

$$\theta_{S,R,k} \approx \frac{\left(\sum_{i \in S} \sum_{j \in R} c_{k,(i,j),*}\right) + \alpha_k}{\left(\sum_{i \in S} \sum_{j \in R} c_{*,(i,j),*}\right) + \sum_{k'=1}^{K} \alpha_{k'}} \tag{4.29}$$

We conclude that ART addresses, at least to some extent, the problems with online communication data that were discussed earlier in this section, and therefore elect to use it in the following case study and the experiments of chapter 5 and onwards.

4.3 Case Study: Finding Sequential Patterns in Dyadic Communication

In section 2.2.3, we compare different methodological frameworks for the content-based characterization of social relationships, and come to the conclusion that probabilistic topic modeling is the most promising approach, because the resulting representations are interpretable by humans, but also amenable to further computational analysis. An important point in favor of topic modeling is that topic models can jointly draw information from unstructured text and additional observations. This quality is exemplified by ART, which integrates the concept of messages with an observable sender and recipients into the topic modeling process. In this case study, we aim to demonstrate how a bespoke topic model that takes additional facets of online communication data into account can lead to new insights about social behavior.

To that end, we develop a topic model that, when applied to a corpus of dyadic communication data, is able to detect changes in the topic use of a sender-recipient pair over time, and can identify regularities or patterns. In other words, the model provides answers to the following questions:

1. What are the topics that govern the conversation for a particular interval of time?

2. Does the conversation move linearly from one configuration of topics to another, or are certain configurations revisited with some regularity?

The structure of our proposed model builds upon the Hidden Markov Model (HMM; Rabiner, 1989). Our model combines a single-recipient variant of ART with a Bayesian Hidden Markov Model (BHMM; Goldwater and Griffiths, 2007; Johnson, 2007). Each message has a latent state variable, and each sender-recipient pair has one relationship-topic distribution per state and a matrix of state transition probabilities, so that the temporally ordered sequence of messages of a sender-recipient pair forms a Markov chain. The state of a message selects the relationship-topic distribution to be used. This mechanism promotes the assignment of topically similar messages to the same state.

There are two use cases for such a model. The first one is motivated by technical considerations: When working with a corpus of short documents, aggregation can improve the quality of topics. In an experiment on document aggregation (section 4.2.2), we compare several strategies, and find that the best general approach is to separate the messages by sender-recipient pair, and then heuristically aggregate short sequences of messages that are likely to exhibit high topical coherence. What if the topic model could automatically identify periods of high topical coherence, and in that way find an aggregation strategy that is directly supported by the data?

The second use case is motivated by social theory, and requires further elaboration. In the social sciences, there is a general consensus that a relationship is a process that follows a life cycle of initiation, maintenance, and deterioration (cf. Parks, 1997). Parks (1997) identifies six dimensions "along which interaction changes as relationships develop and deteriorate": interdependence in the sense of Kelley (see the introduction to chapter 2 for a definition), breadth or variety of interaction, depth or intimacy of interaction, commitment ("the expectation that a relationship will continue into the future"; Parks, 1997), predictability (due to formation of norms), and coordination of communication (e.g., specialization of language, shift to non-verbal communication). With respect to these dimensions, Parks asserts that "change is rarely smooth or linear", but rather characterized by "considerable, perhaps constant, fluctuation" and "sharp breaks in which major changes in several dimensions occur at the same time", and concludes that "[p]eople may experience these abrupt shifts as 'turning points' and scholars may use them to mark the boundaries of different relational stages."

Parks' analysis suggests that the long-term temporal development of a relationship can be captured by a change point model that divides the observed lifetime of the relationship into discrete "relational stages". In the "stages and levels" model, the characteristics of a relationship at a particular stage are summarized by an aggregate score on an ordinal scale, the *level* (McCall, 1988). The assumption is that a relationship undergoes long-term processes of growth, stagnation, or deterioration, but its defining characteristics remain constant over time. McCall (1988) rejects the notion of relationship development as a one-dimensional process. Like Parks, McCall views a relationship as a sequence of time periods with stable interaction behavior. However, in McCall's model of relationship development, these *phases* of stability follow each other without a predetermined direction or goal. The development of a relationship is understood as an "ongoing self-transformation" (McCall, 1988), and a phase transition may involve arbitrary changes in its characteristics. A transition from one phase

to another can be triggered by external events (shifts in the social environment), or happens for internal reasons, e.g., changing expectations of the participants towards the relationship (Asendorpf and Banse, 2000).

Under the assumption that the various factors that make up social interaction can somehow be measured at arbitrary points in time and meaningfully compared, a phase of a relationship can be defined as a period of time in which the difference between any two measurements does not exceed a specific threshold. It follows that these phases can be recursively subdivided by successively lowering the threshold. In practice, this process is limited by the temporal resolution of the observations: ultimately, one arrives at the level of individual actions. Conversely, it should be possible to start from individual observations, and recover higher-level phases. Marsden (1990) remarks: "To write of social structure as 'a persisting order or pattern of relationships among some units of sociological analysis' presumes some means of abstracting from these empirical acts to relationships or ties." However, not all relevant aspects of a relationship are equally observable or accessible to the analyst. For example, among the six factors named by Parks, determining the intimacy of a relationship from observed communication is a largely unsolved problem, while assessing the breadth of interaction is less complex: Parks (1997) states that breadth may be "conceptualized in terms of the variety of conversational topics" and "[t]he way relational partners introduce, develop and retire topics". After fitting an ART model to the observed communication, breadth is directly reflected by the entropy of the relationship-topic distribution.

Within the scope of this work, the only observable type of interaction is the exchange of electronic messages. Topic models can, by construction, only discover coarse semantic grouping. This limits us to a purely content-based analysis of social relationships, as discussed in section 2.2. While we cannot claim that the intervals detected by our topic model correspond to stages or phases of the relationship lifecycle as defined by McCall and Parks, it stands to reason that their analysis reveals something about the nature of the relationship. Even without drawing an explicit connection to social psychology, one can see that the output of the proposed topic model provides value to the analyst: It allows to distinguish between topics that describe the current conversation, topics that are no longer current, and topics that have a periodic pattern of occurrence.

4.3.1 Related Work

A central element of our proposed model is the integration of an HMM into the topic modeling process, which introduces a Markovian dependency between successive messages of a sender-recipient pair. This draws upon earlier work by Griffiths et al. (2005), whose "composite model" (sometimes called LDA(-)HMM or HMM(-)LDA in the literature) combines LDA with an HMM to introduce a dependency between successive words in a document, thus weakening the bag-of-words assumption. In addition to the topic assignment, each word has an associated latent Markov state $s \in S$. In one particular state, the current word is generated by the topic-word distribution of the assigned topic, in all other states it is generated by the corresponding (discrete) output distribution of the HMM. In other words, the topic model acts as the output distribution of a specific HMM state. The authors show that this model is able to separate sequences of text with a primarily syntactic function, which

are captured by the HMM, from words with high semantic content, which are assigned to topics.

Several authors modify the composite model to distinguish multiple classes of meaning-carrying words. A typical use case for such models is unsupervised part-of-speech tagging. Jiang (2009) and Darling et al. (2012) independently propose a variant where all states $s \in S_{\text{sem}} \subseteq S$ have LDA-like output, each with separate topic-word distributions φ_s. Semantic coherency between topics of different states is promoted by use of the same document-topic distribution θ in each state. Jiang also explores a variant where not only the emissions of the LDA-like states, but also the state transitions, depend on the current topic. The CDHMM (Moon et al., 2010) distinguishes "functional states" with HMM-like output from "content states" that double as topic assignments. The probability of transition from state s to a content state $t \in S_{\text{sem}}$ is proportional to the product of a global transition probability $\pi_{s,t}$ and the document-topic probability θ_t.

All of the above models apply the HMM at the word level, using it to model the sequential dependencies between words in a document. A number of domain-specific topic models introduce Markovian dependencies between higher-level units of organization: The model of Du et al. (2012) specifically targets longer texts that can be subdivided into a sequence of structural elements like chapters or paragraphs, making the assumption that each element is related, content-wise, to its predecessor. For each document m, an initial topic distribution $\theta_{m,0}$ is drawn from a Dirichlet distribution. For the i-th structural element, a topic distribution $\theta_{m,i}$ is drawn from a Pitman-Yor process, using $\theta_{m,i-1}$ as the base distribution, thus forming a Markov chain. Apart from that, the generative process is identical to LDA.

Wei et al. (2007) demonstrate that the same principle can be applied to modeling the temporal development of topic use in a sequence of interrelated documents. If the time line can be discretized, for example, by obtaining documents at equidistant observation points in time or by aggregating observations from successive intervals of equal length into documents, the evolution of the document-topic distributions can be modeled with an HMM. However, due to the limited Markov horizon, this approach is highly sensitive to missing data and data that is insufficient in volume for a robust estimation of the document-topic distribution. Kleinberg (2002) shows that sequences of human interactions are particularly prone to missing data after temporal quantization due to their inherently bursty nature (see also section 3.5.2). If the temporal order of interactions is more important than their exact spacing, a pragmatic solution is to fit an HMM to the sequence of interactions, ignoring their temporal distance. Ritter et al. (2010) apply this technique to the classification of tweets by dialog act, i.e., their functional role in a conversation. Their model distinguishes global and state-specific topics. Each HMM state corresponds to a dialog act, and contributes its state-specific topic-word distribution to the topic model that generates the tweet. This distribution is expected to capture words that are associated with the dialog act. Compared to the models discussed here, our proposed topic model is most similar to that of Ritter et al., the main differences being that the HMM state selects a sequence-specific document-topic distribution, and all states use the same topics.

One of the use cases of our proposed model is the explorative analysis of online communication datasets, with a particular focus on learning about social interactions. In the literature, one can find several examples for task-specific topic models that specifically target the anal-

ysis of social behavior. Chang et al. (2009a) present a topic model that extracts information about the relationships between named entities from unstructured text, and therefore can, for example, generate an annotated social network of the characters that appear in a book, or enrich an existing social network with information from an external document. A common application of topic modeling is to identify topical communities, that is, groups of actors that are densely connected in the network graph and are associated (e.g., as senders or recipients of a message) with documents that are similar in terms of topic use (Wang et al., 2005; Zhou et al., 2006). The Community-ART model (Pathak et al., 2008) particularly stands out, being a natural extension of ART. The Probabilistic Social Annotation model (Kashoob et al., 2009) identifies topical communities in a social tagging setting (see section 2.1.3), where resources are annotated with tags by different users.

4.3.2 The Message Sequence Topic Model

We propose a novel topic model that operates on message sequences, the Message Sequence Topic Model (MSTM). A sequence is defined as an ordered set of messages of a single sender-recipient pair. We assume that a message only has one recipient, but later discuss how the model could be adapted to handle multiple recipients. The main conceptual difference between ART and MSTM is that the former considers all messages with the same sender and recipients to be exchangeable, while the latter introduces a Markov dependency between successive messages. To do this, the model has to impose an order on the messages of a sequence, but does not prescribe a particular ordering relation. Considering that the goal of this case study is learning about the temporal development of relationships, chronological order is a natural choice.

The graphical model is shown in figure 4.5. Note that the loop edge of variable s is a non-standard extension to plate notation, which we introduce to keep the graph comparable to those of LDA (figure 4.1) and ART (figure 4.4). Assume that each of the M instantiations of the "message plate" is identified by an index $1 \leq m \leq M$ that matches the position of the message in the sequence, and let s_m denote variable s of the m-th instance. The loop edge then corresponds to an edge from s_m to s_{m-1} for each $m > 1$. Figure 4.6 explicitly shows the relationships between the per-document variables.

A separate matrix $\pi \in \mathbb{R}^{S \times S}$ of state transition probabilities is maintained for each sequence. The matrix π can be decomposed into row vectors π_u^T of probabilities for transitions from source state u, each with a Dirichlet prior distribution: $\pi_u \sim \mathrm{Dir}(\gamma_u)$ with $\gamma \in \mathbb{R}^{S \times S}$ While it would be possible to use a scalar parameter γ that just controls the overall sparsity of π, having a separate parameter vector for the prior of each π_u allows us to specify our prior knowledge about the structure of π, e.g., to either promote or penalize self-transitions. Each state u has an associated relationship-topic distribution $\theta_u \sim \mathrm{Dir}(\alpha)$, and each message in a sequence has a latent state assignment s, which selects the relationship-topic distribution. In any situation that would require a reference to the state of a message before the first or after the last, we assume that this state is 1. In consequence, the transition probabilities from state 1 also act as the initial state probabilities of the sequence.

The remaining parts of the model are highly similar to LDA: a latent topic assignment z for each observed word y, and a global set of topics φ. The overall structure of the model

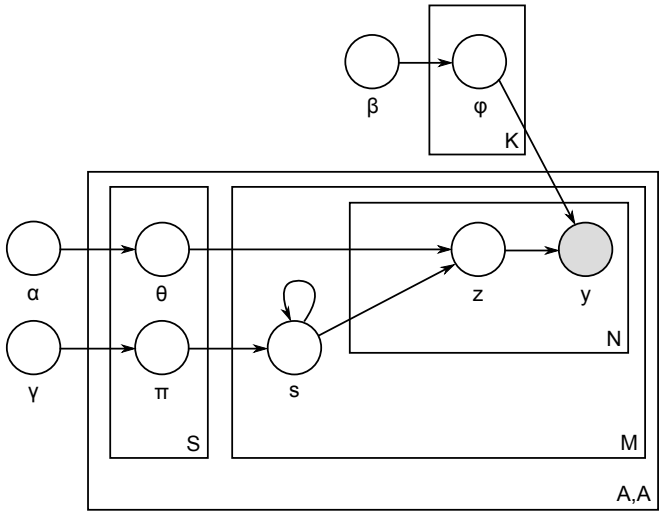


Figure 4.5: Graphical structure of the Message Sequence Topic Model. The loop edge indicates a dependency on the corresponding variable of the previous message.

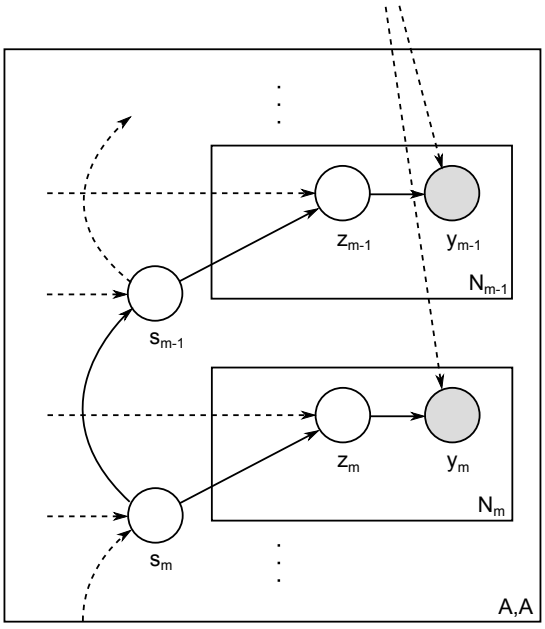


Figure 4.6: Detail view of the "sequence plate" of the Message Sequence Topic Model. The per-document variables are explicitly instantiated, so that the relationships between the document state variables s can be shown.

is less complex than ART, because each message can only have a single recipient and all messages in a sequence belong to the same sender-recipient pair. Since the identity of a sequence is equivalent to the identity of a sender-recipient pair, the messages of a sequence can be directly associated with their state transition probabilities π and relationship-topic distributions θ.

For the sake of notational brevity, we introduce q as a shorthand for the sender-recipient pair (i, j) of a message. The generative process of the MSTM can then be described as follows:

1. For each topic index $1 \leq k \leq K$:
   a) Sample the parameters φ_k of the topic-word distribution from $\text{Dir}(\beta)$.
2. For each sequence, identified by its sender-recipient pair $q := (i, j)$ with $1 \leq i, j \leq A$:
   a) For each state $1 \leq u \leq S$:
      i. Sample the parameters $\theta_{q,u}$ of the relationship-topic distribution from $\text{Dir}(\alpha)$.
      ii. Sample the transition probabilities $\pi_{q,u}$ from $\text{Dir}(\gamma_u)$.
   b) For each message index $1 \leq m \leq M_q$:
      i. Sample the message state $s_{q,m}$ from $\text{Disc}(\pi_{q,s_{q,m-1}})$ if $m > 1$, from $\text{Disc}(\pi_{q,1})$ otherwise.
      ii. For each word index $1 \leq n \leq N_{q,m}$:
         A. Sample the topic $z_{q,m,n}$ from $\text{Disc}(\theta_{q,s_{q,m}})$.
         B. Sample the word $y_{q,m,n}$ from $\text{Disc}(\varphi_{z_{q,m,n}})$.

The equivalent joint probability is given in equation 4.30.

$$
\begin{aligned}
& p(\theta, \varphi, \pi, s, z, y | \alpha, \beta, \gamma) \\
&= p(\theta|\alpha) \cdot p(\varphi|\beta) \cdot p(\pi|\gamma) \cdot p(s|\pi) \cdot p(z|\theta, s) \cdot p(y|\varphi, z) \\
&= \prod_{k=1}^{K} \text{Dir}(\varphi_k|\beta) \cdot \prod_{i=1}^{A} \prod_{j=1}^{A} \left[\prod_{u=1}^{S} \left(\text{Dir}(\theta_{q,u}|\alpha) \cdot \text{Dir}(\pi_{q,u}|\gamma_u) \right) \cdot \right. \\
&\quad \left. \prod_{m=1}^{M_q} \left(\text{Disc}(s_{q,m}|\pi_{q,s_{q,m-1}}) \cdot \prod_{n=1}^{N_{q,m}} \left(\text{Disc}(z_{q,m,n}|\theta_{q,s_{q,m}}) \cdot \text{Disc}(y_{q,m,n}|\varphi_{z_{q,m,n}}) \right) \right) \right]
\end{aligned}
\tag{4.30}
$$

The MSTM can be thought of as fitting a separate HMM to each sequence. The "output distributions" of these HMMs are topic models that generate entire documents. Each topic model has its own document-topic distribution θ, which it uses for all the documents it generates, while the topics φ are shared by all models and across all sequences. This construction promotes the assignment of topically similar messages to the same state. The state of a message depends on the state of the previous message in the sequence via the transition probability π, so one can use the prior distribution of π to influence the sequence of state assignments. For example, by choosing a prior that favors self-transitions, one can

obtain longer runs of successive messages that are assigned to the same state, in exchange for more overlap between the topic distributions θ_u of the different states. Conversely, one can estimate π (via its posterior predictive distribution) from the state assignments of a message sequence to learn about its temporal development. Note that the MSTM has a limited concept of time: Since the time stamp of a message only enters the model indirectly via the order of messages in a sequence, the model is insensitive to the temporal distance of messages. Two successive messages always have the same effect on the estimated model parameters, regardless of whether they were sent two minutes or two days apart.

The full conditional distribution of a Gibbs sampler that collapses out θ, φ, π is defined in equation 4.31 (see appendix A for the derivation). The word count $c_{k,q,u,w}$ is defined as the number of times word w is assigned to topic k in a message in state u that belongs to sequence q. The superscript $c^{-(q,m,n)}$ indicates that the word at position m, n in sequence q is excluded from the count. The transition count $g_{q,u,v}$ is defined as the number of messages of sequence q in state v with the state of the previous message being u. The superscript $g^{-(q,m)}$ indicates that document m of sequence q is excluded, and therefore the transition from document $m-1$ to document m and from document m to document $m+1$ is not counted. The full conditional as defined in equation 4.31 is a joint distribution of s and z, which can be used to implement a blocked Gibbs sampling scheme. Note that each time $s_{q,m}$ is resampled from the full conditional, the matrix of state-topic counts has to be updated to reflect that all words in message m are now assigned to a different state.

$$
\begin{aligned}
&p(s_{q,m}, z_{q,m,n} | s_{-(q,m)}, z_{-(q,m,n)}, y, \alpha, \beta, \gamma) \\
&\propto \frac{c^{-(q,m,n)}_{z_{q,m,n},q,s_{q,m},*} + \alpha_{z_{q,m,n}}}{c^{-(q,m,n)}_{*,q,s_{q,m},*} + \sum_{k=1}^{K} \alpha_k} \cdot \frac{c^{-(q,m,n)}_{z_{q,m,n},*,*,y_{q,m,n}} + \beta_{y_{q,m,n}}}{c^{-(q,m,n)}_{z_{q,m,n},*,*,*} + \sum_{w=1}^{W} \beta_w} \cdot \\
&\frac{(g^{-(q,m)}_{q,s_{q,m-1},s_{q,m}} + \gamma_{s_{q,m}}) \cdot (g^{-(q,m)}_{q,s_{q,m},s_{q,m+1}} + I(s_{q,m-1} = s_{q,m} = s_{q,m+1}) + \gamma_{s_{q,m+1}})}{g^{-(q,m)}_{q,s_{q,m},*} + I(s_{q,m-1} = s_{q,m}) + \sum_{u=1}^{S} \gamma_u}
\end{aligned}
\tag{4.31}
$$

Finally, one can derive estimators for the parameters of the multinomial distributions θ (equation 4.32), φ (equation 4.33), and π (equation 4.34). The marginal likelihood with latent variables s, z marginalized out (equation 4.35) involves a sum over all possible state assignments s'_q of a sequence q, which can be efficiently computed using the forward algorithm (Rabiner, 1989). The likelihood of unseen data can be computed using the LR algorithm.

$$
\hat{\theta}_{q,u,k} = \frac{c_{k,q,s,*} + \alpha_k}{c_{*,q,s,*} + \sum_{k'=1}^{K} \alpha_{k'}} \tag{4.32}
$$

$$
\hat{\varphi}_{k,w} = \frac{c_{k,(*,*,*),w} + \beta_w}{c_{k,(*,*,*),*} + \sum_{w'=1}^{W} \beta_{w'}} \tag{4.33}
$$

$$
\hat{\pi}_{q,u,v} = \frac{g_{q,u,v} + \gamma_v}{g_{q,u,*} + \sum_{v'=1}^{S} \gamma_{v'}} \tag{4.34}
$$

$$p(y|\theta,\varphi,\pi,\alpha,\beta,\gamma) = \prod_{i=1}^{A}\prod_{j=1}^{A}\sum_{s'_q}\prod_{m=1}^{M_q}\left(\pi_{s'_{q,m-1},s'_{q,m}} \cdot \prod_{n=1}^{N_{q,m}}\sum_{k=1}^{K}\left(\theta_{q,s'_{q,m},k}\cdot\varphi_{k,y_{q,m,n}}\right)\right) \tag{4.35}$$

Like in the case of ART, the FastLDA algorithm for accelerated sampling can be applied directly, but implementations of AD-LDA and the LR algorithm need to take into account that the smallest units of conditional independence, given the topics φ, are individual sequences.

Model Variants

As it is currently implemented, the MSTM takes a one-sided view on relationships, but one could just as easily treat sender and recipient as an unordered pair, so that messages in both directions contribute to the characterization of the relationship. The model is also trivially adapted to be more LDA-like and process sequences of regular documents instead of messages.

Extending the MSTM to support messages with multiple recipients is more difficult. Each author-recipient pair identifies a temporally ordered sequence of messages, so, for a given message m, the identity of its predecessor $m-1$ and successor $m+1$ is a function of its author and recipient. Handling multiple recipients like ART, by sampling a recipient for each word, makes it impossible to define such a function. One potential solution is to sample one recipient per message, but, under the modeling assumptions of ART, this makes inefficient use of large messages with many recipients (cf. the discussion in section 4.2.3). A more general solution could be to have one state assignment per message and recipient, thus allowing a message to be part of multiple sequences simultaneously.

Instead of learning about the temporal development of individual relationships, one might ask if it is possible to identify general patterns of interaction in a corpus of message sequences. The MSTM can be modified to use a single transition matrix π for all sequences, while keeping the per-state relationship-topic distributions θ local to each sequence. The joint distribution of an MSTM with such a global transition matrix is given in equation 4.36.

$$\begin{aligned} &p(\theta,\varphi,\pi,s,z,y|\alpha,\beta,\gamma) \\ &= \prod_{k=1}^{K}\mathrm{Dir}(\varphi_k|\beta)\cdot\prod_{u=1}^{S}\mathrm{Dir}(\pi_u|\gamma_u)\cdot\prod_{i=1}^{A}\prod_{j=1}^{A}\left[\prod_{u=1}^{S}\mathrm{Dir}(\theta_{q,u}|\alpha)\cdot\right. \\ &\left.\prod_{m=1}^{M_q}\left(\mathrm{Disc}(s_{q,m}|\pi_{q,s_{q,m-1}})\cdot\prod_{n=1}^{N_{q,m}}\left(\mathrm{Disc}(z_{q,m,n}|\theta_{q,s_{q,m}})\cdot\mathrm{Disc}(y_{q,m,n}|\varphi_{z_{q,m,n}})\right)\right)\right] \end{aligned} \tag{4.36}$$

A major flaw of this model is the non-identifiability of the global transition matrix π. Intuitively speaking, in a model with a local transition matrix, the meaning of each state u is defined by the associated relationship-topic distribution θ_u. With a global transition matrix, there is no structural element that promotes topical consistency of states across different sequences, so there may be many arbitrary combinations of global π and local θ that explain

Table 4.3: Perplexity of ART and MSTM on held-out data

model	perplexity	
	all sequences	length > 1
ART	9 087.3	9 070.0
MSTM	3 923.7	3 917.4
MSTM (shuffled)	3 887.1	3 880.5

the data equally well. There is no reason to believe, *a priori*, that one combination is a "more natural" explanation of the data than the others, so a global transition matrix, as defined here, is not interpretable.

4.3.3 Evaluation

The MSTM is evaluated on the Twitter dataset, specifically, on the messages sent by the 30 000 user subset within the two-month observation period. The evaluation is carried out in two parts: First, we compare the held-out likelihood of the MSTM to that of ART, in order to assess how well the model structure is suited to online communication data. The second part is an explorative study of the Twitter dataset with the aid of the MSTM. The goal is to identify basic patterns of dyadic interaction on Twitter, and at the same time, to demonstrate that the learned model parameters are amenable to human interpretation.

For the first part of the evaluation, the Twitter data is randomly split into a training and a test set. 80% of the data is used for parameter estimation, the remaining 20% is held out for estimation of the predictive likelihood. Non-addressive communication is represented by treating the sender of a non-addressive message as its recipient, so that all non-addressive messages of a particular sender form a sequence. ART and MSTM are fit to the training set. The MSTM has five states and a uniform transition prior $\gamma = 1.0$. Common hyperparameters of the two models and other experimental variables are chosen to be as similar as possible: In both cases, $K = 150$, $\alpha = \frac{50}{K}$, $\beta = 0.01$. Parameter estimation is performed via 2 000 iterations of Gibbs sampling, and the likelihood is computed using the LR algorithm with 100 particles and the logarithmic resampling strategy. Subsequently, training and test set are shuffled: in each message sequence, the order of messages is randomized. A second instance of the MSTM is fit, and its predictive likelihood is estimated, using the randomized training and test set. We compare MSTM and ART to find out if the ability to make use of information about message order improves model fit, and therefore potentially improves topic quality. We compare the fit of the MSTM to the original and shuffled data to test if the improvement, if any, can be attributed to information that is encoded into the order of messages, or is simply caused by the higher representational capability of the larger (in terms of parameter count) model.

Table 4.3 compares the perplexity of ART and MSTM given the test data. The MSTM generally outperforms ART, and when excluding sequences of length one from the test set (but not from the training data), the perplexity of ART rises slightly, while the perplexity

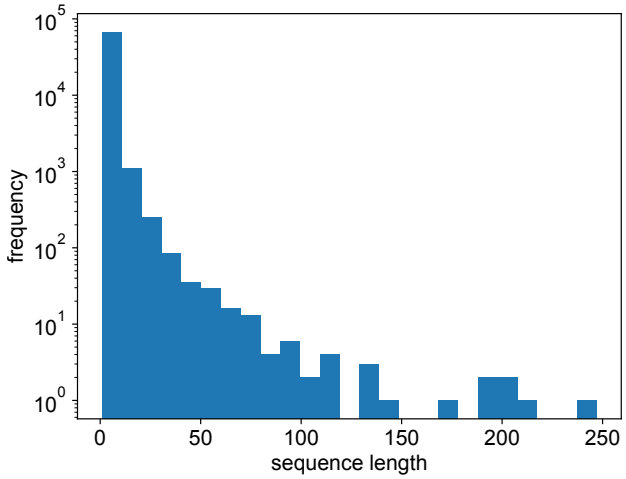


Figure 4.7: Distribution of the length of message sequences (addressive communication) in the Twitter dataset

of the MSTM decreases. However, the performance of the MSTM is not negatively affected by shuffling the sequences; perturbing the chronological order of messages even has a positive effect on perplexity. We therefore cannot attribute the improvement in performance relative to ART to specific temporal characteristics of online communication, e.g., topical consistency over longer spans of time, that the MSTM would be able to exploit. Rather, we have to attribute it mainly to the model's capability of grouping similar messages together, regardless of their temporal order, and assigning individual topic distributions θ_s to each group.

For the second part of the evaluation, we fit another instance of the MSTM to the Twitter data, but focus on interpretation of the model parameters. Of particular relevance are the HMM-like transition probabilities π, which reflect the temporal development of topic use. Since we are mainly interested in dyadic relationships and the associated communication behavior, we explicitly exclude non-addressive communication from this analysis. However, to make the best possible use of the data and to obtain high-quality topics, the model is still fit to the entire dataset. While the overall size of the corpus affects the topic quality, we expect the accuracy of estimation of the transition probabilities to depend on the sequence length. Below a certain threshold, there is not enough data to identify even substantial changes in behavior. The distribution of sequence length in the dataset is plotted in figure 4.7. While longer sequences exist, the majority of sequences is very short, consisting of ten messages or fewer (note the logarithmic scale of the y axis). The Twitter dataset is not an optimal choice for this study: a more suitable dataset would contain a higher proportion of long sequences, and a lower amount of non-addressive messages.

Hyper-parameters are chosen with interpretability in mind. Due to the low average sequence length, we choose an MSTM with three states and thus limit the analysis to coarse patterns of behavior. An asymmetric prior γ on the transition probabilities ($\gamma_{i,j} = 1$ if $i = j$, 0.5 otherwise) encourages self-transitions. This choice of prior reflects our belief, consistent

with the theories of Parks and McCall, that social interaction is characterized by periods of stability, during which the behavior of the actors does not change substantially. Additionally, encouraging self-transitions mitigates, to some extent, the effect of specifying an overly high number of states: all states beyond the first will remain unused unless the prior is overruled by a sufficient amount of evidence. Beyond that, we have no prior belief about the temporal development of social interaction, and therefore no reason to prefer linear development over recurrence, i.e., the repeated adoption of the same behavior patterns. The remaining hyper-parameters have the same values as before, and the same split into training and test set is used. The fitted model exhibits a perplexity of 5 961.9 on the test set, which is consistent with the results from the previous part, if one considers the reduced representational capability due to the lower number of states.

Visually exploring the space of transition matrices requires reducing their dimensionality. Here, we use "classical" multi-dimensional scaling (MDS; Mead, 1992), which embeds high-dimensional data in a lower-dimensional Euclidean space under preservation of the metric distances between the data points. A suitable distance function for transition matrices can be constructed by taking the average Jensen-Shannon distance (square root of the JSD, which satisfies the metric axioms; Endres and Schindelin, 2003) of their rows: $d(A, B) = \frac{1}{S} \sum_{s=1}^{S} \sqrt{\mathrm{JSD}(A_s \parallel B_s)}$ However, this metric is not invariant to permutation of the states. In an MSTM, the "meaning" of a state s, i.e., its effect on the output when viewing the model from a generative point of view, is defined by its associated topic distribution θ_s. Although these distributions are determined by the data, the order of their assignment to the states is arbitrary, i.e., neither constrained by the model nor by the data. After fitting an MSTM to a corpus of message sequences, two sequences that are highly similar content-wise may receive transition matrices that are highly dissimilar in terms of JSD, only because the association between topic distributions and states is permuted.

In the present setting, there is little potential for permutation, because of the low number of states and the fixed initial state. Therefore, we only briefly discuss the suitability of the *divergence rate* of Yang et al. (2020) as an alternative to the JSD-based distance metric. The divergence rate represents the dissimilarity of the output of two HMMs, and is invariant to permutation of states by design, since the output is jointly determined by the transition and output probabilities. However, the application of the divergence rate in this setting poses two problems: First, it violates the triangle inequality, so it is not appropriate for use with classical MDS. There are variants of MDS that do not require a metric distance. A more important problem is that the divergence rate, when applied to the per-sequence parameters of an MSTM, is dominated by differences in the topic distributions θ_s. After clustering the sequences in embedding space, the most central points of each cluster tend to have transition matrices that assign a self-transition probability close to one to the initial state. In effect, the clusters represent differences in overall topic use rather than different temporal patterns.

Dimensionality reduction of the MSTM transition matrices is performed according to the following procedure: Duplicate transition matrices are eliminated. The remaining matrices are embedded into a two-dimensional space using classical MDS with the JSD-based distance metric. Figure 4.8 shows the result of this process as a scatter plot (68 780 points before, 1 481 after deduplication). Each point corresponds to the embedding of one sequence, and is colored with an intensity proportional to the sequence length. One can see several trajectories

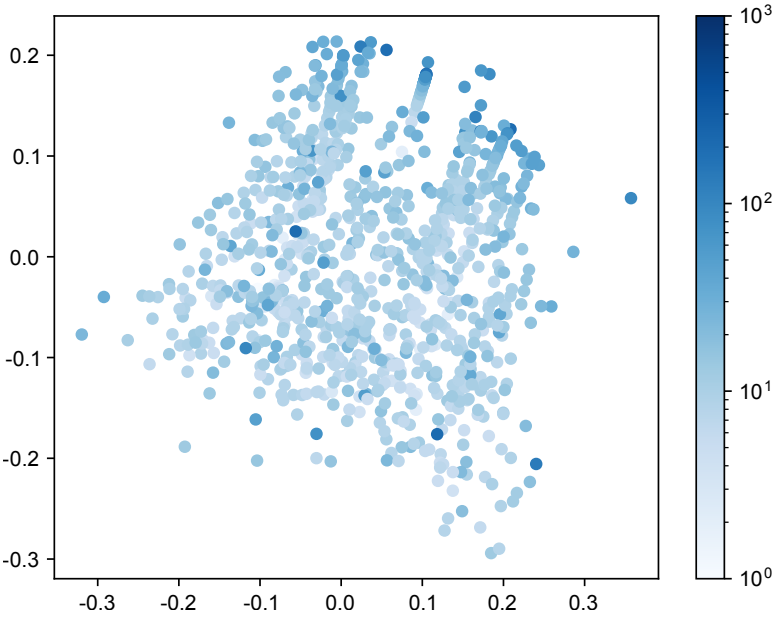


Figure 4.8: Two-dimensional embedding of MSTM transition matrices of sequences from the Twitter dataset; color intensity proportional to sequence length

of sequences with highly similar temporal development in the rectangular area between $(-0.1, 0.2)$ and $(0.3, 0.0)$. Their transition matrices are moving away from the prior as more and more evidence becomes available. Generally, transition matrices that differ from the expected value of the prior appear to be located farther away from the origin.

In order to take inventory of the different patterns of temporal development that are identified by the MSTM, we perform the embedding procedure a second time, but exclude all sequences that contain fewer messages than the MSTM has states. This restricts our analysis to the subset of sequences that are able to visit each state at least once. Clustering with the k-means++ algorithm (Arthur and Vassilvitskii, 2007) is performed on the points in the embedding space. The number of clusters is fixed to an arbitrary, low value (six). For each cluster, the data point closest to the centroid is chosen as an exemplar that represents the cluster. Note that the k-medoids algorithm (Kaufman and Rousseeuw, 1990), which is able to cluster the transition matrices directly, i.e., without prior embedding into a Euclidean space, is a more principled alternative to this procedure, if one is not interested in joint clustering and visualization of the matrices.

The result of the clustering is shown in figure 4.9 (13 708 points before, 1 471 after deduplication). For each cluster, the transition matrix corresponding to its exemplar is visualized as a graph (identified by letters A–F). An important property of the MSTM is that one can easily check if a state is not used, i.e., never assigned to any message of a sequence, by inspection of the latent state assignments. The transition probabilities π_s of an unused state, as well as its topic distribution θ_s, are equal to the expected values of their respective priors. Unused states are omitted from the visualization, and the probabilities of the remaining states are re-normalized.

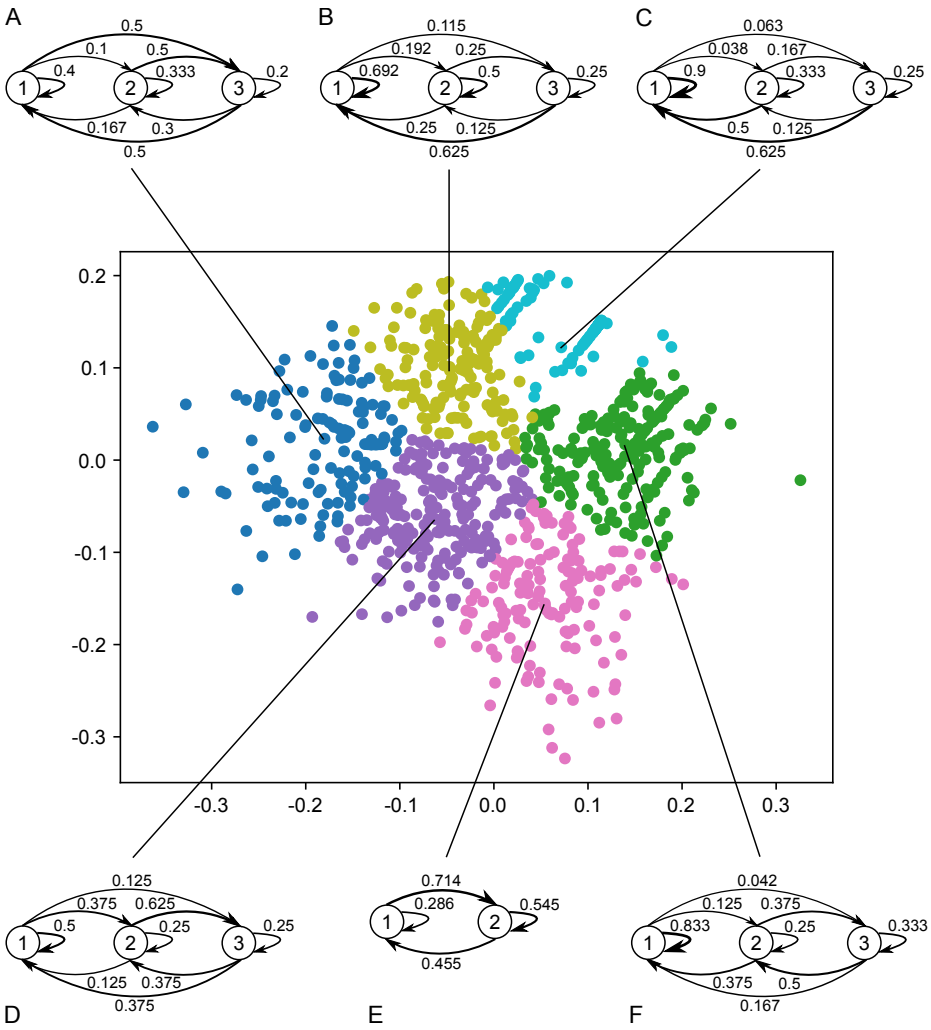


Figure 4.9: Two-dimensional embedding of MSTM transition matrices of sequences from the Twitter dataset; color indicates cluster assignment

The most simple pattern is 'C': the model almost exclusively remains in the initial state, so the sequence does not exhibit any substantial variation in topic use. If a transition to a different state does happen, it returns to the initial state quickly. Pattern 'F' is similar, in that it mostly remains in the initial state, but if it does transition to the other two states, it spends more time there. The observed "trajectories", groups of sequences that appear to converge to a certain point in embedding space with increasing length, all belong to these two clusters. This indicates that a large group of sequences is characterized by behavior that does not change over time, and are therefore best described by a single topic distribution. Pattern 'B' also prefers to remain in the first state, and otherwise tends to traverse the second and third state linearly before returning to the first.

Pattern 'E' corresponds to a sequence that switches between two topic distributions. The second topic distribution tends to remain in use for longer periods of time than the first. Pattern 'A' switches between the first and the third state, and occasionally visits the second. Pattern 'D' tends to move linearly from the first to the second and third state, and spends more time in the first state than in the other two.

Based on these observations, we can identify three main forms of topic use over time: First, topic use remains mostly constant ('C', 'F', 'B'). Second, topic use moves freely between two or three configurations ('E', 'A'). Third, topic use tends to move linearly from one configuration to the next ('D'). Any deeper, socio-psychological interpretation of these results must necessarily be very cautious, but the patterns found here bear a closer resemblance to McCall's recurrent phases than to Parks' linear stages of development.

In conclusion, the present case study demonstrates the utility of the MSTM for exploring and learning about the temporal dynamics of topic use. Furthermore, there is evidence that the MSTM is useful as an alternative to heuristic aggregation strategies when operating in a setting where short documents are common.

5 Influence in Online Social Networks

Verbatim and near verbatim quotations from prior publications (Hauffa et al., 2016, 2019) are highlighted in gray.

Online social platforms enable a new kind of public discourse that crosses geographical boundaries, but have lately become subject of critical investigation after the observation of adverse effects on discussion culture, such as the creation of virtual "echo chambers" where dissenting opinions are drowned out, the spread of misinformation, and even deliberate manipulation of the public opinion. A pertinent example are the results of the 2016 US presidential election, which are suspected to have been affected by automated social media posts, steering the discourse (Bessi and Ferrara, 2016). Learning how people exert influence on and receive influence from others via their online social networks is a necessary first step towards understanding and counteracting the negative aspects of online discussion culture.

Interpersonal social influence or *peer influence* is a longstanding subject of research in the social sciences. In an early work on this topic, Kelman (1958) defines influence as change in a person's "attitude and belief" that is "brought about by a particular communication or type of communication" and identifies three kinds of "process[es] whereby the individual accepts influence". While this definition of social influence continues to be generally accepted, the nature of the influence process remains the subject of debate and active investigation (Rashotte, 2007).

With the rise of online social network services, social interaction has become observable outside of constrained experimental settings and accessible to large scale data mining. In longitudinal interaction data, change in behavior can be directly identified, which enables reasoning about the underlying changes in attitude and belief and the process that drives these changes (Cosley et al., 2010). Online social networks have been shown to be similar to networks based on real-world interaction in many respects (Wellman, 1997; Petróczi et al., 2007). Of particular interest are Christakis and Fowler's studies on social contagion (2013), which demonstrate how social influence can induce changes in physical and emotional state (obesity, happiness) as well as in behavior (cooperation). These results have been partially replicated on online social networks (Coviello et al., 2014). The outcome of these studies suggests the existence of a general social influence process, which is independent of the medium and of the type of observable effect it produces. By analyzing online communication data in large volume, focusing on interpersonal influence and the detection of its effects in observed interactions, we attempt to identify fundamental characteristics of this process, while treating the process itself as a black box.

In this and the following two chapters, we discuss two experiments that are explorative in nature, in that they aim to provide insight into the social influence process by simultaneously testing a range of plausible hypotheses. Each experiment examines the social network graph at a particular level of scale: In chapter 6, we attempt to recover a graph of dyadic

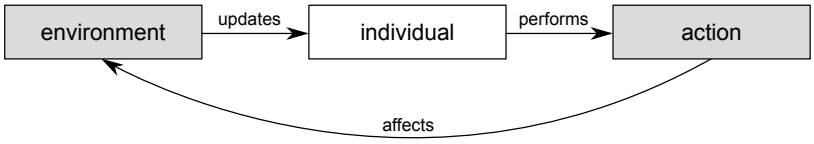


Figure 5.1: A simple model of human cognition

influence relationships from observed communication, which places the experiment on the micro level of analysis. In chapter 7, we ask to what extent people are influenced by their social environment. The main goal of this experiment to identify local substructures of the social network that contain information about the future behavior of an actor, so we locate it on the meso level of analysis. Both experiments make use of a topical representation of social relationships as argued for in chapters 2 and 4, and therefore serve as extrinsic validation for the hypotheses put forth in these chapters. The remainder of the current chapter is devoted to building a common foundation for the two experiments. One part of that is proposing a working definition of social influence and a basic model of the social influence process. By having each experiment refine that common, basic model of the influence process, we ensure that the results are comparable, at least on the conceptual level. The other part is defining a concrete topical representation for the communication data obtained from the various social media described in chapter 3.

5.1 A Basic Model of Social Influence

We begin by deriving an abstract model of social influence process from first principles. This model should make no more *a priori* assumptions about the nature of social influence than are necessary to ensure that, within the model, the analysis of influence is tractable. In the following chapters we iteratively refine and concretize this model.

In a primitive model of human cognition, as shown in figure 5.1, the behavior of an individual is determined by its internal state, which is constantly updated by perception of the environment. Change of behavior in reaction to events in the environment is the most general form of influence. The internal state is not observable, but observing both the environment and the behavior of an individual enables inductive reasoning about their relationship, and by extension about the underlying cognitive process. These inferences can be tested by using them to make predictions about future behavior. Social influence can then be defined as the subset of updates to the internal state caused by interpersonal interaction, particularly verbal and non-verbal communication, and their effects on future interactions. Leenders (2002) cautions that the relationship between internal state and produced behavior can be complex: "A change in some of the attitudes and beliefs does not automatically lead to changes in behavior. [...] Moreover, actors with different beliefs might well behave similarly."

This definition allows for different types of influence: By looking at the effect of observed interaction on produced interaction, one can distinguish simple forms such as *additive* and *subtractive influence*, where a person either adopts or drops a behavior pattern after observ-

ing it, from arbitrary influence, where observed behavior induces change in multiple facets of behavior, which may be unrelated to the observed behavior. Other characteristics of influence are the amount of exposure to observed behavior that is required to affect the internal state, and the time it takes for a change in the internal state to manifest itself in produced behavior. These two characteristics separate *short-term* from *long-term influence*. In this thesis, we limit our analysis to additive short-term influence.

From an outside perspective it is not possible to distinguish the effects of social interaction from the effects of general sensual perception on the observable behavior. If anything can be a potential source of influence, the amount of data that can be gathered in a practical setting will invariably be insufficient for reasoning. Data collection is usually limited to a subset of social interactions on a particular social platform. To make inference tractable, we need to make an additional assumption we call *locality of influence*[1]: The strength of influence of perceived behavior x on one's own behavior y is proportional to the similarity of the social context where x was observed and the context where y is produced. It follows that a local source of influence may override even strong external influences, but the resulting change in behavior may also be limited to a particular social context. The principle of locality can be applied at different levels of abstraction. In general, social behavior is mainly influenced by social interaction. Social groups are usually formed by individuals with similar interests or a shared purpose, so the focused communication among the members exerts a stronger influence on the behavior within the group than communication with outsiders. Under the assumption of locality, meaningful conclusions about the influence process can be drawn from a temporally and spatially limited set of observations of social interaction.

A direct consequence of the principle of locality is the importance of the social network for learning about influence. According to our definition, being influenced by someone requires *exposure* to and conscious perception of that person's actions, a condition which we summarily call *awareness* of that person. Leenders (2002) more generally states that "the availability [...] of information about the attitudes or behavior of other[s]" is a "precondition for social influence". The existence of an explicit social tie as a potential channel of communication is weak evidence for exposure, actually receiving a message and demonstrating awareness by acknowledging the receipt successively strengthen the case. As discussed in section 2.1.1, stronger evidence for awareness can be obtained from the frequency of interaction. To avoid having to choose an appropriate threshold for the number of interactions (Tsur and Lazer, 2017), one can require mutual interaction, i.e., an edge is removed from the interactive network graph if there is no edge in the opposite direction.

Influence can cause information or behavior patterns to diffuse through the network via a mechanism we call *indirect exposure*: If person a interacts with b, the content of the interaction may be (completely or partially) reflected in the future interactions of b with third parties c due to influence. Thus, even if there is no direct connection, c may be indirectly exposed to the behavior of a. In analogy to direct influence, *indirect influence* can be defined as the effect of indirect exposure on the internal state, which becomes visible as correlated behavioral changes in nodes that are not directly connected. Christakis and Fowler (2013) observe indirect influence up to a path distance of three.

[1] Not to be confused with *social influence locality* as defined by Zhang et al. (2015), which is not directly related.

In most practical settings, not all nodes and edges of a social network graph are observable. If there are strong influencers among the unobserved nodes, their confounding effect on the observed network may appear as spurious direct or indirect influence: if unobserved a influences b and c, and b and c are directly connected, their subsequent behavior may be indistinguishable from the effect of influence of b on c or vice versa. In certain cases, indirect exposure makes it possible to detect strong unobserved sources of influence that act upon a particular node. Due to homophily, members of a densely connected group of nodes in the network graph are likely to have similar sources of influence, observed and unobserved. In accordance with the principle of locality, these influencers have a similar effect on the behavior of each member, so aggregating the behavior of a group smoothes over individual differences, but preserves information about strong, homogenous influence effects. It follows that the aggregated behavior of people who are socially close to a node c should be to some degree representative of the behavior c is exposed to. More generally, indirect exposure imbues our model with some robustness to unobserved (or principally unobservable) strong influencers.

Concepts related to locality can be found in the literature: From a sociological point of view, locality of influence subsumes a number of phenomena, including, but not limited to the conformity to norms of behavior in a social group, familiarity, which is the preference of interaction with group members over outsiders, and homophily, the preferred association with individuals perceived as similar to oneself. Latané's dynamic theory of social impact (1996) contains the principle of "immediacy", which is conceptually similar to locality: "[...] if other factors are held constant, influence is directly proportional to the immediacy of the source of influence." Latané defines immediacy as a combination of variables, including "the clarity or richness of the communication channels" and geospatial distance. Empirical support for locality of influence can be found in the work of Myers et al. (2012), who attribute only 29% of information in a complete record of Twitter activity over one month to "external events and factors outside the network". The role of local graph structure for information diffusion in social networks is attested by several studies (Weng et al., 2013; Zhang et al., 2015, 2017). Finally, Lerman et al. (2011) point out a relationship between the similarity of the behavior of two Twitter users and their proximity in the follower graph, thus empirically linking social phenomena like homophily and influence to locality in terms of social network graph structure. However, if one considers influence to be the amalgamation of various cognitive processes, not all of them will exhibit the same degree of locality: By definition, imitation (Dijksterhuis and Bargh, 2001, section V) is highly local, while the process of mind-set formation, where successful behavior patterns are repeated in different situations, is highly non-local (Wyer Jr and Xu, 2010). It follows that building a model upon the assumption of locality may restrict the variety of manifestations of influence that can be represented by that model.

5.2 Topical Representation of Social Behavior

Both of our experiments involve predicting the content of future social interactions using a model of social influence trained on past observations. On all social platforms investigated

in chapter 3, unstructured text is a major, if not the main type of content. For that reason, a suitable numeric representation of unstructured text is required. In section 2.2, we present arguments for the use of topic models to build content-based intermediate representations of social behavior that are suitable for further computational analysis. Additional arguments in favor of topic modeling can be derived from the nature of the prediction problem: Consider a range of representations of rising dimensionality, starting from a binary representation ("does / does not contain a specific piece of information") analogous to information diffusion models (Kempe et al., 2003), and ending with a vector space model with one component per unique word. One can see that the dimensionality of the representation is directly related to the complexity of the prediction problem. Given a sufficiently high-dimensional representation, predicting future behavior is equivalent to predicting the exact wording of future messages, which is clearly infeasible. Conversely, a too low-dimensional representation is not sufficiently expressive for the manifestations of influence we are interested in (Brahim et al., 2013). Given evidence that a person's potential to exert influence depends on the topic of conversation (Barbieri et al., 2012), reasoning about the content of communication in terms of such topics is desirable. Related work indicates that a representation obtained from a probabilistic topic model is appropriate for observing the effects of social influence (Barbieri et al., 2012; Liu et al., 2010).

Characterizing the social influence process requires a corpus of observed social interaction that is large in volume and not restricted to a particular social group or subject matter. In chapter 3, we discuss how to obtain such datasets from three different social platforms: Twitter, Facebook, and e-mail. Our analysis of this data starts from the raw textual communication among users ("messages"), and only considers the set of core users (as defined in section 3.4) and their interactions. Given a set of K conversation topics, either chosen manually or automatically extracted from the data using probabilistic topic modeling (Blei, 2012), a message can be viewed as a mixture of these topics and represented by a K-dimensional real vector on the unit simplex Δ^{K-1}. We refer to that vector as the *topic distribution* of the message. By extension, a temporally ordered sequence of messages can be represented as a multivariate time series. Depending on the experiment, a sequence either contains all messages sent by one user, or the messages sent by one user to one particular recipient. To ensure the comparability of different time series, we partition the observation period into N intervals of equal length. For each interval and sequence, aggregate topic distributions are computed from all messages sent within that interval. The result is a longitudinal dataset consisting of one time series of length N per sequence.

A canonical observation period of 56 days (approximately two months) is defined for each dataset in section 3.5.3. Choosing the length of the intervals into which the observation period is subdivided is necessarily a compromise: Any kind of quantization constitutes loss of information, as Cosley et al. (2010) empirically confirm in the context of information diffusion models, so an argument can be made for choosing the shortest possible interval length. However, robust identification of the topic distributions requires a minimum amount of messages within each interval, considering that individual social media posts tend to be brief. Depending on the choice of interval length, phenomena on different time scales become visible or invisible. For example, long intervals will obscure short-term changes in behavior. The concrete choice of interval length depends on the nature of the experiment

and is discussed alongside the experimental setup in the following chapters.

After discarding all interactions outside of the chosen observation period, a substantial number of users will be inactive, i.e., no longer have any observed interactions with others. Furthermore, the bursty temporal distribution of messages (see section 3.5.3) implies that the shorter the quantization intervals, the higher the proportion of users that do not send any messages within an interval, which manifests as missing data in the time series. It follows that any approach to the detection and analysis of social influence in this setting has to be robust towards missing data, both in the form of inactive nodes and edges and in the form of unsystematically missing values in the time series of topic distributions. Before processing can continue, missing data must be addressed, which typically involves imputation and the exclusion of series with a high rate of missing data. Imputation directly imparts bias on the data, the exclusion of users (or user pairs) does so indirectly: In addition to the potential confounding effects of unobserved users, the removal of mostly inactive users might cause the characteristics of highly active users to be overrepresented in the data. For example, active users are more frequently exposed to other users' actions, and therefore might be generally more susceptible to influence. Since the actual quantity of missing values in the time series, their effect on the experiments, and possible strategies for their mitigation all strongly depend on the experimental setup, further discussion of these points is left to the following chapters.

Information sharing behavior, e.g., retweeting, is indicative of information flow, but difficult to fit into a model of social influence. Without knowing the sentiment of the sharer towards the original message, be it acceptance, rejection, or simply an objective and factual interest, one cannot infer the effect of the message on the sharer's state of mind. Aiming to achieve a clear separation between interpersonal influence and information diffusion, and considering that retweeting and similar information sharing behavior has already been thoroughly studied within the information diffusion framework (e.g., by Galuba et al., 2010), we remove all shared messages from the datasets before fitting the topic models. However, the classification of a message as "shared information" relies on meta-data that may not be perfectly accurate. For example, in the case of Twitter, "modified retweets" cannot be reliably distinguished from regular tweets.

Transforming the collected messages into time series of topic distributions is a two-step process. First, we fit an ART topic model (see section 4.2.3) to the data to identify K topics that best describe the data over the whole observation period. We choose ART over the MSTM (section 4.3), because the latter cannot handle messages with multiple recipients and introduces a new hyper-parameter (the number of states S), which would have to be tuned. To obtain the actual topic distributions for a given interval of time, the previously fit topic model is *queried* by repeating the parameter estimation process exclusively for the messages within that interval while keeping the topics fixed. If the interval does not contain any messages for a particular user (or pair of users), the resulting topic distribution will not be informative, so the absence of data is recorded in the time series. The resulting topic distributions are comparable across intervals and series, because they are mixture distributions that have a common set of topics as their support.

Querying an ART model takes a set of messages as input, each with a single sender and one or more recipients from an overall set of actors, and assigns a topic distribution to each

pair of actors. As discussed in section 4.2.3, the level of aggregation of the resulting topic distributions can be controlled via the mapping of observed senders and recipients (as defined in section 3.4) to values of the corresponding variables of the topic model. For the micro-level experiments, we use the observed sender of a message both as its sender and its sole recipient within the model, in order to obtain *actor-topic distributions* that represent the aggregate behavior of each observed actor. For the meso-level experiments, we distinguish between addressive and non-addressive messages. Addressive messages have one or multiple recipients. We use the actual sender and recipients of these messages to obtain a *relationship-topic distribution* for each pair of actors from the content of their interactions. Non-addressive messages do not have a well-defined recipient, so we fall back to the convention of treating the observed sender of a non-addressive message as its recipient, to obtain actor-topic distributions that specifically represent the content of non-addressive communication. From a social network point of view, these actor-topic distributions, either derived from the non-addressive communication or the aggregate behavior of individual actors, are associated with the nodes of the communication network graph, while addressive communication and relationship-topic distributions are associated with its edges. While this approach is directly motivated by our earlier attempts to characterize relationships by the content of the associated communication (see section 2.2.3), it is also, in some sense a departure: We do not analyze the bidirectional communication between two people as a whole, but look at the behavior of each individual, i.e., the contribution of that individual to the relationship, separately.

Values of the hyper-parameters of the topic model are chosen in accordance with recommendations from the literature: The number of topics K is fixed at 150. α and β are the parameter vectors of the Dirichlet prior on the topic distributions θ and the topic-word distributions φ, respectively, and control their sparsity. All components of β are set to 0.01 (Steyvers and Griffiths, 2007) to obtain a symmetric Dirichlet prior for φ, while α is determined in a data-driven way, allowing the prior of θ to be asymmetric. Parameters are estimated via Gibbs sampling after 2 000 iterations of burn-in. The document-topic distributions are computed by averaging over 20 samples with a lag of 5. Since we do not directly analyze the resulting model parameters, but use the topic distributions as an intermediate representation of the input data, we are to a certain degree insensitive to the choice of hyper-parameters. We only assume that the model is sufficiently expressive to preserve, at least in part, the influence effects that are present in the data.

One particular issue with this approach deserves elaboration: Since the quality of the estimated topics directly depends on the size of the dataset (Tang et al., 2014), we want to use data from the whole observation period for the initial parameter estimation. In consequence, topics are generated from data associated with a long time frame. From the perspective of a particular interval, this includes past data that may no longer be relevant due to its age, and future data that is only available in a retrospective experimental setting. In a setting where past and future observations are available, the topics of a model fit to the whole dataset potentially leak future information (Kaufman et al., 2011). A principled solution to this problem would involve either incrementally updating a topic model as described in section 4.1.1, "Updating the Model", or using a topic model that tracks the evolution of topics over time. We hypothesize that the overall effect of data leakage through topics is negligible,

and therefore do not specifically address the issue in the micro-level experiment. In the meso-level experiments, we determine the effect of data leakage on prediction accuracy by performing the experiments with a regular ART model and repeating the experiments under the same conditions, but using an incrementally updated topic model.

Other potential sources of information leakage are the social network graphs. In the case of explicitly declared social relations, the date of establishment of a relational tie is unknown, so we only know that an edge existed at the time of data acquisition (crawling). The implicit communication network is derived from communication within the observation period. While it is possible to compute the subgraphs associated with particular intervals, for the sake of simplicity and comparability with the explicit network we treat the graph as fixed over the entire observation period. As before, we assume that the effect of leakage is negligible.

5.3 Related Work

Existing approaches the analysis of social influence can be broadly divided into two categories, models of information diffusion and more general influence models. The study of information diffusion is concerned with one or more discrete and immutable units of information, which are transmitted from person to person. If there are multiple units, the diffusion of one unit is usually assumed to be independent of all others. Information enters the social network, and the nodes of the network graph are exposed to it by means of social interaction over a sequence of discrete time steps. Upon exposure to a unit of information, a node decides according to a probabilistic process whether to adopt or reject it. This decision process can be interpreted as a model of influence. The most basic diffusion models employ an analogy from epidemiology. In the SIR model, a node is in one of three states: susceptible, infective (has adopted the information), or recovered. At each time step, a proportion of infective nodes comes in contact with randomly chosen nodes anywhere in the network graph, infecting any susceptible node. Another subset of infective nodes recovers and thus acquires immunity (Newman, 2002b). Variants with different states and transition schemes exist. Since these models completely ignore the structure of the network graph, they can only be applied to the study of population-wide effects.

Kempe et al. (2003) discuss information cascade models, built on the assumption that information is propagated along the edges of the social network graph. Real-valued weights that represent the strength of influence are assigned to the edges. In the Independent Cascade model, an infected node gets one chance to infect each neighbor with a probability equal to the edge weight. In the Linear Threshold model, each node has an activity threshold. A node becomes infected once the weight sum of edges to already infected neighbors exceeds that threshold. The two models exemplify the difference between simple and complex contagion: In the complex case, the probability of an exposure to cause an infection depends on prior exposure, while in the simple case the probabilities are independent. The weights or transmission probabilities are usually assumed to be uniform (Bonchi, 2011, section III), but can be learnt from data (Saito et al., 2008; Goyal et al., 2010). Cascade models can be applied to data without a known social network graph, inferring the edges along which information

is transmitted from the temporal order of observed adoptions (Myers and Leskovec, 2010; Gomez-Rodriguez et al., 2012). The canonical use case for cascade models is influence maximization, which is the identification of a set of seed nodes that, when initially exposed to information, will maximize the number of infected nodes.

Online social platforms commonly provide mechanisms for information dissemination (e.g., forwarding, retweeting), so observational data of exposure to and sharing of information is readily available and has been studied under the assumption of an association between dissemination and influence. In some instances information sharing is treated as evidence for influence (Leskovec et al., 2006; Bakshy et al., 2011), others treat sharing as ground truth against which predictors of influence are evaluated (Kumar et al., 2016). Controversially, Brahim et al. (2013) argue that the assumption of immutability of single units of information renders diffusion models unable to express more complex influence phenomena found in communication data. They apply a cascade model to blog posts connected by hyperlinks to recover the pathways of information diffusion and manually analyze the textual content to quantify the amount of information transmitted from one post to another. They find the posts of a cascade to be unexpectedly heterogeneous in content, which can be taken as an argument for a process of partial contagion, where a node is permitted to modify the received information to a certain degree before transmitting it to other nodes.

Different ways of adapting information diffusion models to topical representations of interaction content have been proposed. Basic information diffusion models can be used without modification by classifying multiple pieces of information by topic and studying the differences in adoption on the topic level (Romero et al., 2011b). Barbieri et al. (2012) introduce topic-aware generalizations of the cascade models of Kempe et al. (2003) to account for people's varying degree of interest in different topics. They represent each piece of information by its topic distribution, and their model includes a latent variable for the topical influence strength of each observed edge between two actors. Grabowicz et al. (2016) use topic modeling to determine the interests of users from the content of the messages they send. They find that the probability of propagation of a message by a particular user is positively correlated with the similarity between the user's interests and the content of the message.

A variety of influence models exist outside of the information diffusion framework, and are therefore not built around the transmission of discrete, observable, and immutable units of information. One class of influence models is based on the "influentials" hypothesis from communication theory, which states that some individuals, by virtue of their personality or social role, exert a disproportionally large amount of influence over others (Bakshy et al., 2011). Attempts have been made to correlate the ability to exert influence with measurable quantities, e.g., centrality in a social network graph (Kiss and Bichler, 2008). Just like information diffusion is usually focused on the adoption of discrete behaviors, influentials are either assumed to exert influence of uniform strength, independent of the behavior that is to be stimulated in their targets, or exhibit varying strength of influence with respect to a discrete set of topics (Weng et al., 2010).

When dropping the assumption that the strength of influence is independent of the target or the behavior that is to be stimulated, the problem becomes one of recovering the latent influence graph from interaction data. The approaches of Tang et al. (2009) and Liu et al. (2010) share some conceptual similarities: Both use graphical models to relate an observed

social network graph to a collection of documents, with each document being represented as a probability distribution over discrete topics. The latent variables of either model induce an influence graph, in which vector-valued edge weights represent the per-topic strength of influence. The main difference is that the directed graphical model of Liu et al. is generative, that is, it defines an idealized procedure for generating documents in the presence of social influence. According to their model, the presence of an influence relationship makes the influencee more likely to copy content from the influencer, therefore influence manifests as textual similarity between documents. The direction of influence can be inferred from connections between the documents (e.g., conversation flow or academic citations). The undirected graphical model of Tang et al. learns about influence form numeric features like topical similarity of documents, interaction volume, and path distance of nodes in the social network graph. Other approaches to graphical modeling of social influence include the work of Bi et al. (2014), who try to identify topically coherent subgraphs of followers on a Twitter-like social platform, and Guo et al.'s model of language use over time (2015), which is founded on the assumption that in turn-based conversations, influence manifests as short-term linguistic accommodation.

The autocorrelation model of Friedkin and Johnsen (1999) describes the convergence of opinions within a small group of people, under the assumption that its members are not exposed to any external influence for the time it takes to reach a state of equilibrium. In the most simple setting, each person can freely interact with any other. Under these assumptions, the pairwise influence strength within the group can be expressed as a complete, weighted graph.

The detection of influence relationships among actors in a social network can also be regarded as a causal inference problem. Liotsiou et al. (2016) propose a theoretical framework based on graphical causal modeling. Within their framework, given data obtained from a suitably designed experiment, social influence can be distinguished from confounding factors such as homophily and sources of influence external to the observed social network. Practical applications of causal reasoning to the detection of social influence typically employ a more limited notion of causality. A number of works build upon Granger's definition of causality (GC; Granger, 1980) to reconstruct an influence graph from observed interactions. While GC offers a useful framework for modeling causal relationships in time series data, its implementations suffer from susceptibility to external confounding, the very problem causal reasoning sets out to solve (Maziarz, 2015). Chua et al. (2015) use a custom topic model to track the temporal evolution of topics in the titles and abstracts of papers within an academic citation network. They do not explicitly construct an influence network, but apply GC to the time series of topic distributions to test several hypotheses about influence flow, e.g., "first author influences second author". Acknowledging the issue of unobserved confounders, they do not claim to detect causal effects, but rather a superset of social phenomena they call "social correlation".

Transfer entropy (TE) is a measure of information transfer that can be understood as a model-free generalization of GC (Barnett et al., 2009). The application of TE to multivariate, topical representations of communication data is demonstrated by Ver Steeg and Galstyan (2013). They construct an influence graph for a sample of Twitter users by computing the topic distributions of their individual messages, computing the TE for all user pairs with

sufficient data, and connecting each pair for which the TE is greater than zero. The influence graph is evaluated by comparing it to a graph of users who frequently @-mention each other, which is interpreted as "a weak proxy for online influence" (Ver Steeg and Galstyan, 2013). To reduce the computational complexity of TE estimation, observations can be transformed into a time-quantized, binary representation, where each value indicates the presence or absence of a particular piece of information within an interval of time (Ver Steeg and Galstyan, 2012; Bauer et al., 2012; He et al., 2013). McKenney and White (2017) note that a non-zero TE does not necessarily correspond to a meaningful influence relationship. They propose a scheme for deriving a threshold from limited information about the true, unobserved influence graph, and evaluate it on various synthetic networks.

The main difference between our work and other studies of social influence is our intention of learning about the influence process. While epidemic models can be used to characterize the influence process at the macro level, no explanative model of influence exists on the level of local structures or individual nodes. Like other multi-faceted models of influence, we represent observed social interactions as probability distributions over topics, and define a transmission process that allows information to be modified in the process of transmission. While not all of the models mentioned above fully specify how observed interactions shape future interactions, those that do appear to be consistent with the abstract model given in section 5.1. Furthermore, the studies cited in this section implicitly assume that observed interactions among a closed group of people are sufficient to learn about influence with reasonable accuracy, prefer exposure by direct interaction, and nodes of low path distance over distant nodes, which is consistent with the principle of locality of influence.

6 Social Influence at the Micro Level

This chapter expands upon the results of the bachelor's theses of Wolfgang Bräu (2015) and Lisa Lörinci (2016), which were jointly supervised by Jan Hauffa and Georg Groh. The experimental results have been in part previously published at the 2019 Workshop on Social Influence, held in conjunction with the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2019; Hauffa et al., 2019). Verbatim and near verbatim quotations from prior publications (Hauffa et al., 2019) are highlighted in gray.

Social influence is often studied through the lens of information diffusion (Kempe et al., 2003): An actor in a social network is exposed to a discrete, immutable piece of information and may, under certain conditions, decide to adopt it and share it with others. This ultimately results in a cascade of adoptions that represents the diffusion through the social network. If both the act of receiving and the subsequent sharing of a particular piece of information can be observed unambiguously, these observations are evidence for a causal relationship in which one person exhibits a change in behavior in response to interaction with another. This agrees with a broad definition of social influence as "change in an individual's thoughts, feelings, attitudes, or behaviors that results from interaction with another individual or a group." (Rashotte, 2007)

The equation of information diffusion and influence can be challenged on two counts. First, restricting the scope of analysis to immutable pieces of information severely limits the range of influence phenomena that can be detected, yet the assumption of immutability is necessary for reliable attribution of a particular change in behavior to the prior action of another person. Second, the act of sharing a piece of information constitutes a change in behavior, but not necessarily in attitude. Among the conceivable motives for sharing information (Tufekci, 2014) there is wholehearted agreement and the resulting desire to share, but the decision might also be independent from any personal interest in the information and driven purely by social obligations. If an interaction only results in mechanical forwarding of the received information, the sender cannot be said to hold more than shallow influence over the recipient.

The empirical results of Brahim et al. (2013) support this point of view: After identifying cascades of blog posts connected by hyperlinks and annotating them according to their topic, they find that blog posts do not consistently adopt the topic of their parent in the cascade, particularly when the cascade is more linear than star-shaped. Petrović et al. (2011, section 5.1) find that human annotators perform above chance in identifying tweets that are highly likely to be retweeted, even when information about the social context of the tweet is withheld. This points towards the existence of purely content-wise characteristics of a tweet that strongly determine its "retweetability", i.e., features that are likely to trigger a reaction in a diverse range of people, independent of their social environment. Naveed et al. (2011) expand on this result by evaluating the utility of several content-based features in a machine

learning system for retweet prediction. They report above-chance accuracy in predicting whether a given tweet will be retweeted. The most predictive features match our intuition about what makes a tweet retweetable: addressive tweets are less likely, tweets containing URLs more likely to be retweeted. The content of a tweet also affects its retweetability, with some topics being more predictive of future retweeting than others.

While the connection between social influence and the diffusion of discrete units of information is tenuous, it can be argued that there is a much closer link between social influence and topical information diffusion. If we define topical diffusion, in analogy to discrete information diffusion, as the adoption of a topic after being exposed to it, the observable effect of adoption can take different shapes. Assuming, without limiting generality, that the exposure happens via a single, highly topical message, these include direct retransmission of that message, sending an original message with high topical similarity, and incorporating the topic into multiple forthcoming messages. The latter two cases indicate a stronger, and longer-lasting effect of exposure on the exposed person's attitude towards the topic.

In this chapter, we build upon these considerations, and propose a model of dyadic social influence based on topical information diffusion. If the implicit social networks formed by discrete information diffusion admit an interpretation as networks of influence (Gomez-Rodriguez et al., 2012), it should be possible to make a much stronger case for the presence of an influence relationship on the basis of observed topical diffusion. To test this hypothesis, we construct a framework for the inference of social influence from topical information diffusion and evaluate it on social media data obtained from Twitter (see section 3.1).

We specifically address two issues we believe have not been treated with sufficient depth in related work. First, we argue that an intuitively appealing measure of time-lagged topical similarity satisfies Granger's definition of causality and is therefore subject to the same theoretical limitations and confounding effects. Second, while causal effects in general and social influence effects in particular can be measured, it is not known to what extent these measurements correlate with the presence of what an average human observer would perceive as social influence. We discuss the difficulty of obtaining ground truth data for systematic evaluation and propose heuristic tests that do not rely on human annotation.

6.1 Topical Social Influence

Assume that the individual behavior of actors in a social network can be represented as a set of time series of K-dimensional topic distributions as described in section 5.2. Let the behavior of actor b at time t be denoted by $\theta^{b,t}$. Given two successive topic distributions $\theta^{b,t}, \theta^{b,t+1}$ from the time series of actor b, $\theta^{b,t+1}$ can be trivially decomposed into additive change, subtractive change and inertia, i.e., the tendency not to deviate from past topics:

$$\begin{aligned} \theta^{b,t+1} &= \theta^{b,t} + \delta^{+} - \delta^{-} \quad \text{with} \\ \delta_i^{+} &= \max\{0, \theta_i^{b,t+1} - \theta_i^{b,t}\} \\ \delta_i^{-} &= -\min\{0, \theta_i^{b,t+1} - \theta_i^{b,t}\} \end{aligned}$$

By normalization, the additive change can be expressed as a topic distribution θ^+:

$$m = \|\delta^+\|_1 = \|\delta^-\|_1 \quad \text{since } \theta^b \in \Delta^{K-1}$$
$$\theta^{b,t+1,+} = \frac{1}{m}\delta^+$$

We measure the strength of additive influence $I_{a\to b}(t)$ of an actor $a \neq b$ on b's actions at time $t+1$ by comparing the past topic distribution of a to the additive change of b:

$$I_{a\to b}(t) = m \cdot (1 - \mathrm{JSD}(\theta^{a,t}, \theta^{b,t+1,+})) \tag{6.1}$$

This reflects to what degree the change in topic use of b is determined by the past topic use of a, i.e., the amount of topical diffusion from a to b. By multiplying with m we ensure that $I_{a\to b}$ is proportional to the magnitude of change effected by a. If at least one of $\theta^{a,t}, \theta^{b,t}, \theta^{b,t+1}$ is unobserved, $I_{a\to b}(t)$ is defined to be zero.

The symmetry of the Jensen-Shannon divergence implies that $I_{a\to b}(t)$ is maximal when $\theta^{a,t}$ has no information not in $\theta^{b,t+1,+}$ (behavior that was observed, but not adopted), and $\theta^{b,t+1,+}$ has no information not in $\theta^{a,t}$ (adoptions that can not explained by observations). In other words, $I_{a\to b}$ takes on a value of one if, given two topics i, j ($i \neq j$), a exclusively writes about i and b exclusively about j, followed by b completely switching over to i. If instead the behavior of b does not change or b switches to k ($k \neq i$), $I_{a\to b}$ is zero. Conceptually, this approach is a straightforward generalization of the information diffusion model, with the JSD-based similarity of topic distributions replacing the direct tracking of discrete pieces of information.

The above definition of $I_{a\to b}$ contains the assumption that the potential influencing effect of a's behavior is constant over time: As long as a talks about a particular topic, a exerts influence with respect to that topic. An alternative hypothesis is that change in b's behavior is mainly triggered by exposure to change in a's behavior. The strength of influence exerted by a peaks when a starts talking about a new topic, and decays quickly afterwards. This can be approximated by replacing $\theta^{a,t}$ with $\theta^{a,t+1,+}$ in equation 6.1.

$I_{a\to b}$ can be aggregated over multiple time intervals to identify temporally stable influence relationships. In the following experiments we use the average strength of influence:

$$\bar{I}_{a\to b} = \frac{1}{N-1}\sum_{t=1}^{N-1} I_{a\to b}(t) \tag{6.2}$$

More complex weighting schemes are possible. For example, a temporally decaying weight $w_d(t)$ with a parameter $d \in [0, \infty)$ that interpolates between uniform weight and placing all weight on the most current observations can be realized with an exponential or power-law distribution (Sekara et al., 2016, S3.1):

$$w_d(t) = \frac{f_d(N-1-t)}{\sum_{u=0}^{N-2} f_d(u)} \quad \text{with} \tag{6.3}$$
$$f_d(t) = \exp(-dt) \quad \text{or}$$
$$f_d(t) = (t+1)^{-d}$$

We apply this measure of influence strength to the problem of *influence network recovery*: Given a social network represented by a graph (V, E) and the observed communication behavior of its actors within a predefined period of time, we want to construct the influence network, a directed graph with vertices V and edges (a, b) induced by the relation "a exerts influence on b". We call a the *influencer* and b the *influencee*. A real-valued edge weight in the interval $(0, 1]$ represents the strength of influence. Maximal weight indicates that future behavior of b is completely determined by past behavior of a.

6.1.1 Connection to Granger Causality

We define social influence as a causal relationship between influencer and influencee, so it makes sense to compare it to other, more general mathematical definitions of causality. A closely related definition of causality in time series, due to Granger (1980), rests on three assumptions: The future cannot cause the past, causal relationships remain constant in direction over time, and there exists a redundancy-free representation Ω_t of all knowledge up to the current point in time t. As long as these assertions hold, Y can be said to cause X if it contains unique information about the future of X, that is, $P(X_{t+1} \mid \Omega_t) \neq P(X_{t+1} \mid \Omega_t \setminus Y_t)$. We refer to this definition as *Granger's principle of causality* to distinguish it from its various operationalizations that are known as Granger causality (GC).

Any operationalization needs to introduce additional limiting assumptions, particularly with regards to the definition of Ω. By replacing Ω with the set $\{X, Y\}$, one can test for the limited notion of *prima facie* causality that ignores potential confounding effects of other variables. The hypothesis test involves fitting two linear vector autoregression (VAR) models, one to X and one to the column-wise concatenation of X and Y, and comparing the variability of the residuals (Geweke, 1982). If the variability of the concatenated model is significantly lower, the corresponding improvement in prediction can be attributed to Y, and causality is established. This procedure is commonly referred to as testing for Granger causality (GC), and is limited to detecting causal effects that manifest as lagged correlation of time series. GC has a long history of being applied to the study of human social behavior. For example, Kenny (1988) describes a test for GC in the special case of a dyadic relationship, where each actor is represented by a time series of univariate observations.

Transfer entropy (TE) is a measure of directed information transport that implements Granger's principle of causality. It is model-free and therefore can be estimated without making assumptions about the distribution of the observation data, particularly without requiring the interaction effects to be linear, and does not require temporal quantization. The higher expressivity comes at the cost of estimation procedures that are more complex, both computationally and in terms of the required sample size (Gencaga et al., 2015). In the specific case of Gaussian variables, TE and GC are equivalent (Barnett et al., 2009).

The strength of social influence $\bar{I}_{a \to b}$ can also be interpreted as an implementation of Granger's principle of causality. By using the JSD to compare past behavior of a to the additive change in behavior of b, we ask whether the past of a contains information about the future of b that is not present in the past of b, which is equivalent to Granger's notion of prima facie causality. Alternatively, the detection of social influence from a to b can be viewed as the constrained induction of a directed graphical model, as shown in figure 6.1,

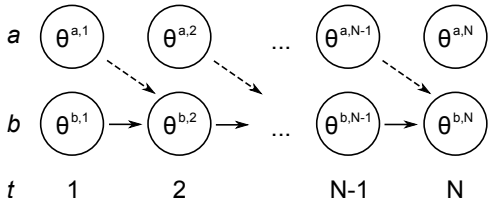


Figure 6.1: General structure of a graphical model representation of a's influence on b

where dependencies between past behavior of a and present behavior of b (dashed edges) are to be established (Lörinci, 2016). Our definition of social influence strength is conceptually equivalent to operationalizations of Granger's principle of causality that are similarly limited to pairwise observations, so the well-understood limitations of Granger causality (Maziarz, 2015) apply equally to influence strength. We discuss the main issues in the sequel.

Granger's second assumption requires social influence to remain constant in direction, which is not necessarily true in social relationships, where, for example, the roles of "conversation leader" and "follower" may switch over time. In such cases, unless the time frame of analysis matches the temporal characteristics of that dynamic, the result of causality testing is indeterminate. The detectability of causal effects also depends on the temporal quantization. If it is too coarse, observations corresponding to cause and effect end up in the same interval and become indistinguishable. However, when dealing with communication data one cannot arbitrarily shorten the intervals, considering that a certain amount of messages per interval is required for robust estimation of topic distributions. We therefore have to treat the interval length as a model parameter and attempt to experimentally find suitable values.

Granger's third assumption stipulates that true causality can only be established if all potential causes can be observed and entered into the model. If two variables X, Y have a common cause Z, GC may detect spurious causality between X and Y unless the confounder Z is controlled for. *Conditional GC* (Geweke, 1984) can test for causality with respect to an arbitrary number of possible confounders, but computational and sample complexity increase with the number of observed variables. Outside of experiments on synthetic data, some variables will always be unobservable. In consequence, GC captures a weaker notion of causality than, for example, a properly constructed graphical causal model (Liotsiou et al., 2016). However, conditional GC has been successfully applied to the problem of discovering connections between financial institutions (Basu et al., 2017), where a reasonably representative proportion of actors can be observed. Strong unobserved confounders that have the same effect on all observed variables result in redundancy in the VAR representations of the observations, comparable to our definition of indirect exposure in section 5.1. A method called *partial GC* (Guo et al., 2008) can exploit this redundancy to eliminate the correlation between the observed variables caused by the confounders.

We suspect that unobserved confounders are a major issue for the analysis of social influence, given that interpersonal communication is usually spread across many different channels of communication, and considering the existence of organizational influencers such as mass media. Granger (1980) suggests approaching the issue with a Bayesian mindset: to

first assess the prior probability of a causal relationship according to a domain-specific theory, and then adjust this belief according to the results of causality testing. In addition to spurious causality, the family-wise error associated with multiple hypothesis testing further increases the false positive rate. While influence relationships may exist among all $O(n^2)$ possible edges between the observed actors, it is therefore not reasonable to indiscriminately test all edges for influence. Due to the combined effects of spurious causality and multiple hypothesis testing, it is difficulty to justify a claim of influence with observational data alone. Within our model of social influence, being influenced requires exposure to the behavior of another actor, so our strategy is to construct an initial influence network according to evidence of exposure ("candidate network") and then refine it by causality testing.

A related issue is distinguishing the effects of social influence and homophily. Shalizi and Thomas (2011) show that when defining homophily as the statistical dependency of the probability of presence of an edge on latent characteristics of the individuals it connects, a broad class of probabilistic models is principally unable to make this distinction. GC belongs to the class of models that cannot distinguish influence from homophily, even in the hypothetical case where all actors are observable and entered into the model. Constructing a candidate influence network from a given social network, implicit or explicit, introduces a sensitivity to the confounding effects of homophily.

The findings of Shalizi and Thomas remain relevant even if we were to abandon GC in favor of a more principled framework for causal inference. Asking if a change in the behavior of a is responsible for change in the behavior of b is fundamentally a question of causality. A direct test for causality requires counterfactual evidence: how would b have behaved if the behavior of a had not changed? Since this kind of information is rarely available, one usually resorts to an experimental design that yields approximate counterfactual evidence. For example, in a randomized controlled trial, a homogeneous group of participants is split randomly into a test and a control group. Members of the test group are subjected to an active intervention, while the control group provides the counterfactual information. Under certain circumstances, causal inference is possible from observational data alone, without requiring experimental intervention. However, Shalizi and Thomas show that as long as the detection of social influence is cast as a problem of causal inference from observational data, it will always suffer from the confounding effects of homophily. Their findings are in accordance with the earlier verdict of Kenny (1988): "Causal priority issues, which are difficult to resolve in other disciplines, are virtually insoluble in the study of two-person relationships."

6.1.2 Evaluation Strategy

Two arguments might lead us out of the dilemma posed by causal non-identifiability in the influence network recovery setting: First, rather than as methods of causal identification, we might understand Granger-causal approaches as tools for ruling out specific classes of non-causality, and accept the resulting systematic error that goes beyond the statistical uncertainty imposed by the limited sample size. Second, one might ask if true causal identification is actually necessary for the task at hand. After observing a sufficiently large part of a social network, constructing a set of *plausible* influence relationships, that appear causal to a hu-

man judge, but have not been controlled for confounding, may be sufficient for identifying relevant phenomena on a higher level, e.g., attempts to manipulate the public opinion.

Either way, since causal identification is principally unreliable in the influence network recovery setting, it becomes necessary to assess the accuracy that can be obtained on actual social media data. Influence network recovery can be viewed as an information retrieval problem. From a large set of candidate edges, a comparatively small subset of influence relationships is to be identified by a binary classifier. The performance of said classifier can be fully characterized in terms of its positive predictive value (PPV) and true positive rate (TPR), also known as precision and recall. Both are independent of the number of negative examples, which represents the assumption that an arbitrary number of unobserved additional negative examples exist. Influence network recovery can also be compared to medical screening, where a large at-risk group is tested for a particular disease. Screening procedures are usually evaluated in controlled trials, a setting where the prevalence of positive examples is known in advance. In that case, a characterization in terms of TPR and true negative rate (TNR), which are independent of the prevalence, is preferred. Both approaches require an annotated reference dataset.

Among studies with a setting similar to influence network recovery, we note the lack of a common evaluation methodology and in particular the absence of a standardized dataset for evaluation. Obtaining ground truth via manual annotation is difficult: When presented with two sequences of messages and no further information about the relationship of the authors, an annotator can only assess the content-wise similarity of the messages over time, and is thus subject to the same confounding effects as a GC test. Since short-term influence effects, as analyzed in this study, may be at least in part caused by unconscious mental processes, even interviewing the actors about the reasons for their behavior is not guaranteed to uncover the true causal relationships. However, background knowledge and awareness of the social context enable a human observer to identify certain types of external confounders. An annotator can also distinguish influence from unrelated short-term phenomena such as conversational back-and-forth, and is able to assess the topics of a conversation more accurately than a statistical topic model.

Consider, for example, person a, who is interested in soccer and regularly tweets about the matches of a particular club. On game day, a listens to a live radio broadcast at noon and tweets a running commentary. Person b is not much of a fan, and has not tweeted about soccer before, but hears about the game on the evening TV news and decides to comment. From the perspective of GC, it looks as if a influenced b, while a human observer is likely to identify the media coverage as a common confounder in the absence of information about b's way of exposure to the game. Further discussion on the effect of exposure to the same information sources can be found in the study of Bakshy et al. (2012).

While we have reason to believe that human annotation produces meaningful reference data, the cost of annotation is high. Reading and evaluating all tweets of a pair of users (on average 150 tweets / month in our dataset) is time-consuming. The lack of objective, dichotomous criteria for the presence of social influence makes us expect high uncertainty in the decision process, which manifests as low inter-annotator agreement. Furthermore, we expect social influence to be rare, even among the heuristically selected set of candidate edges. When limiting the annotation to a sample of the whole dataset, it is likely that the

sample size would have to be chosen unfeasibly high to obtain a sufficient amount of positive examples.

In consequence we only perform minimal annotation. Given multiple sets of candidate edges and multiple classifiers, the performance of each combination is to be evaluated. For each candidate set, we annotate a random sample of size N^- under the expectation that most or all annotations will be negative. These annotations let us estimate the TNR of each classifier and provide us with a loose upper bound on prevalence. In the case of the Twitter dataset, annotation of $N^- = 50$ edges for each network type yielded no positive examples, which places an upper bound of 0.02 on the prevalence of influence edges. If a particular classifier is found to perform well according to TNR, one can then annotate a sample of edges classified as positive to estimate its PPV, and obtain a lower bound on TPR via Bayes' theorem:

$$\frac{(1 - \mathrm{TNR}) \cdot (N^- - 1) \cdot \mathrm{PPV}}{1 - \mathrm{PPV}} \leq \mathrm{TPR} \leq 1$$

This evaluation scheme is complemented by two heuristic tests, which do not require annotated reference data. The first test can be viewed as an approximation of the PPV. For a given classifier, one can compute the positive classification rate with respect to the candidate set. The candidate set consists of edges with a high prior probability of influence, so for purposes of comparison we construct an equally sized set of edges with low prior probability. Since each candidate set is derived from a social network graph, a suitable set can be constructed by random sampling without replacement from its complement graph. If c^+ and c^- are the positive classification rate of the high and low probability set, respectively, the ratio $r = c^+/(c^+ + c^-)$ exceeds 0.5 if the classifier finds more influence edges within the high probability set, as expected.

The second test makes use of the fact that all classifiers have the same underlying model of social influence as a Granger-causal process, so the classification results should exhibit a certain degree of consistency, which can be established by computing the pairwise similarity of the sets of positively classified edges. One measure of similarity is the Jaccard index, which for two sets A and B is defined as the ratio $|A \cap B|/(|A| + |B| - |A \cap B|)$. Since the Jaccard index penalizes pairs of classifiers proportional to the difference in their positive classification rates, it is most informative when comparing classifiers with similar rates. Even if two classifiers have highly different rates of positive classification, they may still exhibit consistency in terms of acting as successively narrower filters, i.e., the set of positive edges of one classifier is a subset of the other's. We therefore also compute the Szymkiewicz-Simpson overlap coefficient, defined as $|A \cap B|/\min(|A|, |B|)$. An overlap of 1 indicates that the smaller set is an exact subset of the larger.

6.2 Implementation of Influence Network Recovery

As established earlier, influence network recovery requires a set of candidate edges with a high *a priori* probability of social influence. In the following experiments, we compare the suitability of mutual following and mutual interaction as criteria for candidate selection. For the 30 000 user subset considered in the following experiment, the mutual follower graph

contains 1 544 553 and the mutual reply graph 37 444 edges. If one considers influential individuals and, by extension, influence relationships to be rare, even mutual communication is a rather optimistic prior.

Topic distributions that represent the behavior of a user in an interval of time are obtained from a topic model as described in section 5.2. If a user has not authored any tweets within an interval, the absence of data is recorded in the time series. For an interval length of 2 days we observe 53.8% of missing data per user on average (standard deviation $\sigma = 36.7\%$), for an interval length of 7 days 37% of data is missing ($\sigma = 39.9\%$). The amount of missing data, even for longer interval sizes, indicates that many users take longer breaks from social media. When shortening the interval to less than two days we expect a higher rate of gaps due to day / night rhythms.

Influence network recovery can be implemented in terms of the similarity-based measure of influence strength $\bar{I}_{a\to b}$ as well as by straightforward application of Granger causality. Either approach has a number of implementation details that require further discussion.

6.2.1 Influence Measurement

Making a decision about the presence or absence of influence along a candidate edge (a, b) according to the measured influence strength $\bar{I}_{a\to b}$ requires a frame of reference. In particular, we want to test whether a non-zero value actually corresponds to the presence of influence, or can be explained by alternative hypotheses such as random coincident changes in behavior. The method of surrogate data (Schreiber and Schmitz, 2000) (also known as "randomization testing"; La Fond and Neville, 2010) provides a Monte Carlo (MC) framework for formulating and testing such hypotheses. From each observed time series, we derive N synthetic series that are indistinguishable from the original except for the absence of the quality of interest, in this case social influence. The process for generating synthetic data is the MC equivalent to an analytic null hypothesis. If the influence strength of e synthetic series exceeds the influence strength of the (single) real series under consideration, one can estimate $p = (e + 1)/(N + 1)$.

The attainable accuracy of estimation is limited by the available computational resources: Individual iterations of the MC test may be expensive, and the number of iterations required to reach the required numeric precision and confidence in the accuracy of estimation of p increases super-linearly as α approaches zero. Gandy and Hahn (2014) describe a method for stopping early, once there is sufficient confidence to not reject the null hypothesis. If a set of hypotheses H is to be tested, subject to a fixed overall "iteration budget", then early stopping will allocate a larger part of that budget to the borderline-significant cases that benefit the most from additional iterations. In the following experiments we choose $N = 2\,000$ and use the early-stopping procedure with an overall budget of $N \cdot |H|$.

The *topic drift* null hypothesis states that the topic distributions of influencer and influencee coincidentally changed in a way that resulted in a high observed influence strength. Synthetic data is generated implicitly: For computation of $I_{a\to b}(t)$, $\theta^{b,t+1}$ (if not missing) is resampled from a Dirichlet distribution with the same concentration parameter α as used by the topic model. It is reset to its original value for subsequent computation of $I_{a\to b}(t+1)$. Figure 6.2 illustrates this process.

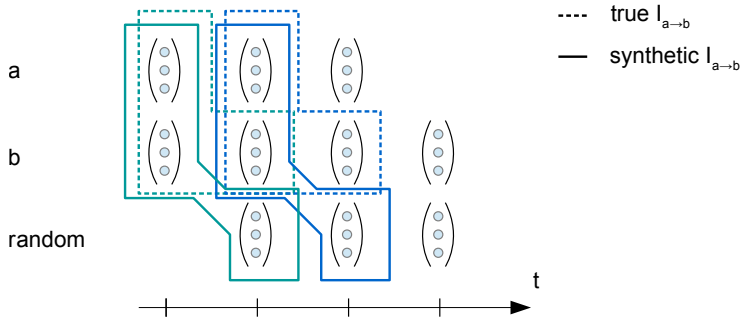


Figure 6.2: Generation of synthetic data for the topic drift null hypothesis

Temporal permutation (Anagnostopoulos et al., 2008; Bossomaier et al., 2013) is a common method for perturbation of data so that influence relationships are weakened, but other characteristics are preserved. Temporal permutation of the influencer's time series weakens Granger-causal effects by changing the temporal order of cause and effect. During permutation, missing data is kept in place. If due to missing data $I_{a\to b}(t)$ cannot be computed for at least six time intervals, there are not enough possible permutations to sample from, and we choose to fail to reject the null hypothesis.

Through the combination of topic drift test and temporal permutation test, we aim to filter out spurious influence effects of low magnitude. When testing this kind of composite hypothesis, where non-rejection of a partial hypothesis does not reject the null hypothesis, correction for multiple testing is not necessary. To address concerns about multiple testing on the network level, we perform a node-wise correction of the significance level as suggested by Bossomaier et al. (2013): Controlling the false positive rate for the entire social network would require the adjustment of α, and consequently the increase in false negatives, to be proportional to the number of candidate edges. To avoid this issue, we apply the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to the set of hypothesis tests for the incoming edges of each actor. The magnitude of correction is thus proportional to the actor's in-degree, which in real-world social networks is bounded above by the limited capacity of human actors to concurrently maintain social relationships (Gonçalves et al., 2011).

Conceptually, we treat the MC procedure as a decision mechanism with a result that is to be evaluated against ground truth, rather than as an actual hypothesis test, where the result would stand on its own. Adjusting α moves the decision boundary, i.e., lower α will increase precision at the expense of recall. To enable a comparative evaluation of all methods of influence detection, we use a fixed $\alpha = 0.05$ for the MC tests and the GC hypothesis tests discussed in the following section.

In addition to the hypothesis test, we evaluate a less complex alternative procedure for binarizing $\bar{I}_{a\to b}$. As described in section 6.1.2, given a social network graph, one can construct a set of random edges with low prior probability of influence. A fixed threshold τ can be derived from the distribution of influence strength $\bar{I}_{a\to b}$ of these edges. We choose $\tau = \text{mean} + 3 \cdot \text{SD}$ to account for the possible presence of a low amount of true influence edges in the set. Separate random graphs are constructed for computation of the threshold

and evaluation of the classifier.

6.2.2 Testing for Granger Causality

In analogy to influence measurement, we test for GC from θ^a to $\theta^{b,+}$, which involves fitting a VAR model to the time series. The VAR model requires stationary data, so the topic distributions should be time-invariant in the sense that they do not contain trends or seasonal variation, or the model may detect spurious interactions among the variables (Seth, 2010). A subset of non-stationarity caused by the presence of unit roots can be eliminated by transforming a time series X to a series X' of first differences, so that $X'_t = X_t - X_{t-1}$. The extraction of additive change from a series of topic distributions can therefore be interpreted as a means of improving stationarity.

Before GC can be computed, missing values have to be eliminated. From the paired time series of influencer and influencee, we extract the longest sub-sequence in which neither series is missing any values, with the exception of the last time step, for which the topic distribution of the influencer is allowed to be missing. For an interval length of 2 days, the average proportion of remaining observations per edge is 17.5% (standard deviation σ = 23%). For an interval length of 7 days the average is 48.4% (σ = 40.3%). To be consistent with the procedure for influence measurement, we exclude edges from the experiment if no more than six consecutive paired observations are present.

GC is equivalent to its non-linear generalization TE for variables that follow a Gaussian distribution (Barnett et al., 2009). The time series under investigation are made up from topic distributions θ that are Dirichlet-distributed per specification of the topic model. A Dirichlet distribution can be approximated by a logit-normal distribution (Aitchison and Shen, 1980)[1], which is a transformation of the multivariate normal distribution onto the unit simplex, so a Dirichlet distribution that is subjected to the inverse transformation is approximately normal (Srivastava and Sutton, 2017, section 3.2). The inverse transformation $\mathbb{R}^K \rightarrow \mathbb{R}^{K-1}$ is the additive log-ratio (ALR) transformation $y_i = \log(x_i/x_d)$ with $i \in [1, K-1], d = K$. While the transformation itself depends on the choice of d, Brunsdon and Smith (1998) show that the parameters of a time series model fit to transformed data are invariant up to permutation.

Since the ALR transformation requires all values to be non-zero, we first remove all components that are constant over time and therefore do not have any effect on the parameters of the time series model. However, this case is rare, with only 0.26% of rows across all experiments being time-constant. The remaining zeros are replaced with small positive values according to the scheme of Martín-Fernández et al. (2003). While the components of topic distributions obtained from ART are never exactly zero due to the regularizing effect of the prior, zeros can arise from the transformation of a time series. In the Twitter dataset, a series of additive change θ^+ contains 38% of zeros on average, indicating that the behavior of users is often rather static over time. We subject all observations to the ALR transformation, maximizing the effects visible to the linear VAR model, and therefore minimizing the gap in expressivity between GC and the more computationally expensive TE. The gap could be

[1] However, the authors do not provide empirical results for distributions with concentration parameter $\alpha \ll 1$ that commonly arise in the context of topic modeling.

closed completely by using a topic model that generates logistic-normal topic distributions, e.g., the Correlated Topic Model (Blei and Lafferty, 2006).

After these preprocessing steps, a VAR(1) model is fit to the time series data, matching $\bar{I}_{a\to b}$'s implicit Markov horizon of one. Due to the high dimensionality of the topic distributions, the number of parameters p exceeds the number of observations n, so parameter estimation via ordinary least squares is not possible. However, the hyper-parameters of the topic model have been chosen so that the topic distributions are sparse, and Ver Steeg and Galstyan (2013) show that this sparsity can be exploited to facilitate the estimation of TE, so we expect the same to be true for GC. The traditional GC hypothesis test involves fitting two separate VAR models, one to the column-wise concatenation of the time series of influencer and influencee ("full model"), the other to the time series of the influencee only ("restricted model"), and comparing the residuals. ℓ_1-regularized least squares methods, such as LASSO, can fit a VAR model to high-dimensional, sparse data. However, in that case the restricted model is no longer guaranteed to be nested, that is, to be a sub-model of the full model, which is a prerequisite of the traditional hypothesis test.

An alternative test for GC (Chaudhry et al., 2017; Basu et al., 2017) uses the de-biased LASSO of Javanmard and Montanari (2014) for fitting the full model. The set of coefficients of this model can be partitioned according to the source and destination of the effect they represent: from the past of the influencer or influencee on the future of the influencer or influencee. Individual hypothesis tests are performed for the subset of coefficients that represent an effect from influencer to influencee. If at least one such coefficient significantly differs from zero, a Granger-causal relationship is established. The fitting of a VAR model can be decomposed into separate least squares estimation tasks for each component of the response variable (Lütkepohl, 2005, p. 72), so the computational effort can be halved by limiting the estimation to the components of the influencee.

Testing multiple related hypotheses necessitates control of the family-wise error rate (FWER). Since we only distinguish between the non-rejection of all null hypotheses and the rejection of at least one null hypothesis, the FWER is equivalent to the false discovery rate (FDR). Chaudhry et al. (2017) note that the Benjamini-Hochberg procedure is not applicable in this setting and propose a custom FDR correction scheme[2], which we use in the following experiments.

Finally, we define a score for ranking edges by magnitude of influence. If c_a is the sum of absolute coefficients from influencer to influencee, and c_b the sum of absolute coefficients from influencee to influencee, then the score $M_{a\to b} = c_a/(c_a + c_b)$ or 0, if both c_a and c_b are zero. Like $\bar{I}_{a\to b}$, the score can be interpreted as relative confidence within the set of candidates, but not as a probability of accurate prediction.

Robustness to Violation of Modeling Assumptions

The procedure for detecting Granger causality in time series of topic distributions is composed of a number of individual statistical methods. Each method makes assumptions about the nature of the input data, and the assumptions that are likely to be violated by topical representations of social media activity turn out to be closely related. Testing for GC involves

[2] An equivalent scheme was independently developed by Javanmard and Javadi (2019).

fitting a VAR model, which can be expressed as repeated linear regression. Regardless of the dimensionality of the problem and the presence or absence of ℓ_1 regularization, linear regression is sensitive to strong correlation among the variables that make up the design matrix X (Raninen and Ollila, 2017). In the context of our experiment, these variables are the univariate time series that make up the vector time series.

In the low-dimensional case ($p \leq n$, usually $p \ll n$), solving the least squares parameter fitting problem involves inverting $X^T X$, which is a Gramian matrix. If the variables in X are mean-centered, the Gram matrix is approximately proportional to the covariance matrix. It is positive definite if and only if the variables are linearly independent. In consequence, if X contains strongly correlated variables, the parameter fitting problem is ill-conditioned, and if some variables are collinear, it does not have a unique solution. In the high-dimensional case ($p > n$), a similar result can be obtained for sparse design matrices in ℓ_1-regularized linear regression via the restricted eigenvalue condition (Bickel et al., 2009). Here, collinearity is a sufficient, but not necessary condition for the non-uniqueness of the solution.

The de-biasing procedure of Javanmard and Montanari operates under the assumption that the restricted eigenvalue condition holds, and is therefore sensitive to strong correlation. To obtain the initial, biased parameter estimates, the procedure uses the square-root LASSO (Belloni et al., 2011), which can be implemented in a computationally efficient way via proximal optimization. Since the optimization problem posed by the LASSO is neither strongly convex nor smooth, Li et al. (2020) propose a path-wise optimization scheme to ensure that the parameters remain within a region of local strong convexity and smoothness throughout the optimization process, given that the restricted eigenvalue condition holds. This makes proximal first and second order methods applicable, with first order methods being slower in convergence, but more robust to strongly correlated variables, which may cause the Hessian of the optimization problem to not be positively definite. We therefore use proximal gradient descent, which has a negligible error rate: The optimizer fails to converge for 0.02% of GC tests across all experiments.

The propensity for linear dependency among the components of the time series of topic distributions, sometimes taking the shape of constancy over time, can be in part attributed to a particular property of LDA-style topic models. According to the generative model specification, topic distributions, or more precisely, the parameter vectors θ of multinomial distributions, are drawn from a Dirichlet distribution, and therefore their components take on continuous real values. When performing parameter estimation via collapsed Gibbs sampling, θ is estimated from the number of words that have been assigned to a particular topic. It follows that θ can only be treated as drawn from a continuous distribution in the asymptotic case of the overall number of words approaching infinity. In practice, the volume of textual material may be low (user does not post often, size limitation of messages) or limited by division of the observation period into short intervals. In these cases, θ is practically discrete, in the sense that each component can only take on a finite number of different values. This directly increases the likelihood of linear dependence among components. A secondary effect is that, even in the absence of collinearity, the solution of the LASSO estimator is no longer unique with probability one if the variables are not continuous (Tibshirani, 2013).

If there are multiple solutions, we have to assume that the optimization will converge to an arbitrary one. Yet, GC can only be said to be present if all solutions assign a non-zero

coefficient to at least one component of the topic distribution of the influencer. If there is at least one solution that does not involve the influencer, the behavior of the influencee can be explained equally well (in terms of least-squares error) using only data from its own past, so the influencer does not add any new information. Tibshirani (2013, section 4.2) describes a computationally inexpensive algorithm that classifies a variable as *indispensable*, if it has a non-zero coefficient in every solution, or *dispensable* otherwise. At least one variable of the influencer being indispensable implies the presence of GC, but non-existence of an indispensable influencer variable does not imply the absence of GC, as the set of solutions could be partitioned so that each subset has a different active influencer variable. Conversely, one cannot define absence of GC as dispensability of all influencer variables. A test for GC that is robust to multiple solutions has to enumerate all solutions, which may not be computationally tractable (cf. Tibshirani, 2013, appendix A.3).

In consequence, there are circumstances under which GC cannot be reliably tested for. In the absence of theoretical or experimental results on the effect of strong correlation on the de-biasing of the LASSO estimates, we cannot preclude that it may cause spurious classification of coefficients as significant, which may result in spurious detection of GC. The existence of multiple LASSO solutions has a similar effect on the detection of GC: In the absence of all other sources of error, if there are multiple solutions and GC is present, all solutions consistently assign a non-zero coefficient to at least one variable of the influencer. Therefore, the presence of multiple solutions may result in a false positive, but never a false negative detection.

A principled solution to the problems of correlation and non-uniqueness can be found in the sparse group LASSO (Simon et al., 2013), which prefers a solution where all weight mass is concentrated in one group of coefficients. Groups can be defined arbitrarily; when testing for Granger causality, variables of the influencer would be assigned to one group, and variables of the influencee to another. Although grouping the coefficients does not guarantee uniqueness of the solution (Roth and Fischer, 2008), Meinshausen (2015) shows that testing a hypothesis over a group of variables requires weaker assumptions than testing individual variables. Most notably, the contribution of a group of variables can be detected even if the variables within the group are strongly correlated. However, there is no existing framework for group hypothesis testing that can be directly applied to GC testing. The proposed test of Meinshausen uses a *data splitting* strategy, which requires assumptions about the distribution of the true coefficients that cannot be reasonably assumed to hold in the GC case (Dezeure et al., 2015). Mitra and Zhang (2016) discuss de-biasing the group square-root LASSO, but only its non-sparse variant, which does not allow for sparsity within a group (compare Stucky and van de Geer, 2017, sections 4.2 and 4.4). Stucky and van de Geer (2018) build a more general theoretical framework for de-biasing LASSO variants, including the sparse group LASSO, but do not provide a generally usable implementation.

The complexity of building a robust test for Granger causality, which can be largely attributed to the modeling assumptions that are required to make linear regression work in the high-dimensional regime, is an argument in favor of the model-free transfer entropy. While transfer entropy is equivalent to GC in the case of Gaussian variables, in the sense that the TE of two variables is zero if and only if there is no Granger-causal relationship, there is no known significance test for TE estimates (Barnett et al., 2009). McKenney and

White (2017) apply TE to the problem of influence network recovery, but their approach for deriving a suitable threshold requires information about the true influence network.

For the following experiments, we choose a more pragmatic approach, and address the problem at the stage of fitting the topic model: When estimating a topic distribution from n posterior samples, increasing n not only improves the accuracy of estimation, but also reduces the likelihood of collinearity by enlarging the range of possible values of each component (see section 4.1.1, "Estimating the Parameters of the Document-Topic and Topic-Word Distributions"). However, the computational cost of estimation grows linearly with n, and correlation can only be reduced, not completely eliminated. A possible alternative might be dequantization by adding noise, which is successfully applied in a different case of fitting a continuous model to discrete data (Theis et al., 2016).

6.3 Experimental Evaluation

The methods of influence network recovery discussed in the previous section have been subjected to preliminary tests on synthetic time series exhibiting either strong influence (b's actions are a lagged copy of a's, i.e., the behavior of b is completely determined by a) or no influence at all (b's actions are independent of a). Each method is able to correctly classify these synthetic examples.

To evaluate the performance on real data, we perform influence network recovery on the 30 000 user subset of the Twitter dataset with varying experiment parameters and compare the results. The parameters and their possible values are:

Method The two basic methods of influence network recovery, measurement of influence and testing for Granger causality, can be subdivided into four variants: The influence strength $\bar{I}_{a\to b}$ can either be tested for statistical significance (*M+S*) or subjected to network-wide thresholding (*M+T*). To ensure comparability of influence measurement and GC, we always use the average influence strength (equation 6.2) instead of a weighting scheme that introduces temporal decay (equation 6.3).

All three variants, *M+S*, *M+T*, and *GC*, can be modified to use the additive change θ^+ in place of the observed behavior θ of the influencer (*+δ*). As discussed in section 6.1, this effectively replaces the model assumption that influence is constant over time with the assumption that influence is strongest at the time of a change in behavior. In analogy to biological neurons, we use the term *spiking influence* for this kind of model. In total there are six variants.

Candidate network We compare two criteria for choosing candidate edges with high prior probability of influence: mutual following (*mF*) and mutual replies (*mR*).

Time interval length For transformation into time-quantized series of topic distributions, messages are aggregated over intervals of equal length. For this experiment we consider intervals of 2 and 7 days.

Table 6.1: Results of the influence network recovery experiments (constant influence; Hauffa et al., 2019)

experiment								
method	net.	int.	rejected	pos. class.	c^+	c^-	r	est. TNR
M+S	mF	2	44.11%	37	< 0.1%	0%	1	~ 1
M+S	mF	7	51.97%	10	< 0.1%	0%	1	~ 1
M+S	mR	2	7.13%	61	0.2%	< 0.1%	0.924	~ 1
M+S	mR	7	18.91%	9	< 0.1%	< 0.1%	0.900	~ 1
M+T	mF	2	24.71%	210 710	13.6%	2.3%	0.856	0.820
M+T	mF	7	21.91%	150 326	9.7%	1.7%	0.851	0.899
M+T	mR	2	0.67%	19 284	51.5%	2.3%	0.957	0.755*
M+T	mR	7	0.47%	8 869	23.7%	1.6%	0.936	0.906*
GC	mF	2	64.94%	244 004	15.8%	7.5%	0.679	0.831
GC	mF	7	55.05%	618 604	40.1%	18.1%	0.689	0.674*
GC	mR	2	25.26%	12 243	32.7%	11.3%	0.743	0.792*
GC	mR	7	21.92%	26 591	71.0%	19.0%	0.789	0.642*

Table 6.2: Results of the influence network recovery experiments (spiking influence)

experiment								
method	net.	int.	rejected	pos. class.	c^+	c^-	r	est. TNR
M+S+δ	mF	2	52.29%	35	< 0.1%	0%	1	1
M+S+δ	mF	7	58.81%	5	< 0.1%	0%	1	1
M+S+δ	mR	2	12.35%	61	0.2%	< 0.1%	0.884	1
M+S+δ	mR	7	27.20%	3	0.1%	< 0.1%	0.750	1
M+T+δ	mF	2	31.30%	212 157	13.7%	2.6%	0.840	0.854
M+T+δ	mF	7	27.02%	135 456	8.8%	2.0%	0.810	0.899
M+T+δ	mR	2	1.64%	17 671	47.2%	2.7%	0.946	0.792*
M+T+δ	mR	7	1.23%	6 897	18.4%	2.1%	0.898	0.906*
GC+δ	mF	2	69.65%	280 873	18.2%	7.9%	0.697	0.775
GC+δ	mF	7	61.06%	600 148	38.9%	16.8%	0.698	0.685*
GC+δ	mR	2	31.67%	15 245	40.7%	10.2%	0.800	0.774*
GC+δ	mR	7	30.63%	25 922	69.2%	16.5%	0.808	0.660*

6.3.1 Results

The results of these experiments are collected in tables 6.1 (constant influence) and 6.2 (spiking influence). The early rejection rate is the proportion of candidate edges that could not be tested for influence because of missing data. It is noticeably higher for edges from the mutual follower graph. This is particularly visible in the case of *M+T*, where an edge is only rejected if the pair has no overlapping periods of activity at all. The follower network appears to contain a substantial amount of relationships that are either characterized by passive consumption of content rather than active conversation, or by disjoint bursts of activity. In the case of *M+S*, across all experiments, the rejection rate is higher for longer intervals. This is because coarser quantization yields shorter sequences, while the required minimum length stays the same. In addition to imposing a natural upper limit on the interval length, this rejection strategy could result in bias against strong influencers who rarely interact with others.

The positive classification rate on the candidate set, which contains edges that are believed to have a high prior probability of presence of influence, is denoted as c^+, while c^- is the positive classification rate among edges with low prior probability. The generally higher c^+ of experiments on the mutual reply network is consistent with the assumption that interaction is a better indicator of influence than passive exposure. In the case of influence measurement, shorter intervals result in a higher positive classification rate, which may point to short-term effects that are rendered invisible by aggregation over longer intervals, but the opposite can be observed for GC. We suspect that GC testing is particularly sensitive to sample size, and the low number of observations in the case of 7 day intervals causes a higher rate of false positives, reflected by a higher c^-. Non-stationarity of a time series can also cause the erroneous detection of GC, but the variant *GC+δ*, which is designed to improve stationarity, does not exhibit a consistently lower c^+. Similarly, the variant *GC* is theoretically sensitive to a broader class of influence phenomena than *GC+δ*, given that the former operates on the original observations and the latter only on the additive change, but this is not reflected by c^+.

The consistently low c^- and high r of the *M+T* experiments shows that a threshold derived from a randomly constructed set of likely to be non-influential edges generalizes, at least to other sets of the same construction. However, both the absolute number of positive classifications and c^+ show that significance tests act as a much stronger filter than thresholding. With the exception of one experiment, *GC*, *M+T*, and *M+S* are successively more selective. The estimate of TNR shows that all three methods to some extent agree with the human annotator. However, TNR becomes less useful with lower rates of positive classification, given that an optimal TNR can be trivially achieved by rejecting every edge. The expected TNR when selecting $c^+ \cdot |E|$ edges randomly is $1 - c^+$, so TNR has no discriminative power for *M+S*, where c^+ is consistently low. In tables 6.1 and 6.2, the estimated TNR is marked with an asterisk if it outperforms the expected TNR of a random classifier by a margin greater than the resolution of estimation (using 0.02 as a conservative estimate). Comparing the constant and spiking influence settings, spiking GC performs slightly better in terms of c^- and *TNR*, but apart from that there are no discernible systematic differences.

Table 6.3 compares the positive classification rates c^+ and c^- of *M+S* to those of modified

Table 6.3: Positive classification rate of *M+S* by choice of hypothesis

experiment		both			topic drift only			permutation only		
net.	int.	abs.	c^+	c^-	abs.	c^+	c^-	abs.	c^+	c^-
mF	2	37	< 0.1%	0%	135 445	8.8%	1.1%	18 113	1.2%	0.6%
mF	7	10	< 0.1%	0%	57 486	3.7%	0.5%	12 759	0.8%	0.4%
mR	2	61	0.2%	< 0.1%	7 570	20.2%	1.1%	639	1.7%	0.5%
mR	7	9	< 0.1%	< 0.1%	1 621	4.3%	0.5%	437	1.2%	0.4%

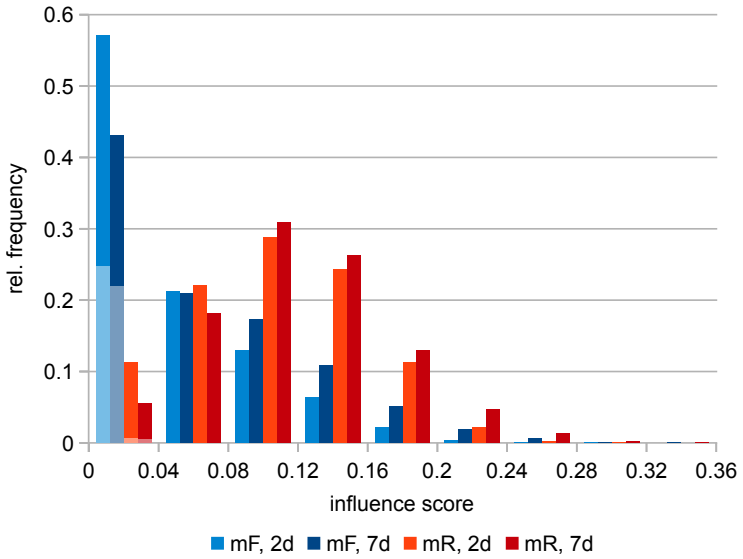
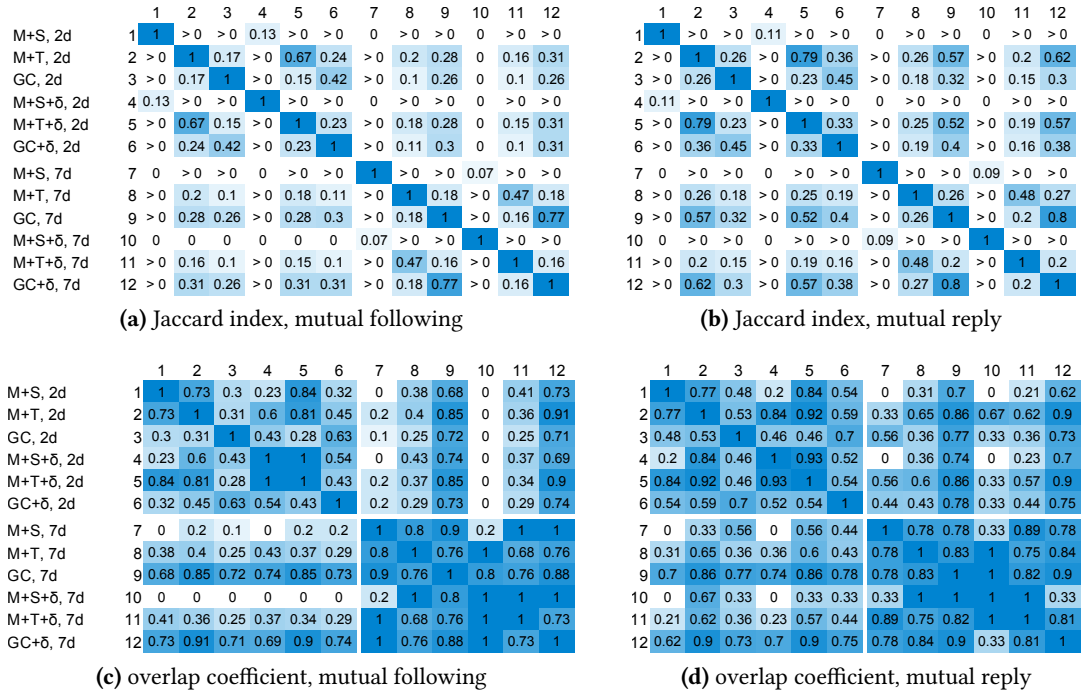


Figure 6.3: Empirical distribution of $\bar{I}_{a\to b}$ (Hauffa et al., 2019)

variants that only test one of the two original hypotheses. When applied individually, topic drift and permutation tests have a much higher positive classification rate than combined, which implies that the sets of edges classified as positive by each test have little overlap. This confirms that the tests represent distinct null hypotheses. Since combining the tests leads to a lower c^-, each null hypothesis individually contributes to the characterization of social influence.

The empirical distribution of $\bar{I}_{a\to b}$ provides further insight into the properties of our definition of influence strength. Figure 6.3 compares the histograms of the distributions arising from the four *M+S* experiments. The proportion of exact zeroes in the first bin is indicated by use of a lighter shade. In each of the four cases, a goodness-of-fit test using the MC procedure described by Clauset et al. (2009, section 4.1) shows that the data is appropriately described by a unimodal Beta distribution (250 iterations, $p < 0.01$). There is no obvious clustering into groups of edges exhibiting high and low magnitude of influence, which highlights the need for a decision procedure that separates low-magnitude noise from potential influence



(a) Jaccard index, mutual following

(b) Jaccard index, mutual reply

(c) overlap coefficient, mutual following

(d) overlap coefficient, mutual reply

Figure 6.4: Jaccard index and overlap coefficient of experiment pairs

on a per-edge basis.

Figure 6.4 consists of heat map visualizations of Jaccard index and overlap coefficient. Experiments with different candidate networks are not directly comparable, so we provide a separate chart for each network. There are no visible structural differences between the mutual following and reply network. As expected, the Jaccard index largely reflects the relative selectiveness of the classifiers, with *GC* and *M+T* as the least selective being somewhat similar to each other. While *M+T* and *GC* exhibit consistency across different interval sizes, the result sets of *M+S* are completely disjoint. Each of the three main methods *M+S*, *M+T*, and *GC* exhibits some consistency with respect to transformation of the time series (constant vs. spiking). Visualization of the overlap coefficient confirms the existence of a hierarchy of successive refinement, but only within the individual interval sizes. There are two clusters: *M+S* refines *M+T* with and without transformation and for each interval length, but the results of *M+S* are completely disjoint across different interval lengths. *M+T* refines *GC* across all experiment variants, but to a lesser degree.

In summary, all three methods of influence network recovery can detect influence effects in completely synthetic data, but the high cost of manual annotation makes it difficult to assess to what extent this capability generalizes to real data. Via estimation of TNR, we show, as far as it can be supported by our limited resources, that *GC* and *M+T* are successively more consistent with the human annotation. The estimation does not provide meaningful results for methods with a low positive classification rate, most notably *M+S*. However, given *M+S*'s high degree of overlap with the less selective *M+T*, its true TNR is unlikely to be lower.

6.3.2 Follow-up Experiments

The inconclusive nature of the preceding experiments motivates some additional investigation. Considering the theoretical shortcomings of Granger-causal inference, particularly the sensitivity to external confounders, the low positive classification rate of *M+S* is unexpected. To learn about the factors that contribute to the classification decision, the topic drift test can be divided into its components, which can then be analyzed separately. As described in section 6.1, given topic distributions θ^t, θ^{t+1} that represent behavior within successive intervals of time, the additive change in behavior can be expressed as the product of direction and (scalar) magnitude of change. The influence strength $\bar{I}_{a\to b}$ is defined as the product of the influencee's magnitude of change and the similarity of the influencer's past behavior and the direction of the influencee's change. Change in behavior that is intrinsically motivated, i.e., not effected by social influence, appears as random change to an observer. In the MC test for topic drift, synthetic future topic distributions θ^{t+1} of the influencee are drawn from an unconstrained Dirichlet distribution and may therefore vary arbitrarily in both direction and magnitude of additive change. If the influence strength of the synthetic time series exceeds the influence strength of the real data sufficiently often, we fail to reject the null hypothesis that the measured influence is indistinguishable from the effects of random variation in magnitude and direction.

More specific null hypotheses can be derived by placing constraints on the synthetic future topic distribution of the influencee. To learn about the respective contributions of magnitude and direction to the result of the hypothesis test, we fix the magnitude to a constant value. The magnitude m and the topic distributions θ^t, θ^{t+1} are related as follows:

$$
\begin{aligned}
m &= \|(\max\{0, \theta_1^{t+1} - \theta_1^t\}, \ldots, \max\{0, \theta_K^{t+1} - \theta_K^t\})\|_1 \\
&= \sum_{i=1}^{K} \max\{0, \theta_i^{t+1} - \theta_i^t\} \\
&= \frac{1}{2} \sum_{i=1}^{K} |\theta_i^{t+1} - \theta_i^t| \qquad \left(\text{since } \sum_{i=1}^{K} \theta_i = 1\right) \\
\Leftrightarrow 2m &= \|\theta^{t+1} - \theta^t\|_1
\end{aligned}
$$

When sampling from a distribution supported on the unit simplex, for any given $\theta^t \in \Delta^{K-1}$ and m with $0 \leq m \leq 1 - \min_{i=1}^{K} \theta_i^t$, fixing m is equivalent to constraining the support to the set of points at an ℓ_1-distance of $2m$ from θ^t. A sample θ^{t+1} from a Dirichlet distribution subject to this constraint can be generated via Gibbs sampling. The derivation of the Gibbs sampler is provided in appendix B. We investigate two strategies for choosing the value of m in the generation of synthetic data for the hypothesis test: If m is chosen to match the observed magnitude, only the direction is left open to random variation. The resulting null hypothesis is that the measured influence is indistinguishable from random variation in direction. Alternatively, one can fit a Beta distribution to the observed magnitudes and draw m from that distribution, thus restricting the null hypothesis to random, but typical variation in magnitude.

We repeat the set of *M+S* experiments for each of the three variants of the topic drift null

Table 6.4: Positive classification rate of *M+S* for different variants of the topic drift test

experiment		unconstrained		typical magnitude		fixed magnitude	
net.	int.	c^+	c^-	c^+	c^-	c^+	c^-
mF	2	9.2%	1.1%	54.8%	27.8%	56.4%	27.8%
mF	7	4.8%	0.6%	45.7%	20.5%	51.8%	23.1%
mR	2	23.7%	1.0%	95.8%	28.0%	96.1%	28.0%
mR	7	7.0%	0.6%	87.1%	20.5%	91.7%	22.9%

hypothesis. The permutation test is skipped, so that the results directly reflect the outcome of the topic drift test. The results are given in table 6.4. Both of the more specific null hypotheses are much easier to reject. In other words, replacing the future topic distribution of the influencee with a random sample with the same magnitude but different direction rarely leads to a higher influence strength $\bar{I}_{a\to b}$. On one hand, this means that random change in behavior (uncorrelated with the behavior of others, i.e., intrinsically motivated) is unlikely to be detected as spurious influence. On the other hand, the time-lagged similarity of the directions of change, on its own, is not sufficient for distinguishing influence from random variation. The higher selectivity of the unconstrained topic drift test can be explained by its ability to generate synthetic data with atypically high magnitudes of change, thus giving the magnitude more weight in the decision. This suggests a direction for further research: Formulate a null hypothesis that fixes the direction but randomizes the magnitude, effectively choosing an adaptive threshold that depends on the similarity of the behavior of influencer and influencee.

Finally, to get a better idea of the correlation between the measurements of influence strength and the human perception of social influence, a brief qualitative evaluation was performed by manually reviewing the top two edges classified as positive in each of the experiments listed in table 6.1. The edges are ranked by $\bar{I}_{a\to b}$ for *M+S* and *M+T*, and by $M_{a\to b}$ for *GC*. For only one of the 22 annotated edges, the associated message sequence exhibits visible effects that are consistent with our definition of social influence. In other words, there is a surprising lack of consistency between the influence edges identified with highest confidence by the proposed methods of influence network recovery and human intuition of how influence should manifest itself. This is consistent with the negative results of Huang (2014) and Schneider (2014), who attempt to develop a graph visualization technique that aids the user in discovering patterns of social influence, and conclude that the Twitter dataset either does not exhibit the expected patterns, or the patterns cannot easily be visualized.

6.4 Discussion

Building upon a theoretical model of social influence, we propose three Granger-causal methods for the detection of influence in social networks and evaluate them on a sample of Twitter communication data. Our attempt at comparative experimental evaluation reveals major issues on three levels: On the conceptual level, our results raise doubts as to

whether information diffusion and its generalization to a topical representation of behavior is an appropriate model of the effects of social influence. On the implementation level, implementation details and choice of hyper-parameters may have a strong restrictive effect on the influence effects that are detectable. Finally, the evaluation of the proposed methods is limited by the lack of standardized, annotated reference data.

Within our simple model of human cognition (see section 5.1), the detection of social influence is fundamentally a problem of causal inference: An individual produces (social) behavior in response to the (social) environment, therefore establishing an influencer-influencee relationship requires causal attribution of the influencee's behavior to actions of the influencer. Influence network recovery can be seen as "screening" large social networks for influence relationships; however, the kind of controlled experiment that would allow for true causal identification (Liotsiou et al., 2016, section 6) is not possible at scale. We therefore have to settle for a weaker notion of social influence. Chua et al. (2015), who operate in a similar setting of Granger-causal inference applied to topical representations of user behavior, use the term *social correlation*. Barrett and Barnett (2013) advise to view estimates of Granger causality as "magnitude[s] of causal effect" that do not contain any information about the mechanism (here, for example, actual influence, homophily, or other confounding factors) that produced the effect.

Acknowledging these theoretical limitations directly leads to one question: How useful are our proposed methods of influence network recovery, e.g., for applications that do not require causal identification, or as "pre-screening" in preparation for more expensive interventional studies that are capable of stronger causal identification? Surprisingly, a follow-up experiment showed that even edges with a high predicted magnitude of influence rarely exhibit any influence effects that are visible to a human annotator. A possible explanation is model misspecification: Peer influence may be present in the Twitter data, but differ from the type of influence described by our model in one or more aspects. For example, influence might be the result of stimuli accumulated over a longer period of time, similar to the linear threshold model of Kempe et al. (2003). In a setting where a single influencer usually does not generate sufficient stimuli to trigger a response, our monocausal model is clearly inappropriate. In the context of information diffusion, non-monocausal models have been discussed under the moniker of *complex contagion*. A body of empirical evidence for complex contagion in social networks can be found in prior studies (Romero et al., 2011b; Hodas and Lerman, 2015; Mønsted et al., 2017). Since our model is based on the textual content of messages, it fails when the main motivation for adopting behavior from others is social, i.e., the probability of adoption mainly depends on the person of the sender (Sousa et al., 2010). Even if influence is content-driven, topic distributions may not be suitable as an intermediate representation, as they only capture broad and highly prevalent topics, so influence-induced change in behavior may not register as a change in topic. Finally, we have to consider that, in the influence network recovery setting, a much stricter selection of the candidate set may be necessary, so that the individual tests for Granger causality retain sufficient statistical power after correction for multiple testing.

Observational data of user behavior in social network services is naturally sparse due to bursty patterns of individual activity, and the choice of temporal quantization interval further affects the sparsity of the resulting time series. Testing for Granger causality re-

quires temporally overlapping activity of influencer and influencee, and quantization with shorter intervals is more likely to produce unmatched observations which cannot be exploited. Identifying variable-length periods of activity by temporal clustering might yield a more data-efficient representation of user behavior, but the resulting *irregular* time series are no longer directly comparable and require alignment akin to dynamic time warping. The *gapped GC* test of Bahadori and Liu (2012) can handle arbitrary irregular time series.

The cost of annotation severely limits the scope of evaluation against ground truth, thus preventing us from fully characterizing the performance in terms of precision and recall, and leaves a high amount of variability in the evaluation results. Evaluation against more easily available data, such as the number of times a retweets b or adopts a hashtag earlier used by b, would require demonstrating that this surrogate ground truth actually reflects social influence, which again requires expensive manual annotation. A way of evaluation not explored in this study is extrinsic evaluation, where the ability of a person to affect the behavior of others is measured in a concrete, real-world use case, and the results are tested for correlation with the strength of influence computed from social media data (Melville et al., 2010). Extrinsic evaluation may discover influence phenomena that are too minute to be noticed by a human observer, but still have a substantial aggregate effect on a derived measure. The availability of a reference dataset or a standardized extrinsic evaluation task would advance the field, first by codifying a common understanding of social influence and its visible effects, and second by allowing researchers to compare their results to baselines and the state-of-the-art.

We conclude that the detection of topical influence is a much harder problem than the tracking of information diffusion, to the point that it cannot be satisfactorily solved by current methods. Even if a user pair is known to be frequently involved in information diffusion, the conclusive detection of topical influence from one user on the other remains difficult. Information diffusion in the form of unmodified retransmission can be reliably detected, but is only weakly linked to social influence. Topical diffusion directly implies some sort of influence, but is much harder to detect and separate from confounding effects. Taken at face value, our experimental results cast doubt upon whether binary features such as retweeting or adoption of hashtags and the diffusion cascades they produce are indicators of actual social influence. If social influence rarely has a directly observable effect, we have to look at other social forces to explain the large cascades associated with "viral" diffusion.

Since the detection of social influence at the dyad level is fraught with theoretical limitations and practical issues, the line of research presented in the following chapter focuses on the aggregate influence an individual receives from the direct social environment.

7 Social Influence at the Meso Level

The series of experiments described in this chapter was conducted in the context of the master's theses of Benjamin Koster (2013), Florian Hartl (2013), Valeria Köllhofer (2013), Julia Strauß (2013), and Shruthi Padma (2014), as well as the bachelor's theses of Gregor Semmler (2013), Johannes Feil (2014), Monika Ullrich (2014), Matthias Wadlinger-Köhler (2015), and Felix Sonntag (2015), all of which were jointly supervised by Jan Hauffa and Georg Groh. The experimental results have been, in part, previously published at the 2nd International Workshop on Social Influence Analysis, co-located with the International Joint Conference on Artificial Intelligence (IJCAI 2016; Hauffa et al., 2016). Verbatim and near verbatim quotations from prior publications (Hauffa et al., 2016) are highlighted in gray.

In the previous chapter, we identify a number of barriers to the identification of influence relationships between individual users of online social networking platforms. Instead of asking if an individual receives influence from another, we may ask more generally if individuals are influenced by their social environment, and how exactly that environment can be defined, thus moving on to the meso level of analysis.

We cast the detection of social influence as a problem of predicting the future behavior of individuals, given their past behavior and the behavior of their respective environments. As in the previous experiment, we analyze behavior in terms of topical representations of communication. The main contribution of this chapter is a predictive model of influence in social networks, which learns to attribute changes in the content of communication to multiple factors: *Inertia*, a person's tendency not to deviate from the current topics of conversation, effectively constitutes a resistance to influence. *Direct exposure* to other people's attitude or belief is a natural consequence of any form of social interaction. Our model also includes the communication behavior within a person's social environment, which plays a double role: One one hand, it is a source of direct exposure. On the other hand, the collective behavior of the environment is a proxy for unobserved behavior of people inside or outside the observed network. This mechanism is known as *indirect exposure*. The model is evaluated on communication data from different social platforms, and the model's predictive accuracy is taken as evidence of its explanatory power. Although the model is predictive in nature, we are more interested in learning about general characteristics of the social influence process than in optimizing the accuracy of predictions.

This experimental setup addresses multiple issues of the previous micro-level experiments: For the purpose of evaluating the model's predictive accuracy, the observations can be arbitrarily divided into past and future, which eliminates the need for labelled reference data. By defining influence as the result of all interactions within an individual's social environment, aggregated over an interval of time, we eschew some of the problems associated with causal inference. In particular, we do not assume that influence effects are monocausal, i.e., explainable by a single causal link between an influencer and an influencee. Further-

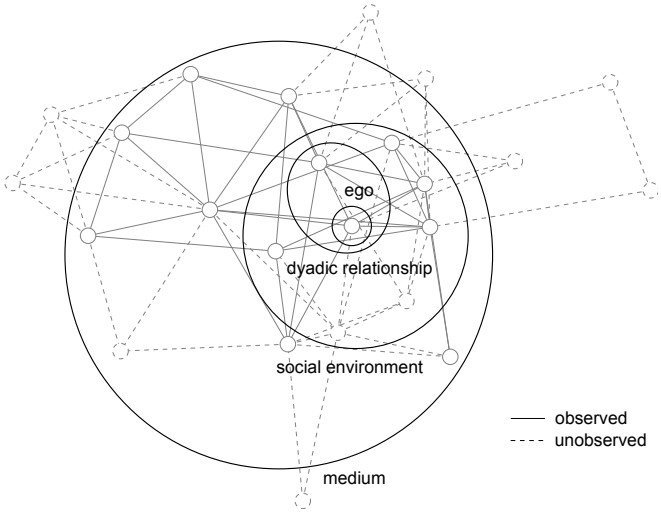


Figure 7.1: A social network as a hierarchy of social circles

more, the effects of unobserved strong influencers (external confounding) are explicitly considered in the model design.

7.1 The Social Content Influence Model

In the following, we develop a predictive model of social behavior. Since we operate on topical representations of the content of social interaction via textual messages, and consider social influence to be the principal force behind any change in behavior, we call this model the Social Content Influence Model (SCIM). The model's fundamental assumption is that all behavior can be fully explained by the presence or absence of social influence, ignoring the potential effects of any other internal cognitive or external social process. It follows that a person's future behavior can be expressed as a combination of *inertia* and *exposure* to others' behavior via participation in or conscious perception of social interaction. In a social network graph induced by social interaction, all first-degree neighbors of a given node are sources of exposure. If exposure is potential influence, inertia is the counterbalancing force of individual resistance to influence. Inertia is defined as a person's tendency not to deviate from past behavior. Empirical evidence for the role of inertia in modeling social influence in the context of information diffusion is presented by Romero et al. (2011a), who improve the accuracy of retweet prediction by ranking users by their passivity, i.e., resistance to influence, in addition to their potential for exerting influence on others.

Figure 7.1 shows an exemplary social network graph. From the perspective of an individual node or node pair (*ego* and *alter*) connected by an edge, it appears as a hierarchy of social circles of decreasing locality. While interactions outside of the studied social platform are unobservable by definition, interactions on the platform typically cannot be completely observed either. To account for randomly missing observations within the network and principally unobservable external actors, we introduce the concept of a node's *social envi-*

ronment, a group of peers which may extend past the node's direct neighbors in the graph. This group is a source of direct exposure, but, as discussed in section 5.1, a source of indirect exposure as well, given that the aggregate behavior of the group potentially reflects strong unobserved influencers.

As a part of the model, we specify a network-wide rule or template for generating an individual influence network for each user, which we refer to as the user's *social neighborhood*. Note the difference between the social environment and the social neighborhood: The environment is a general term that encompasses an individual, any direct interaction partners, and any kind of groups or communities he or she participates in. In the context of the experiments performed in this chapter, the social neighborhood refers to a concrete specification of that environment as a weighted subgraph of the social network graph that explicitly excludes the individual. A particular goal of this experiment is to find a rule for constructing the social neighborhood that works equally well for all members of the social network. Additionally, influence from outside the medium is approximated by the aggregate behavior of the whole network, which is expected to capture strong trends that originate from other media.

A conceptually similar attribution of influence to different parts of the social network in order of decreasing locality can be found in the social recommender system of He and Chu (2010), which recommends items according to user preference, the preferences of friends, and information from mass media, and also in the topic model of Xu et al. (2012), which attempts to explain user behavior on Twitter as being driven by a combination of personal interests, the behavior of friends, and current events. This tripartite view of the egocentric social network in turn corresponds to the distinction between interpersonal, peer, and media influence in sociology (Walther et al., 2010).

The social neighborhood is comparable to the egocentric diffusion networks of Zhang et al. (2015). When defining *local network structure* rather broadly as the structure of any kind of subgraph of an egocentric network graph, usually characterized by a bounded shortest path distance between alters and the ego, the connection between local structure and social influence is well-established in the literature: In an information diffusion setting, the presence of a social tie and its strength affect the probability of transmission (Bakshy et al., 2012; Anagnostopoulos et al., 2008). Zhang et al. (2017) identify influence patterns, that is, frequently occurring isomorphic subgraphs that are associated with a high probability of information transmission. In the context of social contagion, Ugander et al. (2012) show that the probability of contagion is positively associated with the number of connected components in the subgraph induced by infected direct neighbors. Finally, Gulati and Eirinaki (2019) provide an extrinsic evaluation of the utility of local structure by identifying influentials and their targets via a diffusion model, and incorporating these local subgraphs as features in a recommender system.

7.1.1 Topical Representation of the Social Environment

For obtaining topical representations of users and their respective social environments, we follow the scheme presented in section 5.2. Given a set of messages sent on a particular social platform during the observation period, and an ART model with K topics fit to that data, we subdivide the observation period into intervals of equal length, and pick two successive intervals as the *interaction period* and *evaluation period*. We obtain separate topic distributions for the two periods by querying the model with the respective subsets of messages, use the topic distributions from the interaction period to make predictions about the future behavior of users, and compare the predictions to the actual behavior in the evaluation period.

Considering the substantial conceptual differences between addressive and non-addressive communication, it stands to reason that they also have different roles in the social influence process, so we analyze the two kinds of communication separately. Separate topic distributions are estimated for addressive and non-addressive communication: the relationship-topic distributions $\theta^{a}_{i,j}$ represent addressive messages from user i to j, the actor-topic distributions $\theta^{n,s}_{i}$ represent non-addressive messages sent by user i. From these two classes of topic distributions, other distributions can be derived by post-hoc aggregation (see section 4.2.3). Table 7.1 lists all topic distributions that are estimated from the messages of one time period, describes the subset of messages they are computed from, and gives the sets of senders S and recipients R they are aggregated over.

The type of aggregation that is performed in order to obtain $\theta^{n,r}_{i}$, the topic distribution of non-addressive messages received by i, depends on the social platform. Some platforms, like Twitter and Facebook, let users explicitly declare their social relationships. On Twitter, non-addressive messages authored by i are mainly visible to i's followers, i.e., direct neighbors in the explicit social graph $G_{\text{exp}} = (V, E_{\text{exp}})$, so we choose $S = N_i$, $R = V$ with $N_i = \{x \in V : (i, x) \in E_{\text{exp}}\}$. When working with data from platforms that lack an explicit social network, we treat messages with a high number of recipients as non-addressive. Any user j that has sent at least one such message to i within the current time period is included in the set of *exposers* P_i. We choose $S = P_i$ and $R = V$. Facebook is a special case: While there is an explicit network, and non-addressive messages are exclusively shown to direct neighbors in the network graph, the exact dissemination process is unknown. Facebook uses a black-box algorithm to decide which activities are shown to friends. We therefore adopt a mixture of the two aggregation strategies presented earlier. If $P_i \neq \emptyset$, then $S = P_i$, otherwise $S = N_i$. In either case, $R = V$.

Due to the Bayesian nature of probabilistic topic models, the estimate of a topic distribution θ is dominated by the prior distribution, unless there is a sufficient amount of evidence against it. We define θ^{p} as the topic distribution in the complete absence of data, which is the mean of a Dirichlet distribution with parameter α, obtained by normalizing α to unit length: $\theta^{p} = \alpha / \sum_{i=1}^{K} \alpha_i$. If the prior is a symmetric Dirichlet distribution, θ^{p} is equivalent to the uniform distribution. After optimizing an asymmetric prior α as part of the model fitting, θ^{p} reflects the common characteristics of all topic distributions in the dataset. We use θ^{p} as a representation of missing data. For the purpose of predicting human behavior, falling back to the conceptual equivalent of the population average appears to be more appropriate

Table 7.1: Topic distributions estimated from online communication data for use in the SCIM

	definition	S	R
$\theta_i^{n,s}$	non-addressive messages sent by i	$\{i\}$	V
$\theta_i^{n,r}$	non-addressive messages received by i	*	*
$\theta_{i,j}^{a}$	addressive messages from i to j	$\{i\}$	$\{j\}$
$\theta_i^{a,s}$	addressive messages sent by i	$\{i\}$	V
$\theta_i^{a,r}$	addressive messages received by i	V	$\{i\}$
θ_G^{n}	all non-addressive messages	V	V
θ_G^{a}	all addressive messages	V	V

than simply expressing the lack of knowledge by a uniform distribution.

7.1.2 The Predictive Model

The topic distributions of addressive and non-addressive communication are predicted separately. The act of prediction can be formulated as computing $\hat{\theta}_{i,j}^{U,a}$ for each edge from i to j or $\hat{\theta}_i^{U,n,s}$ for each node i, so that the Jensen-Shannon divergence (JSD) of the predictions and the actual distributions $\theta_{i,j}^{a}$ or $\theta_i^{n,s}$ in the evaluation time period is minimal. We define the prediction $\hat{\theta}$ for an individual edge or node as a finite mixture of topic distributions θ^k from the interaction period with $1 \leq k \leq C$ (equation 7.1). For a given set S of edges (i, j) or nodes (represented as pairs (i, i)), the estimation of the coefficients c that minimize the JSD can be expressed as a convex optimization problem over $[0, 1]^{C+K}$ (equation 7.2). The convexity of the function to be minimized follows directly from the convexity of the JSD (Burbea and Rao, 1982), convexity of the ℓ_1 and ℓ_2 norm, and the composition rules for convex functions.

$$\hat{\theta} = \left(\sum_{k=1}^{C-1} c_k \, \theta^k \right) + c_d \, \theta^d \tag{7.1}$$

$$\underset{c,\theta^d}{\operatorname{argmin}} \sum_{(i,j)\in S} \mathrm{JSD}(\hat{\theta}_{i,j}, \theta_{i,j}) + \lambda_1 \|\hat{\theta}_{i,j}\|_1 + \lambda_2 \|\hat{\theta}_{i,j}\|_2 \tag{7.2}$$

$$\text{subject to } 0 \leq c_k \leq 1 \text{ for all } 1 \leq k \leq C,$$

$$0 \leq \theta_t^d \leq 1 \text{ for all } 1 \leq t \leq K,$$

$$\sum_{k=1}^{C} c_k = 1, \sum_{t=1}^{K} \theta_t^d = 1$$

The same model structure, as specified by equations 7.1, 7.2, is used for the prediction of addressive and non-addressive communication, but the number of mixture components θ^k differs. Table 7.2 lists and defines a total of 15 components and names the subset of messages they are computed from. Each component represents either inertia, indirect, or

direct exposure, and is associated with a level of locality relative to the subject of prediction (*scope*), as shown in figure 7.1. The four components at relationship scope are only used for the prediction of addressive communication, the remaining 10 are present in both models.

Table 7.2: Mixture components of the SCIM (Hauffa et al., 2016)

	definition	role	scope
$\theta_i^{n,s}$	non-addr. messages sent by i	inertia	personal
$\theta_i^{a,s}$	addr. messages sent by i	inertia	personal
$\theta_{i,j}^{a}$	addr. messages from i to j	inertia	relationship
$\theta_{N(i)}^{a,s}$	addr. messages from i to neighbors	inertia	environment
$\theta_i^{n,r}$	non-addr. messages received by i	direct exposure	personal
$\theta_i^{a,r}$	addr. messages received by i	direct exposure	personal
$\theta_{j,i}^{a}$	addr. messages from j to i	direct exposure	relationship
$\theta_{N(i)}^{a,r}$	addr. messages from neighbors to i	direct exposure	environment
$\theta_j^{n,s}$	non-addr. messages sent by j	indirect exposure	relationship
$\theta_j^{a,s}$	addr. messages sent by j	indirect exposure	relationship
$\theta_{N(i)}^{n}$	non-addr. messages sent by neighbors	indirect exposure	environment
$\theta_{N(i)}^{a}$	addr. messages sent by neighbors	indirect exposure	environment
θ_G^{n}	all non-addr. messages	indirect exposure	medium
θ_G^{a}	all addr. messages	indirect exposure	medium
θ^{d}	estimated from data	indirect exposure	medium

Unlike the other components, θ^d is not a per-node or per-edge constant, but a parameter vector that is estimated from the data as part of the optimization process. It captures all global effects of influence that are either not explicitly represented in the SCIM or not directly observable, and as such complements θ_G^n and θ_G^a, which represent the aggregated observable behavior of the entire medium. The theoretical justification for a data-driven mixture component is that it allows the model to attain a training error of zero if the influence process does not have any individual characteristics, i.e., affects the behavior of all actors in the exact same way.

Similar to elastic net regularization for linear regression (Zou and Hastie, 2005), the additive penalty term combines ℓ_1 and ℓ_2 regularization. A major difference to regularization methods for linear regression is that the penalty is a function of the prediction $\hat{\theta}$ rather than a function of the model parameters. The ℓ_1 term penalizes non-sparse predictions $\hat{\theta}$ and thus indirectly also promotes the sparsity of parameter vector θ^d. It represents our knowledge about the sparsity of the topic distributions generated by the ART. The ℓ_2 term penalizes predictions with large individual components and therefore makes the model prefer conservative predictions that are closer to the uniform distribution. The regularization weights are fixed at a value of $\lambda_1 = \lambda_2 = 0.0001$, which implies that the two terms contribute equally to

the result.

The shorter the interaction period, the more likely it is not to contain any observed interactions initiated by a particular person. In this case, the topic distributions of that person are equal to θ^p, and do not contribute any information to the model. Ideally, the length of the interaction period should match the speed of conversation flow within the medium. If a medium is characterized by rapidly changing topics, we want to keep the interaction period short. To compensate for the lower volume of observation data, we introduce a fallback mechanism. In addition to the original interaction and evaluation periods of length n, topic distributions are also computed for extended periods of length $m > n$. If a topic distribution equals θ^p, it is substituted with the corresponding distribution from the extended period. Instead of making the conservative assumption that a lack of observed interaction implies a lack of knowledge about a person's attitude, we assume that the attitude has not changed since the last observation.

By defining the prediction as a mixture of topic distributions, we implicitly make assumptions about the nature of the influence process: Computing a single set of scalar coefficients that globally minimizes the JSD for all nodes or edges corresponds to the hypothesis that all nodes react in the same way to exposure from different parts of the social network, independent of the topic or their individual social context. In other words, we assume that the influence process is dominated by global, instead of individual or topical, characteristics. While this model can only detect global effects of influence, i.e., effects that can be observed to the same degree across the whole network, it is not limited to additive influence. Arbitrary dependencies between observed and future behavior can be learned, as long as they can be expressed as linear combinations of the mixture components. Focusing on global characteristics is expected to be detrimental to prediction accuracy, but is consistent with our goal of learning about the social influence process at the network level.

7.1.3 Construction of the Social Neighborhood

The social neighborhood $N(i) = (V_i, E_i, W_i)$ of a node i is a node-weighted subgraph of the social network graph (V, E). It is induced by an indicator function $I_i : V \to \{0, 1\}$ and a weight function $W_i : V \to \mathbb{R}^+$, so that $V_i = \{v \in V : I_i(v) = 1\}$ and $E_i = \{(a, b) \in E : a \in V_i \wedge b \in V_i\}$. Given normalized weights $W_i'(v) = W_i(v)/\sum_{w \in V_i} W_i(w)$, the neighborhood mixture components $\theta_{N(i)}$ are computed according to equation 7.3. The temporal fallback mechanism is applied in the following way: If a topic distribution involving a node $v \in V_i$ equals θ^p, it is replaced by the corresponding distribution from the extended interaction period. If that distribution contains no information either, $W_i(v)$ is set to 0 and W_i' is updated accordingly.

$$\begin{aligned} \theta^{a,s}_{N(i)} &= \sum_{v}^{V_i} W_i'(v) \cdot \theta^a_{i,v} & \theta^{a,r}_{N(i)} &= \sum_{v}^{V_i} W_i'(v) \cdot \theta^a_{v,i} \\ \theta^{n}_{N(i)} &= \sum_{v}^{V_i} W_i'(v) \cdot \theta^{n,s}_{v} & \theta^{a}_{N(i)} &= \sum_{v}^{V_i} W_i'(v) \cdot \theta^{a,s}_{v} \end{aligned} \tag{7.3}$$

Within the SCIM, the social neighborhood has a dual role as a source of direct and indirect

exposure. By comparing the prediction accuracy of models with different I_i and W_i, we aim to identify the neighborhood definition that, for any given node, produces a subgraph that fulfills both roles best. As a source of direct exposure and influence on an actor, this subgraph should closely match the group of people who are involved in that actor's process of opinion formation, similar to the concept of a peer group in sociology. The node weight indicates how strongly a node is associated with that group, and how much it consequently contributes to the collective attitude of its members. As a source of indirect exposure, the neighborhood should be exposed to similar external and unobserved sources of influence as the actor at its center. Homophily is one mechanism that links short path distance in the social network graph and similarity of interests and preferences.

Before we can propose a set of candidate indicator and weight functions, we must ask whether the social neighborhood should be derived from the explicit or the implicit social network graph. Even though the explicit social network is a source of first-hand information about a user's social relationships, there are some downsides to using the explicit network as the only source of information: Obviously, not all social platforms provide an explicit network. Among the platforms that do provide explicit social edges, the user-visible effect of establishing an edge varies substantially, and therefore the implied meaning of the presence or absence of an edge is different as well. The high temporal sparsity implies that the presence of an explicit edge not as indicative of actual interaction as one would like. Just as predicted by social network theory (see section 2.1.1), explicit and implicit network coexist, and neither can be said to be the more faithful or "canonical" source of relationship information. The proposed indicator and weight functions therefore make use of either source of information, where available.

We implement and compare seven indicator functions and 29 weight functions, listed in tables 7.3 and 7.4. One family of indicators is based on shortest path distance in the social network graph. This is generally motivated by the principle of locality, and more specifically by the results of Adamic and Adar (2003), who find that the path distance between two nodes in a social network graph is inverse proportional to the similarity of their social contexts, which drops sharply and remains constantly low at a path distance of three or above. Determining whether a relationship exists between two users of an online social platform is not trivial: An explicitly declared relationship usually implies exposure to another person's non-addressive messages, but not necessarily mutual awareness, which is a prerequisite for a social relationship. Addressive communication is a stronger indicator of a social relationship, but a true social relationship might not be necessary for one person to influence another. We therefore propose four separate indicator functions, which define the neighborhood of i as the set of all nodes within a maximum distance of either one or two from i, either in the explicit network graph or the implicit network graph induced by addressive communication. The implicit network is treated as an undirected graph for the computation of the path distance. In the analysis of social platforms that do not have an explicit network, the corresponding indicator functions are omitted.

A second family of indicator functions attempts to identify cohesive social groups. As social cohesion is expected to correspond to local density in the social network graph (Newman and Girvan, 2004), an obvious candidate are cliques, i.e., fully connected subgraphs. A simple indicator function picks the largest maximal clique (fully connected subgraph) con-

Table 7.3: Neighborhood indicator functions of the SCIM

	definition
network	nodes of distance 1 in explicit network
	nodes of distance ≤ 2 in explicit network
	nodes of distance 1 in communication network
	nodes of distance ≤ 2 in communication network
communities	largest maximal clique
	union of communities found by clique percolation
	union of communities found by edge clustering

Table 7.4: Neighborhood weight functions of the SCIM

	definition	transformation (optional)
	constant	-
network	in-degree	$\log(1+x)$
	out-degree	$\log(1+x)$
	ratio of in- and out-degree (after log-tf.)	-
	number of neighbors shared with i	-
	PageRank	-
	betweenness centrality	-
	reciprocity score	-
communities	number of maximal cliques	$\log(1+x)$
	number of clique-based communities	$1/(1+x)$
	number of edge-based communities	$1/(1+x)$
	number of clique-based communities shared with i	-
	number of edge-based communities shared with i	-
communication	number of addr. messages sent	$1/(1+x)$
	number of non-addr. messages sent	$1/(1+x)$
	number of messages shared by others	-
	ratio of msg. being shared and non-addr. msg. sent	-
content	similarity of addr. messages to i	-
	similarity of non-addr. messages to i	-
	similarity of non-addr. messages to medium avg.	1 - x
	focus level	-

taining i as its neighborhood. If there is more than one largest clique, an arbitrary choice is made. A more principled approach is offered by the community detection algorithms described in section 3.6.2. Communities may overlap, which matches the observation that a person is usually member of many different social groups (Ahn et al., 2010). The indicator functions for clique- and edge-based communities define the neighborhood of node i as the union of the communities i is a member of. Here, and for all weight functions based on clique and community membership, we use the implicit network graph, transformed to an undirected graph after removing all unidirectional edges. The reasoning behind this choice is that this graph is available for all social platforms, and is small enough so that all maximal cliques and communities can be found in a reasonable time frame.

A weight function assigns a positive real value to any node j in the neighborhood of i. The most simple weight function assigns uniform weight to all nodes, thus setting a baseline for more complex functions. The other weight functions can be divided into three categories: Functions of the first category derive the weight from structural properties of the social network graph. Both the out- and the in-degree of j in the network graph represent j's connectedness. Node degrees in the network graphs of online social platforms commonly follow a power-law distribution. For example, this is known to be true for the Twitter follower graph (Java et al., 2007). If highly connected nodes are selected by the indicator function, their normalized weight will dominate the neighborhood. To investigate whether a more balanced weight distribution yields a better neighborhood, the set of weight functions includes both the actual degree d and its value after a transformation $\log(1 + d)$, which dampens the weight of high degree nodes, as separate functions. The ratio of in- and out-degree (both log-transformed, i.e., $\log(1 + d_{\text{in}})/\log(1 + \max\{1, d_{\text{out}}\})$) is also included as a commonly used proxy measure of influence (Anger and Kittl, 2011).

Additional weights in this category are the number of common friends of i and j, the PageRank and betweenness centrality of j, and the reciprocity score of i and j. The latter takes on different values depending on how well the two nodes are connected in the explicit and implicit network graphs: 4 in case of a mutual explicit relationship and reciprocal communication, 2 if there is reciprocal communication or a mutual relationship and unidirectional communication, 1 in case of mutual relationship, and 0.5 otherwise. In the analysis of social platforms that do not have an explicit network, the structural properties of the implicit network graph are used instead, and the reciprocity score is omitted, as it depends on the properties of both kinds of network. For social platforms with undirected explicit relationships (e.g., "friendship" on Facebook), the weights derived from out-degree and degree ratio are redundant and are omitted as well.

The weights in the second category make use of community structure, in particular the number of cliques, clique-based communities, and edge-based communities j is a member of. The clique membership count is bounded above by the number of cliques in a graph, which in turn has an upper bound that is exponential in the number of nodes. Analogous to the degree weights discussed previously, we therefore include the log-transformed value as a separate weight. Simultaneous membership in many communities may indicate a well-connected and thus highly influential person, but the behavior of a person bridging many communities is unlikely to be representative for any particular community. To account for both possibilities, we provide separate weight functions for the actual number of communities n and its inverse

$1/(1+n)$. Another weight in this category is the number of shared communities of i and j.

The third category contains weights that represent the communication behavior of j. Some weights are based on communication statistics, such as how often the messages of j are shared by others within the interaction period, the overall communication volume as represented by the number of non-addressive messages sent by j, the number of addressive messages exchanged between i and j, and the ratio of messages shared by others to non-addressive messages sent. Since a high communication volume might be due to automated, repetitive, or otherwise irrelevant messages, we also include weight functions based on the inverse message volume.

The remaining weights are derived from the content of communication, as represented by the topic distributions of i and j: Similarity of non-addressive communication of i and j (defined as $1 - \mathrm{JSD}(\theta_i^{n,s}, \theta_j^{n,s})$), similarity of addressive communication (comparing $\theta^{a,s}$ of i and j), as well as similarity and distance of non-addressive communication of j and the aggregate topic distribution of the medium (comparing $\theta^{n,s}$ and θ_G^n). The topical similarity and distance weight functions always take on a value of zero if one of the involved users is inactive. The final weight function is the focus level, which is defined as the sum of all components of $\theta_j^{n,s}$ at the 90th percentile or above, scaled to the interval $[0, 1]$. This can be interpreted as a measure of the sparsity of the topic distribution.

7.1.4 Theoretical Properties

Feil (2014) proposes a model of topical information diffusion and analyzes its theoretical properties. This diffusion model complements the SCIM in the sense that it provides a controlled environment that operates exactly according to the assumptions of the SCIM. Studying this environment allows us to draw conclusions about the properties of the SCIM under ideal conditions. The model describes the flow of non-addressive communication within a directed social network graph $G = (V, E)$ and its effect on the internal state of the actors. Like in the Twitter follower graph, an edge towards an actor represents a "subscription" to the messages written by that actor, i.e., information flows in the opposite direction of the edges.

For each actor $i \in V$, let $\theta_i(0) \in \mathbb{R}^K$ denote the initial topic distribution, $w_i \in \mathbb{R}_*^+$ an arbitrary, positive weight that represents the strength of the influence i exerts on others, and N_i^+ and N_i^- the set of neighbors along outgoing and incoming edges, respectively. A global parameter $\alpha \in (0, 1)$ specifies the receptiveness of the actors to influence. At time $t \geq 0$, each actor i transmits messages with a topic distribution $\theta_i(t)$ to N_i^-. The actors' topic distributions are updated accordingly:

$$\theta_i(t+1) = (1-\alpha) \cdot \theta_i(t) + \alpha \cdot \frac{1}{\sum_{j \in N_i^+} w_j} \cdot \sum_{j \in N_i^+} w_j \theta_j(t)$$

We assume that $\forall_{i \in V} : |N_i^+| > 0$, so that the value of the update equation is never undefined. This equation defines future behavior as a mixture of inertia and direct influence, without any effects of external influence. An equivalent update rule for the entire graph can be defined as follows: Let $\Theta(t)$ be a $|V| \times K$ matrix, where the i-th row is given by $\theta_i(t)^T$. Let A

be a $|V| \times |V|$ matrix with $A_{i,j} = \frac{w_j}{\sum_{k \in N_i^+} w_k}$ if $(i, j) \in E$, 0 otherwise, i.e., the adjacency matrix of G weighted by w_j after row-wise normalization. The update rule is given by:

$$\Theta(t+1) = P \cdot \Theta(t) \quad \text{with}$$
$$P = (1-\alpha) \cdot I + \alpha \cdot A$$

The matrix P is stochastic, since it is a convex combination of two stochastic matrices. Consider the homogeneous Markov chain M with state space V and transition matrix P. P can be interpreted as the adjacency matrix of a weighted graph T_P, the transition graph of M. An unweighted graph with the same adjacency can be obtained by adding a loop edge to every node in G. Therefore, if G is strongly connected, so is T_P. Since the transition graph is strongly connected, and each node has a loop edge with strictly positive weight, M is irreducible and aperiodic. It follows that M has a unique stationary distribution π, given by the rows of $\lim_{t\to\infty} P^t$. Θ has the closed form representation $\Theta(t) = P^t \cdot \Theta(0)$. Since P^t converges, so does $\Theta(t)$:

$$\Theta^* = \lim_{t\to\infty} \Theta(t) = \lim_{t\to\infty} \left(P^t \cdot \Theta(0)\right) = \left(\lim_{t\to\infty} P^t\right) \cdot \Theta(0) = \begin{bmatrix} \pi^T & \dots & \pi^T \end{bmatrix}^T \cdot \Theta(0)$$

The topic distributions of all actors converge to the same limit $\theta^* = \sum_{i \in V} \pi_i \theta_i(0)$, a convex combination of the initial topic distributions. One can show by simple algebraic manipulation that $\pi^T \cdot P = \pi^T \Leftrightarrow \pi^T \cdot A = \pi^T$, so θ^* does not depend on α, and is only determined by the initial topic distributions, the structure of the graph, and its edge weights. However, α affects the rate of convergence. By comparing the eigenvalues of A and P, we can bound the second largest eigenvalue of P: $|\mu_2| \leq 1 - \alpha \cdot (1 - |\lambda_2|)$, where μ_2 and λ_2 are the second largest eigenvalues of P and A, respectively. Convergence is faster for higher values of α and lower $|\lambda_2|$.

The convergence of all actors' behavior to the same point in the chosen representation space occurs analogously in the autocorrelation model of Friedkin and Johnsen (1999), which aims to describe the process of opinion formation within isolated groups. Since we do not observe this kind of global convergence in the collected social media data, we have to assume that the process that would otherwise lead to convergence is continuously perturbed by external influence. The diffusion model can be extended accordingly. We introduce a global parameter $\beta \in (0, 1)$ that specifies the ratio of internal to external influence. The new update rule for the entire graph is:

$$\Theta(t+1) = Q \cdot \Theta(t) + \Xi(t) \quad \text{with}$$
$$Q = (1-\alpha) \cdot I + \alpha\beta \cdot A$$
$$\Xi(t) = \alpha(1-\beta) \cdot \Theta^{\text{ext}}(t)$$

$\Theta^{\text{ext}}(t)$ represents the external influences that affect each actor at time t. If $\beta < 1$, Q is sub-stochastic (row sum less than 1), so $\lim_{t\to\infty} Q^t = 0$. The closed form of the update rule is:

$$\Theta(t) = Q^t \cdot \Theta(0) + \sum_{i=0}^{t-1} Q^i \cdot \Xi(t-i)$$

With these modifications, $\Theta(t)$ does no longer converge, and the impact of the initial topic distribution, as well as the effect of external influence at a particular point in time, vanishes over time due to the convergence of Q^t to zero. The weighted adjacency matrix A has a constant effect on the process of information dissemination.

The topical model of information diffusion defined here incorporates the effects of inertia and direct influence. Its parameters are per-edge influence probabilities and two mixture coefficients that globally control the balance of inertia, internal influence, and external influence. The model can be used to generate longitudinal sequences of observations with known asymptotic properties. The SCIM performs the opposite task: given longitudinal observations, we aim to recover parameters of the underlying influence process. The effect of social network structure and individual influence strength on the diffusion process further motivates our search for salient neighborhood definitions. Due to its fixed window size, the SCIM is unable to learn about the ratio of internal to external influence, which only manifests in the amount of time for which information is preserved. Given that continued external influence is required to keep the model from converging, the effect of external influence may be stronger than predicted by the hypothesis of locality.

7.2 Experiment Design

A *basic prediction experiment* for the evaluation of the SCIM on actual social media data can be defined as follows: Since predictions for addressive and non-addressive communication are made by different variants of the SCIM, the first step is to construct separate candidate sets of edges and nodes. An edge (i, j) or a node i of the implicit social network graph is eligible if there are observed interactions in both the interaction and the evaluation period, that is, if $\theta^a_{i,j}$ or $\theta^{n,s}_i$ differ from θ^p in both periods, after applying the fallback mechanism if necessary. The candidate sets are then randomly split into training and test sets of equal size.[1] After parameter estimation on the training set, each model is evaluated on the test set by comparing the prediction to the actual interactions in the evaluation period. This basic experiment, visualized in figure 7.2, is performed repeatedly to exhaustively test all combinations of the following five experiment parameters:

Dataset We analyze data from the three online social platforms presented in chapter 3: Twitter, Facebook, and e-mail, represented by the Enron and HackingTeam collections. The HackingTeam collection contains e-mails in English and Italian language. Since topic models cannot directly handle multilingual text corpora, we separate the messages by language and analyze the two subsets separately. Due to limited computational resources, the analysis of Twitter data is restricted to a 30 000 user subset, further excluding another 3 000 users, whose data served as a development set for hyper-parameter tuning. The other datasets are analyzed in full.

[1] Obviously, this strategy is only applicable to the analysis of historical data. In a true prediction setting, where data from the interaction and evaluation period is used for parameter estimation, and the resulting model is applied to the prediction of events in a third, yet unseen period, we expect a higher generalization error.

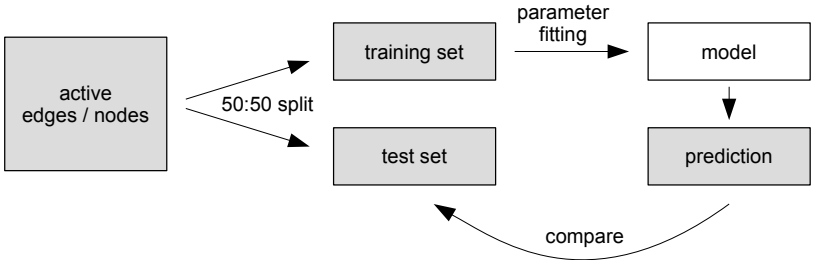


Figure 7.2: Basic SCIM prediction experiment

Time period length The SCIM learns to express future behavior as a mixture of past observations, so the length of the interaction and evaluation period should match the speed of conversation flow. If the interaction period is too long, the observed interactions will include outdated information that is no longer useful for explaining the behavior in the evaluation period. If the interaction period is too short, the number of observed interactions will be too low for the robust estimation of topic distributions. Using shorter periods may also introduce bias towards the characteristics of users who frequently send messages. For instance, active users are also more frequently exposed to other users' actions, and thus might be generally more susceptible to influence.

Interaction and evaluation period always have the same length. Appropriate values can be derived from the temporal analysis of the social media datasets in section 3.5.2: Period lengths of 14 and 2 days match the experimentally determined typical time scales of interaction, while a length of 5 days is included as an intermediate value. One could make a case that the temporal dynamics of social media require periods even shorter than 2 days, but a lower bound is imposed by the sparsity of activity within the network. When discretizing time into intervals of equal length, not all nodes and edges will have at least one associated message in each interval. A node or edge is eligible for a prediction experiment if there is activity in the interaction and evaluation period. Table 7.5 shows the ratio of eligible nodes and edges in each dataset, averaged over the different observation dates (defined below). A more fine-grained analysis may require adaptation of the SCIM to a different temporal discretization scheme. The length of the extended time period of the fallback mechanism is always 14 days, so fallback is effectively disabled for experiments with 14 day interaction and evaluation periods.

Observation date This refers to the point in time that marks the end of the interaction period and the beginning of the evaluation period. The motivation behind varying the observation date is testing the stability of the model over time. Although there may be circumstances which make future behavior easier to predict, e.g., an the upcoming release of a big-budget movie, which is likely to generate social media buzz with clearly recognizable influence patterns, in general one would expect the prediction accuracy not to depend on the observation date. Similarly, the mixture coefficients should exhibit low variability across the different observation dates. Three equidistant dates within a time frame of eight weeks (see section 3.5.3) were chosen so that

Table 7.5: User activity on different social platforms after temporal quantization

	nodes	activity		edges	activity	
		2 days	14 days		2 days	14 days
Twitter	30 000	38.5%	63.1%	3 877 963	0.02%	0.16%
Facebook	16 834	1.9%	8.2%	1 221 273	> 0.01%	0.02%
e-mail (HT-en)	60	5.0%	20.6%	770	3.12%	32.90%
e-mail (HT-it)	60	2.8%	31.1%	950	1.61%	35.93%
e-mail (Enron)	158	1.5%	21.3%	483	0.14%	13.73%

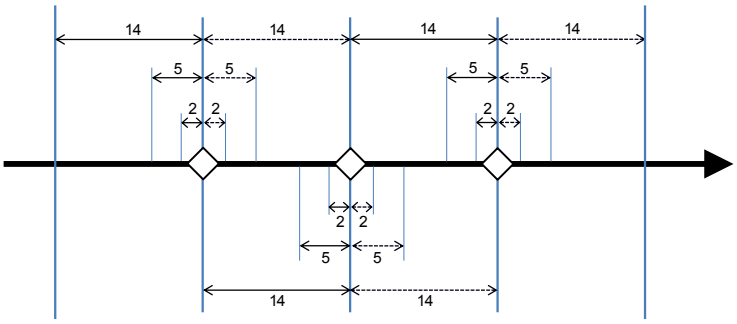


Figure 7.3: Interaction periods (solid) and evaluation periods (dashed) of the SCIM experiments (adapted from Koster, 2013)

there is no overlap: The end of each of the longest evaluation periods coincides with the beginning of the subsequent interaction period. Figure 7.3 illustrates the possible combinations of observation date and time period length. Note that even though we analyze each pair of interaction and evaluation period independently, the observational data as a whole can be viewed as time series of topic distributions, comparable to those used in the micro-level experiments.

Relationship type This parameter is only used in prediction experiments involving addressive communication on a social platform with explicit relationships. When selecting edges for training and evaluation, the parameter controls whether any edge of the implicit social network graph is eligible, as long as there are observed interactions along the edge within both time periods, or if a corresponding edge in the explicit social network graph is required as well. The presence or absence of an explicit relationship affects the visibility of messages in a platform-specific way. Furthermore, a relationship that has been explicitly declared may exhibit different patterns of social influence than an implicit relationship formed by interaction.

Neighborhood The social neighborhood is defined by the combination of an indicator and a weight function, as discussed in section 7.1.3.

When designing the SCIM, one goal was to minimize the number of assumptions about the nature of the influence process that are explicitly built into the model. In consequence, we expect the model to be highly expressive, in the sense that the model parameters and experiment parameters span a large space of hypotheses about the inner workings of social influence. After providing evidence that the SCIM actually learns about the influence process, we perform explorative statistical analysis to narrow down this hypothesis space by identifying the subset of parameters that consistently affect the prediction accuracy. Specifically, we pose the following questions to the experimental results:

1. Baseline predictors provide the necessary context for the interpretation of the predictive performance of the SCIM. Can the SCIM outperform trivial baseline models, thus demonstrating its capability to learn about the influence process?

2. To what degree does the SCIM's performance depend on a single experiment parameter or combinations of multiple parameters? In particular, what is the effect of the neighborhood definition, and is there a set of neighborhoods that consistently outperforms all others?

3. Each prediction experiment yields a set of SCIM mixture coefficients, which can be interpreted as an explanation of the influence process that fits the observed data. How stable are these characteristics across the experiments, i.e., is there a single explanation that generally works well?

We compare the SCIM to three baseline predictors to verify that its parameters actually represent non-trivial information about the influence process. The first baseline method makes a prediction by random draw from a Dirichlet distribution with the same parameter vector α as the Dirichlet prior of the document-topic distributions in the ART model. Performing worse than random chance would constitute a lack of evidence for the fundamental model assumption that future behavior can be explained by a combination of inertia and exposure to the behavior of others. The second baseline method always outputs θ^p, i.e., the topic distribution fully determined by the prior distribution, as the prediction. Not outperforming this baseline would indicate insufficient data: Observed interactions are infrequent or contain too little information, so the resulting topic distributions are too close to θ^p, and the generalization error of the SCIM exceeds the bias of a model that always predicts θ^p. The third baseline method outputs the topic distribution corresponding to observed current behavior, i.e., $\theta_i^{n,s}$ in case of non-addressive and $\theta_{i,j}^{a}$ in case of addressive communication, as a prediction of future behavior. This can be seen as a model of social influence fully driven by inertia. Failing to outperform this baseline would be evidence against exposure as the mechanism by which influence is received, but also against the assumption of locality: If influence is mainly exerted by actors outside of the observed medium, the local exposure captured by the SCIM has little explanatory value.

Given these baseline predictors, the first two questions can be addressed by statistical evaluation of the experimental results using ANOVA (Doncaster and Davey, 2007) in conjunction with a post-hoc test. Given a set of entities (sampling units) with associated measurements, ANOVA ("analysis of variance") is a statistical method for attributing the variation of these

measurements to discrete properties of the entities (factors). Between-subjects factors are properties that differ from sampling unit to sampling unit, while a within-subjects factor corresponds to taking multiple measurements of the same unit under different circumstances. To determine the effect of the experiment parameters on prediction accuracy, we propose an ANOVA design where selected parameters are treated as between-subjects and within-subjects factors, individual nodes or edges are the sampling units, and the prediction error is the response variable. The experiment parameters chosen as between-subject factors are time period length, observation date, and relationship type, while the neighborhood definition is the single within-subjects factor. We use a factorial ANOVA design that can detect the effects of individual factors (main effects) as well as joint effects of multiple factors (interactions).

In the context of ANVOA, a sample (or group) is a set of sampling units that are homogeneous with respect to the between-subject factors. ANOVA requires one sample for each combination of factor levels (the possible values of a factor), and its statistical power benefits from equally sized samples. Furthermore, the samples have to be mutually disjoint. The candidate sets of the basic prediction experiments are constructed accordingly: If a node or an edge is eligible for membership in the candidate set of multiple experiments, its is randomly assigned to a particular one, effectively partitioning the dataset. Afterwards, all candidate sets are shrunk to the size of the smallest set by randomly discarding members in order to obtain a balanced design that is more robust to violations of the model assumptions (Doncaster and Davey, 2007, pp. 237). Due to the partitioning, this procedure requires a large overall sample size in order to retain a sufficient sample size for each combination of factor values. Therefore, we perform the full ANOVA procedure for the Twitter dataset only. For the smaller datasets, we simplify the ANOVA design by dropping all between-subjects factors, and only testing the effect of the neighborhood definition for a single, manually chosen combination of the other parameters. Specifically, we only analyze the results of experiments using data from the most recent observation date, a period length of 14 days, and the most inclusive relationship type, if applicable.

The full ANOVA procedure is performed twice: Initially, we include the results of the baseline predictors by treating the baselines as additional levels of the neighborhood definition factor, in order to be able to test individual neighborhoods for a significant improvement over the baselines. Then the procedure is repeated without the baseline results to investigate the effect of the experiment parameters on prediction accuracy. The simplified procedure always includes the baseline results.

To improve interpretability, the experiment results are filtered before performing ANOVA. We only keep the results of an experiment when there is evidence that the SCIM has successfully learned about the influence process. In addition to the three baseline predictors discussed previously, we introduce two new baseline models, restricted variants of the SCIM, which are only used for filtering the experiment results, specifically, to assess the explanatory value of the fitted coefficients and the neighborhood definitions. In the first variant, the coefficients are fixed to assign uniform weight to all components except θ^d: $c_{1..(C-1)} = 1/(C-1), c_d = 0$. In the second variant, the coefficients are determined as usual, but the neighborhood is always empty. If a neighborhood definition does not consistently outperform these SCIM variants and the three baseline predictors across all combinations of ex-

periment parameters, the results of all experiments using this neighborhood are discarded.

Due to the high number of combinations of indicator and weight function that make up the set of possible neighborhood definitions, treating the choice of neighborhood as a between-subjects factor, and partitioning the dataset accordingly, would require an infeasibly large amount of data. Therefore, we perform experiments with different neighborhood definitions, but otherwise identical experiment parameters on the same sample, which can be viewed as taking repeated measurements of each sampling unit. Regular ANOVA with repeated measurements (rmANOVA) requires sphericity, i.e., the homogeneity of variances of the pairwise differences between repeated measurements (Doncaster and Davey, 2007, pp. 183). The more repeated measurements are taken, the more likely is a violation of sphericity. This potential problem can be avoided by formulating the experiment as a MANOVA design, which, in contrast to ANOVA, admits multiple response variables. Each repeated measurement is treated as an additional response variable (O'Brien and Kister Kaiser, 1985). By providing equally sized samples with a sample size that exceeds the number of repeated measurements, we expect the design to be robust to potential violations of the model assumptions, including multivariate normality (O'Brien and Kister Kaiser, 1985).

MANOVA is negatively affected by multicollinearity of the response variables (Grice and Iwasaki, 2007). In the context of the SCIM experiments, the most likely sources of multicollinearity are neighborhood definitions that produce highly similar neighborhoods for many nodes, particularly definitions that often produce empty neighborhoods. Since the neighborhood is the only point of distinction between two experiments with the same data and experiment parameters, the prediction results will be highly correlated, and MANOVA may fail to detect effects involving the within-subjects factor. We attempt to avoid this issue by designing the indicator and weight functions so that they produce a non-empty neighborhood for the majority of candidate nodes and edges.

The repeated measurements design suffers from two general limitations: First, one cannot test for interaction of individual sampling units with the within-subject factor (Doncaster and Davey, 2007, p. 29), which mirrors the implicit assumption of the SCIM that each neighborhood definition works equally well for all members of the social network. If the effect of the choice of neighborhood varies from person to person, it may be erroneously reported as not significant. Second, we test whether the choice of neighborhood has a significant effect on the prediction accuracy, but the test cannot directly identify individual neighborhood definitions that perform significantly better than others. We address this second issue by means of post-hoc testing. We first perform Tukey's HSD test to find all pairs of neighborhood definitions with significantly different mean prediction error. Then we construct homogeneous subsets, i.e., sets of neighborhood definitions in which no member differs significantly from any other. The construction is equivalent to finding the cliques in an undirected graph, if the nodes correspond to neighborhood definitions, and the presence of an edge indicates absence of a significant difference. In addition, by computing the weakly connected components of that graph, it is possible to test for the existence of non-overlapping subsets, which are significantly distinct from each other. Ranking the subsets by mean accuracy identifies the subset of neighborhoods that generally perform best across all combinations of the other parameters. By including the results of the three baselines in the post-hoc test, we can test for a significant performance improvement over the trivial predictors.

In the ANOVA design described here, all between-subject factors are fixed effect factors. The observation date would usually be modeled as a repeated measurement or a random effect factor, as the chosen levels are representative for a large range of possible dates, and any significant effect would be expected to be replicable for a different choice of dates. However, to avoid the difficulties in modeling and interpretation that are associated with multiple within-subject factors, we treat it as a between-subject fixed factor. This reduces the generality of the ANOVA results to a certain extent (Doncaster and Davey, 2007, pp. 16): Instead of testing for the presence of an effect that is consistent over all possible factor levels, only those levels that are explicitly entered into the model are tested. Since we are repeating the experiment at different observation dates with the specific goal of assessing the stability of prediction accuracy over time, this is an acceptable trade-off. Any significant effect has to be treated as evidence for a time-dependency that requires further investigation.

Guided by the results of ANOVA, we then attempt to answer the third question via purely descriptive statistics. The post-hoc test yields the set of best-performing neighborhood definitions. For each coefficient of the SCIM, we compute mean and standard deviation over all experiments that use one of the neighborhood definitions in that set. We also compute these descriptive statistics for groups of coefficients corresponding to the *roles* and *scopes* defined in table 7.2. The coefficient mean can be interpreted as the explanatory value of the mixture component, while the standard deviation indicates how consistently the component contributes information to the model across the different experiments.

7.3 Results

We repeatedly perform the basic prediction experiment for all combinations of experiment parameters. The experimental results are evaluated with one of the two proposed statistical procedures, depending on the size of the dataset: The Twitter dataset and its variants are sufficiently large, so that the effects of all experiment parameters on the results can be determined using the full ANOVA procedure. The other, smaller datasets require a simplified ANOVA design that is limited to analyzing the effect of the choice of social neighborhood. Since the results of the different ANOVA procedures are not directly comparable, we start with an in-depth analysis of the Twitter dataset, and then conduct a more shallow, comparative analysis of all datasets that aims to identify common characteristics of influence across different social platforms.

7.3.1 Twitter

Given the experimental results, the first step of the analysis is to eliminate neighborhood definitions that do not consistently outperform the baselines. ANOVA is performed on the results of the remaining experiments. In the analysis of non-addressive communication, 81.3% of neighborhood definitions are filtered out. In preparation for ANOVA, the set of sampling units is partitioned into 9 samples of size 261. Table 7.6 summarizes the between- and within-subject effects. In the analysis of addressive communication, 82.3% of neighborhood definitions are filtered out. The sampling units are partitioned into 18 samples of size 261, and the between- and within-subject effects are summarized in table 7.7. The

level of significance is indicated by one or more asterisks, where one asterisk corresponds to $p < 0.05$, two to $p < 0.01$ and three to $p < 0.001$. The reported significances are according to Pillai's trace, which is considered to be a robust test statistic even in designs with low sample size. We note that here and in the following applications of ANOVA, two other common test statistics, Wilk's lambda and the Hotelling-Lawley trace, consistently yield the same significance level. A fourth statistic, Roy's largest eigenvalue, tends to favor higher-order interactions. Following the advice of Carey (1998), we choose to disregard effects that are only significant according to Roy's statistic.

The presence of higher-order interactions makes lower-order interactions and main effects harder to interpret: ANOVA does not admit any conclusion as to whether a significant lower-order effect is meaningful on its own or is already adequately explained by a higher-order interaction. Conversely, interactions may also cause meaningful lower-order effects to appear as not significant.

For both types of communication, there is a significant three-way interaction involving the within-subject factor *neighborhood definition* and the between-subject factors *observation date* and *time period length*, and also a significant two-way interaction of *neighborhood* and *time period length*. This indicates that the utility of some or all of the neighborhood definitions varies over time and is also dependent on the length of the interaction period. The interaction of neighborhood and time period length is particularly interesting, as it suggests that influence processes at different time scales involve different substructures of the network graph. In the case of addressive communication, the two-way interaction of *neighborhood* and *observation date* is significant as well. This interaction is harder to interpret: It implies that the utility of a neighborhood definition for the prediction of addressive communication is not constant over time. We suspect that this is related to the fact that, as established in section 3.5.2, people rarely communicate via a constant stream of messages, but rather in bursts of activity between longer periods of silence. The utility of a neighborhood may therefore depend to a certain degree on the temporal activity patterns of its members. For both types of communication, the choice of neighborhood has a significant effect on its own, which is analyzed further via a post-hoc test.

We also test if the presence or absence of an explicit edge between two users (i.e., the *relationship type*) has an effect on the predictability of their addressive communication behavior. There are two weakly significant interactions ($p < 0.05$) that involve the relationship type, a three-way interaction with *observation date* and *time period length*, and a two-way interaction with *time period length*, but there is no significant main effect. The relationship type is a mensurative factor, in that it is predetermined by the data, and user pairs cannot be randomly assigned to a specific factor level. For that reason, the significance test for this factor is subject to confounding by unobserved covariates (Doncaster and Davey, 2007, p. 18). Considering the weak significance of the interactions, the absence of a significant main effect, and the potential for confounding, we do not feel justified in drawing any conclusions about the role of the *relationship type* from these results.

Figure 7.4 plots the effect of *time period length* on prediction error for non-addressive and addressive communication. In the case of non-addressive communication, one can observe a clear trend towards lower error with longer time periods. No such trend is visible for addressive communication. As becomes apparent in the following analyses, non-addressive

Table 7.6: Between- and within-subject effects for non-addressive communication (Twitter)

factors	d.f.	stat.	F	d.f. 1 / 2	p
intercept	1	0.841	12 332.37	1 / 2 340	< 0.001 ***
observation date	2	0.001	1.66	2 / 2 340	0.190
period length	2	0.042	50.72	2 / 2 340	< 0.001 ***
observation date × period length	4	0.004	2.31	4 / 2 340	0.055
neighborhood	1	0.086	5.86	37 / 2 304	< 0.001 ***
neighborhood × observation date	2	0.037	1.19	74 / 4 610	0.131
neighborhood × period length	2	0.041	1.30	74 / 4 610	0.046 *
neighborhood × observation date × period length	4	0.080	1.28	148 / 9 228	0.013 *

communication is driven by inertia to a large extent, so communication patterns are likely to be stable over longer periods of time. This concludes the interpretation of the ANOVA results.

The choice of neighborhood has a significant effect on the prediction accuracy. This effect can be characterized in greater detail by performing a post-hoc test. Tukey's HSD test identifies significant pairwise differences among the marginal prediction errors associated with the neighborhood definitions. Considering the presence of significant interactions between *neighborhood* and other factors, one has to consider that the test is insensitive to neighborhood definitions that perform above average on specific combinations of between-subject factor levels, and below average on others. We choose to ignore this issue and, in analogy to the initial filtering of the experiment results, focus on identifying neighborhood definitions that generally perform well across different combinations of experiment parameters. For both types of communication, the test yields a high number of homogeneous subsets at significance level $\alpha = 0.01$. All neighborhood definitions that pass the initial filtering also significantly outperform the baseline methods. Apart from the baselines, which reside in individual subsets of cardinality one, all subsets overlap. In other words, there is no single neighborhood or group of neighborhoods that significantly outperforms all others, but a gradual decline in performance from the best to the worst neighborhood definition. While the test allows us to identify a best subset, we cannot eliminate the remaining subsets as strictly inferior.

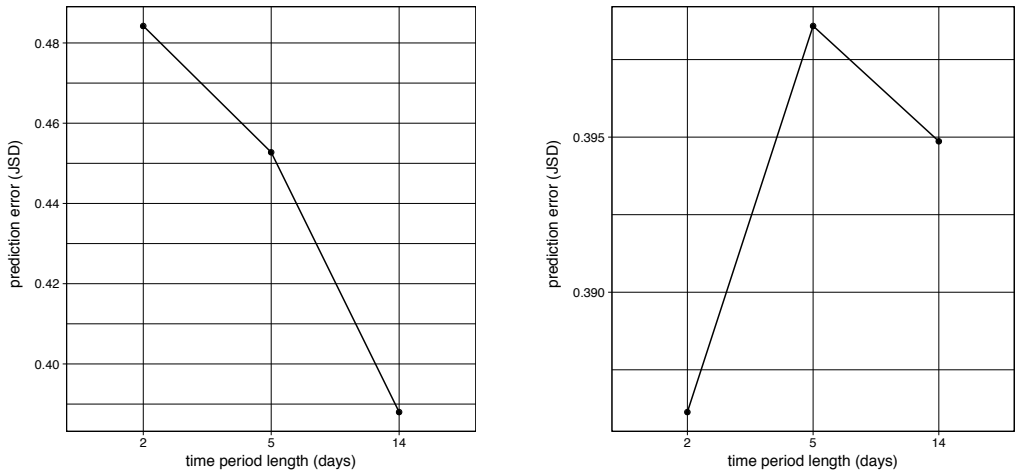
Table 7.8 lists the neighborhood definitions in the best-performing homogeneous subset for non-addressive and addressive communication. Membership in the best subset is evidence for a neighborhood's utility in the prediction of future communication behavior. At this point we cannot determine whether the utility of a particular neighborhood is due to its ability to identify sources of direct, indirect, or both kinds of exposure. However, the indi-

Table 7.7: Between- and within-subject effects for addressive communication (Twitter)

factors	d.f.	stat.	F	d.f. 1 / 2	p
intercept	1	0.89539	40 059.47	1 / 4 680	< 0.001 ***
observation date	2	0.00070	1.63	2 / 4 680	0.196
period length	2	0.00150	3.53	2 / 4 680	0.030 *
relationship type	1	0.00002	0.12	1 / 4 680	0.734
observation date × period length	4	0.00021	0.24	4 / 4 680	0.914
observation date × relationship type	2	0.00088	2.06	2 / 4 680	0.127
period length × relationship type	2	0.00172	4.02	2 / 4 680	0.018 *
observation date × period length × relationship type	4	0.00211	2.47	4 / 4 680	0.042 *
neighborhood	1	0.03437	4.72	35 / 4 646	< 0.001 ***
neighborhood × observation date	2	0.02606	1.75	70 / 9 294	< 0.001 ***
neighborhood × period length	2	0.02162	1.45	70 / 9 294	0.008 **
neighborhood × relationship type	1	0.00683	0.91	35 / 4 646	0.616
neighborhood × observation date × period length	4	0.04745	1.59	140 / 18 596	< 0.001 ***
neighborhood × observation date × relationship type	2	0.01496	1.00	70 / 9 294	0.476
neighborhood × period length × relationship type	2	0.01732	1.16	70 / 9 294	0.171
neighborhood × observation date × period length × relationship type	4	0.03158	1.06	140 / 18 596	0.306

Table 7.8: Best homogeneous subsets of neighborhood definitions (Twitter)

indicator	weight	non-addr.	addr.
explicit network, distance 1	*constant*	✓	✓
	out-degree (tf.)		✓
	number of shared neighbors	✓	✓
	reciprocity score	✓	
	number of clique-based communities (inv.)	✓	
	number of shared edge-based communities	✓	
	number of addr. messages (inv.)	✓	
	similarity of non-addr. messages	✓	
	similarity to medium avg.	✓	
	focus level	✓	
explicit network, distance 2	*constant*		✓
	number of shared neighbors	✓	
	reciprocity score		✓
	number of clique-based communities		✓
	number of shared edge-based communities	✓	
	similarity of addr. messages		✓
	similarity of non-addr. messages		✓
	similarity to medium avg.		✓
	focus level		✓
communities, clique-based	number of shared neighbors		✓
communities, edge-based	*constant*	✓	
	out-degree (tf.)	✓	
	ratio of in- and out-degree	✓	
	number of shared neighbors	✓	✓
	reciprocity score	✓	
	number of maximal cliques	✓	✓
	number of shared edge-based communities	✓	✓
	similarity of non-addr. messages	✓	
	similarity to medium avg.	✓	
	similarity to medium avg. (inv.)	✓	
	focus level	✓	



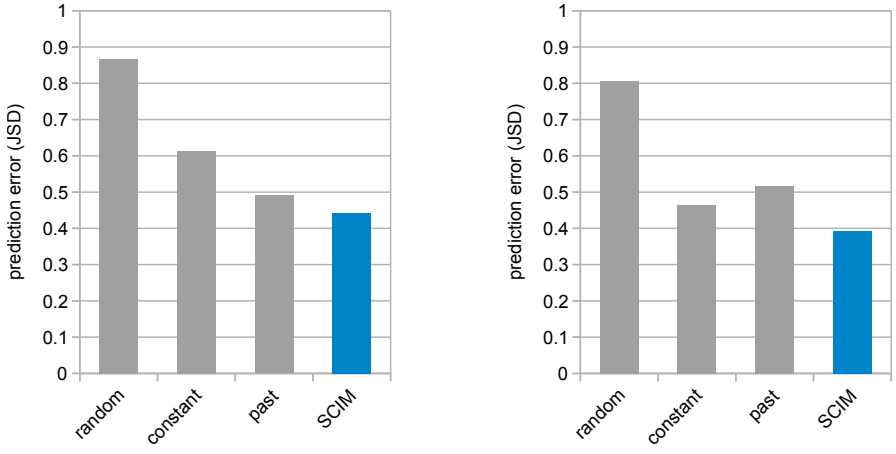
(a) non-addressive communication **(b)** addressive communication

Figure 7.4: Effect of experiment parameter *time period length* on prediction error (Twitter)

vidual contributions of indicator and weight function to the overall utility of a neighborhood can be identified to some extent: If a particular indicator function appears in the best subset in conjunction with the constant weight function, this constitutes a lack of evidence for the benefit of the other, more sophisticated weighting schemes. Most of the utility of the neighborhood definition must then be attributed to the selection of nodes by the indicator function.

For both types of communication, the explicit network provides more information to the predictor than the implicit communication network, and edge-based communities are more informative than cliques and clique-based communities. The best subset for non-addressive communication contains the indicator functions "distance one in the explicit network" and "edge-based communities" in combination with the constant weight function. The predictor can also make use of the larger set of nodes of distance two in the explicit network graph when paired with weighting by the number of shared neighbors or communities, which hints at the importance of socially cohesive local subgraphs. The prediction of addressive communication also benefits from selecting a larger set of nodes from the explicit graph, but there is no evidence for a beneficial effect of weighting. Edge-based communities do benefit from weighting by number of shared neighbors and communities. Weights that appear to be generally useful are the number of shared neighbors, the reciprocity score, the number of shared edge-based communities, content-wise similarity, and the focus level. A possible explanation is that locally dense, topically homogeneous neighborhoods are indicative of homophilous attachment, and derive their predictive value mainly from being sources of indirect exposure.

Conversely, one can look at the indicator and weight functions that do not pass the initial filtering step for most or all combinations of experiment parameters, and are therefore unlikely to build a neighborhood that is useful to the predictor. Complementary to the com-



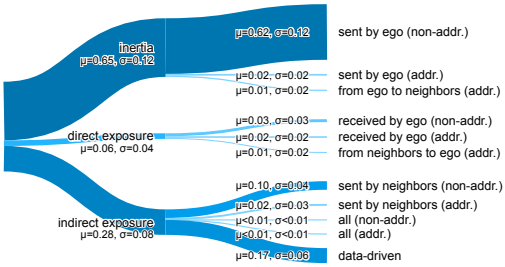
(a) non-addressive communication **(b)** addressive communication

Figure 7.5: Mean prediction error of the SCIM compared to the baseline predictors (Twitter)

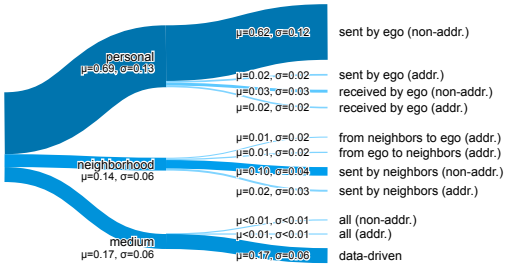
position of the best subset, indicator functions based on the communication graph do not perform well. This includes the indicator functions "largest maximal clique" and "clique-based communities", which use cliques of the communication graph. Given that on Twitter, by design, non-addressive surpasses addressive communication in volume, the corresponding neighborhoods may be too small and too sparsely connected to be useful. This would imply that communities either rarely manifest as fully connected subgraphs in the communication graph, or require a longer interaction period to be visible as such. Looking at weight functions that are frequently eliminated, independent of the type of communication, we find that message volume statistics, such as the number of sent messages or the number of times a message is shared by others, have low utility for the prediction task. The untransformed number of followers or friends is eliminated much more often than the logarithm of the counts, which supports our intuition that a high number of friends does not linearly translate to a high degree of influence.

Figure 7.5 compares the mean prediction error of the best homogeneous subset to the three baseline predictors. The SCIM outperforms all three baselines, with a 10% improvement over the best performing baseline in the case of non-addressive communication, and a 15% improvement for addressive communication. The constant baseline predictor performs much better for addressive than for non-addressive communication, which indicates a generally higher entropy of the topic distributions, likely related to the temporal sparsity problem discussed in section 3.5.3. The SCIM is consistently able to learn about the influence process under varying experimental conditions and make predictions with above-baseline accuracy. We interpret the improvement in prediction accuracy of the SCIM over a baseline model as its gain in explanatory capability.

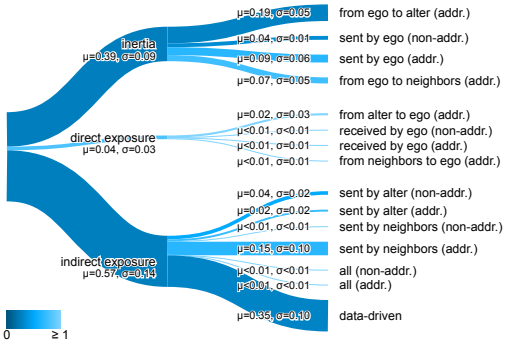
The remaining error can be partitioned in terms of the bias-variance decomposition: Bias refers to error caused by the inability of an overly simple model to fit to the data. Conversely, variance refers to overfitting, i.e., the inability to generalize due training on data



(a) non-addressive communication, grouped by role, $|S| = 22$



(b) non-addressive communication, grouped by scope, $|S| = 22$



(c) addressive communication, grouped by role, $|S| = 14$

(d) addressive communication, grouped by scope, $|S| = 14$

Figure 7.6: Mixture coefficients of the SCIM experiments in the best subset S (Twitter)

that is insufficient in volume or quality. Outside of synthetic problems, a certain amount of irreducible error exists independently of the complexity of the model and the quality and amount of training data. Because of its low parameter count and linear nature, the SCIM is a high-bias, low-variance model. Compared to the error caused by insufficient observational data, we expect that a much larger part of the error can be attributed to properties of the influence process that the SCIM, by design, cannot represent, such as individual differences in the reaction to the different types of exposure. Furthermore, the SCIM does not account for the effects of other cognitive processes besides social influence. However, there is evidence for the existence of general, learnable characteristics of the influence process, which validates the modeling assumptions of the SCIM and motivates the following analysis of the mixture coefficients.

Figure 7.6 visualizes the mixture coefficients as a tree, where each leaf corresponds to a coefficient, and the parent nodes represent either role or scope as listed in table 7.2. The line width is proportional to the mean value of the coefficient across all experiments in the best subset, while the lightness of the color is proportional to the ratio of mean and standard deviation: Minimum lightness corresponds to a standard deviation of zero, and maximum lightness to a ratio of one or above. The darker the color, the less affected is the coefficient by variation of the experiment parameters.

The basic mechanism of influence, as learned by the SCIM, is the same for addressive and

non-addressive communication. Both are clearly driven by inertia to a large degree, and exhibit a surprisingly small effect of direct exposure. According to the principle of locality it should be a strong source of influence, yet the corresponding coefficients are close to zero. The effect of indirect exposure from the neighborhood is in proportion with our expectations. It follows that the social neighborhood derives its predictive utility almost exclusively from its ability to capture the effects of strong unobserved influencers. Most of the remaining weight is assigned to the data-driven component θ^d, hinting at the presence of strong influencers that uniformly act upon the whole network, but are either not directly observable or not covered by any of the proposed neighborhood definitions. Non-addressive communication is mainly influenced by other non-addressive communication, and the same pattern is visible in slightly weaker form for addressive communication. Non-addressive communication is mainly driven by inertia, with the content of past non-addressive messages carrying more predictive value than any other component, while in the case of addressive communication, indirect exposure has a higher predictive value. A notable general effect is that components that aggregate the relationship-topic distributions of a large number of users have little predictive value, independent of the type of communication. This is the case for the topic distribution of non-addressive messages received by the ego, as well as the aggregated distributions of all communication within the medium.

The experimental results support some of the fundamental assumptions of the SCIM, including the decomposability of influence into the effects of inertia and exposure, but do not fully agree with the hypothesis of locality. The strong effect of inertia and the weaker, but substantial contribution of the social neighborhood via indirect exposure are consistent with the hypothesis, but we also observe strong effects that likely originate from external influencers, and almost no traces of direct exposure. We conclude that these two deviations from the predicted outcome deserve further attention: First, strong external influencers, possibly celebrities and news media, appear to play a bigger role in the influence process than expected. Second, the lack of observable, direct influence is consistent with the results of the experiments on micro-level social influence in chapter 6. If locality is not the main factor that determines the strength of the influence of an individual on another, we need to revisit the conclusions of the micro-level experiments and investigate alternative hypotheses, including complex contagion. Still, the SCIM is able to successfully predict future behavior while relying mainly on local information, so locality is the element that makes the detection of social influence tractable. Considering the low average completeness of nodes in the crawled subgraph of the follower graph (see section 3.6.1) and the strong role of indirect exposure in the SCIM, indirect exposure appears to be a promising mechanism for coping with the limited observability of online interactions.

7.3.2 Comparative Analysis of Social Platforms

Are these findings specific to Twitter, or do they generalize to other online social platforms? To answer this question, we perform the basic prediction experiment on data obtained from a wide range of online services. Since these datasets are smaller than the Twitter dataset used in the previous round of experiments, we analyze the results using a simplified ANOVA procedure with fixed experiment parameters *observation date* (most recent), *period length*

Table 7.9: Sample sizes of the SCIM experiments on different datasets

dataset	nodes (non-addr.)	edges (addr.)
Twitter	8 910	3 412
Facebook	659	120
e-mail (HT-en)	8	145
e-mail (HT-it)	10	184
e-mail (Enron)	10	41

(14 days), and *relationship type* (all edges). The only variable is the choice of neighborhood, which is a within-subjects factor, so no partitioning of the sampling units is required. Compared to the full ANOVA procedure, we lose the ability to evaluate the stability of the model over time and the effect of the interval length on the influence process. The sample size of each dataset, i.e., the size of the test set of the associated experiments, is given in table 7.9. Clearly, partitioning the smaller datasets is not feasible. The rules for classifying a message as addressive or non-addressive depend on the platform and are necessarily subjective to some extent. Still, the large difference in sample size for the two forms of communication is consistent with our observation that each platform has a dominant mode of communication. The modern social network services encourage non-addressive communication, while e-mail is addressive by default.

The indicator and weight functions are designed to be applicable to a broad range of online social platforms, with the exception of certain indicator and weight functions that depend on the presence of an explicit social network, or require the edges of the network graph to be directed. The communication graph is a directed graph by definition, but we observe that in the graphs of all three e-mail datasets, each edge has a counterpart in the opposite direction. Presumably, one-sided messaging is rare in a workplace communication setting. Since the direction of edges in the e-mail communication graphs provides no information, we treat these graphs as undirected for the purpose of the experiments.

Prior to the experiments, it is important to confirm that the neighborhood definitions produced by the indicator and weight functions are meaningful in the context of each dataset. A heuristic to gauge the utility of a neighborhood definition is to check if it produces non-empty neighborhoods for the majority of users. The indicator and weight functions appear to be well applicable to the HackingTeam e-mail data. In the associated experiments, the ratio of empty neighborhoods consistently remains below the threshold of 0.5. When applied to the other datasets, a number of functions frequently produce empty neighborhoods: Indicator “clique-based communities” and weight “number of shared clique-based communities” probably suffer from the sparseness of the observable communication network. The weight functions related to sharing behavior are frequently non-useful, because observable sharing is rare outside of Twitter.

ANOVA reveals that the effect of neighborhood choice is generally significant with $p < 0.001$ for all datasets except HT-en, where $p < 0.01$. The lower the sample size, the harder it is for the post-hoc test to identify small, clearly delineated homogeneous subsets. This is

particularly noticeable with the e-mail datasets: Although the SCIM outperforms the baseline predictors on all datasets, regardless of the chosen neighborhood definition, we cannot establish the statistical significance of that improvement in the non-addressive case. For the Enron dataset, even in the addressive case there is not enough statistical power to establish the significance of the improvement over the "previous behavior" baseline. It does significantly outperform the other two baselines.

The statistical evaluation of the Twitter experiments benefits from the higher sample size compared to the previous experiments. The higher sample size yields a better characterization of the relative utility of the different neighborhood definitions. In the non-addressive case, we are able to identify a single neighborhood that significantly outperforms all others, "nodes of distance one in the explicit network, weighted by similarity of non-addressive messages". In the case of addressive communication, we cannot strongly separate the best subset from the others, but the subset is very small, consisting of the two neighborhoods "nodes of distance two in the explicit network, weighted by similarity of addressive messages" and "nodes of distance two in the communication network, weighted by similarity of addressive messages".

The experiments on the non-Twitter datasets identify best subsets that are almost mutually disjoint, with only one neighborhood definition appearing in more than two subsets. In the best subset associated with non-addressive communication on Facebook, most neighborhoods are built using "nodes of distance two in the explicit network" or "nodes of distance one/two in the communication network" as the indicator function. The latter two appear in combination with the constant weight function. The best subset for addressive communication invokes the indicator functions "nodes of distance one in the explicit network", "largest maximal clique", and "clique-based communities". Both the explicit and the implicit social graph appear to be useful for the identification of social neighborhoods.

The best subsets associated with the e-mail datasets HT-en and HT-it are unexpectedly different: In the best subset of HT-en, most neighborhoods are built using the indicator functions "largest maximal clique" and "clique-based communities", while the best subset of HT-it only contains the two neighborhoods "nodes of distance one in the communication network, weighted by number of addressive messages sent" and "edge-based communities, weighted by number of non-addressive messages sent". The best subset associated with the Enron dataset is similar to that of HT-en, mostly consisting of neighborhoods that use "largest maximal clique" as the indicator function. For the comparatively small e-mail datasets, which are also rather homogeneous in terms of social context, locally dense subgraphs appear to be more informative than the plain communication graph.

We compare the predictive accuracy of the SCIM on the different datasets. Since the performance of the SCIM as well as the performance of the baselines depend on the dataset, the results can only be meaningfully compared in terms of improvement over the strongest baseline, as shown in figure 7.7. Compared to the previous experiments on Twitter data, the SCIM performs below average for the chosen parameters in the prediction of non-addressive communication. With the exception of non-addressive communication on Twitter, the strongest baseline in all sets of experiments is previous behavior. This implies that the bottleneck in prediction is not insufficient information, which would manifest in topic distributions that are dominated by the prior distribution, and result in a strong "constant" baseline.

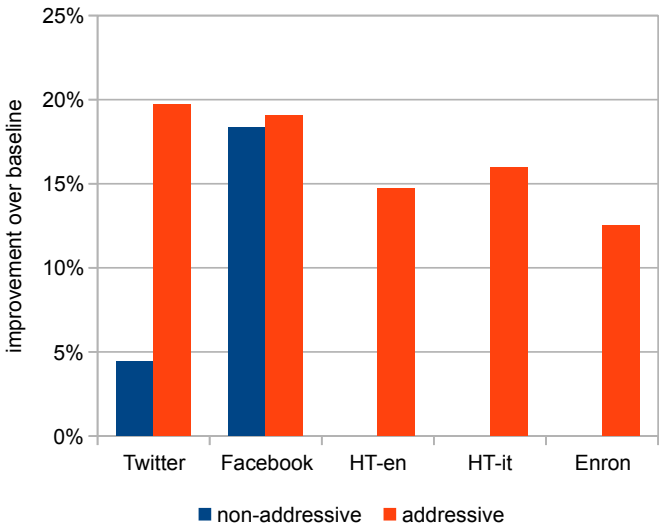
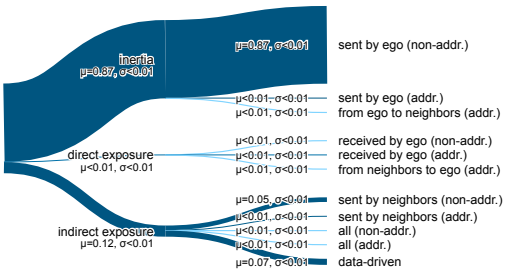
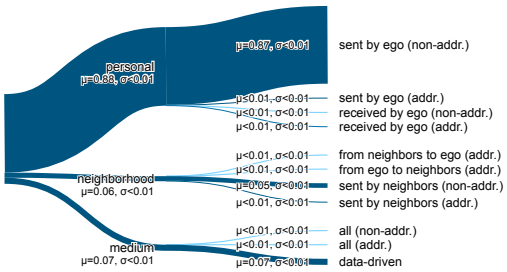


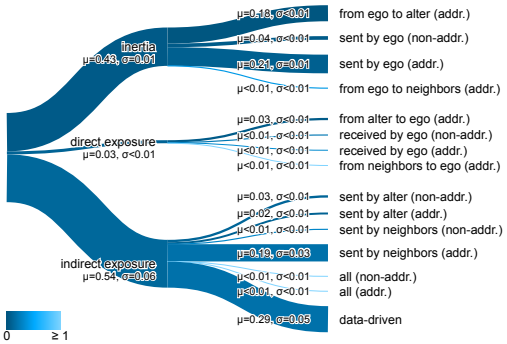
Figure 7.7: Comparison of SCIM performance (% improvement over strongest baseline) for different datasets



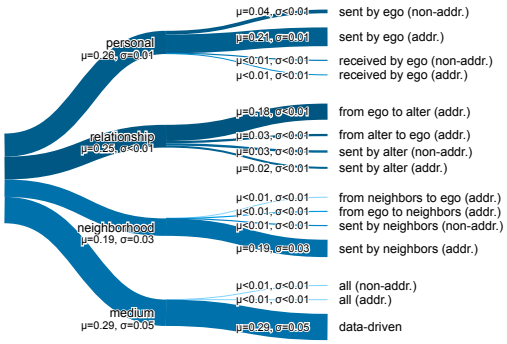
(a) non-addressive communication, grouped by role, $|S| = 1$



(b) non-addressive communication, grouped by scope, $|S| = 1$

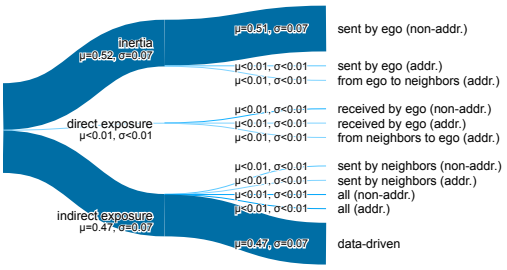


(c) addressive communication, grouped by role, $|S| = 2$

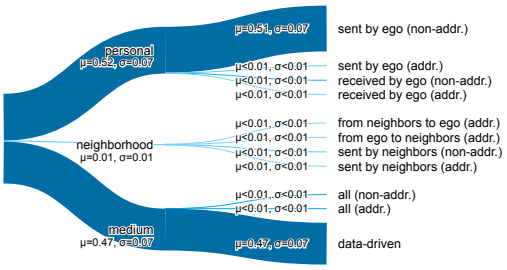


(d) addressive communication, grouped by scope, $|S| = 2$

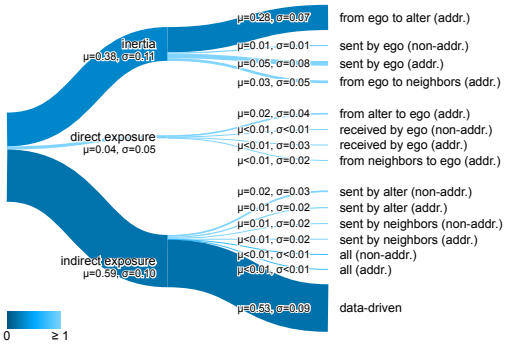
Figure 7.8: Mixture coefficients of the SCIM experiments in the best subset S (Twitter, simple ANOVA)



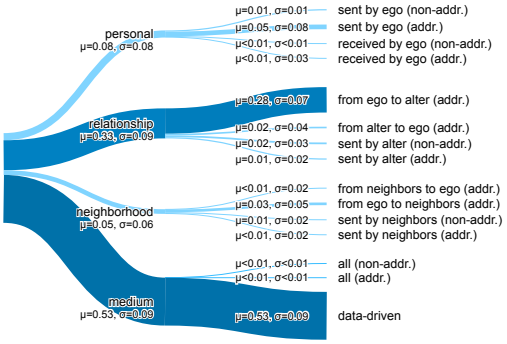
(a) non-addressive communication, grouped by role, $|S| = 19$



(b) non-addressive communication, grouped by scope, $|S| = 19$

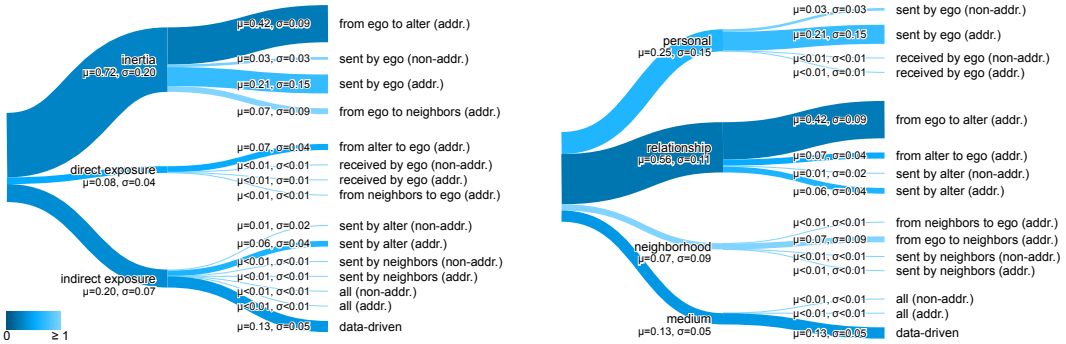


(c) addressive communication, grouped by role, $|S| = 15$



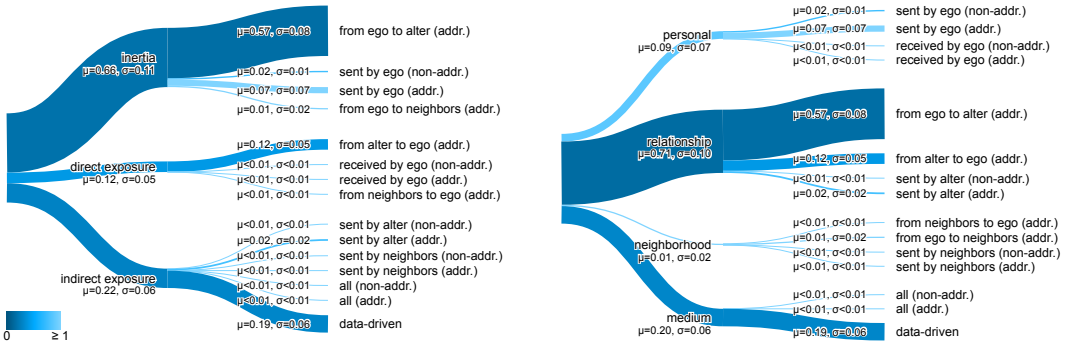
(d) addressive communication, grouped by scope, $|S| = 15$

Figure 7.9: Mixture coefficients of the SCIM experiments in the best subset S (Facebook)



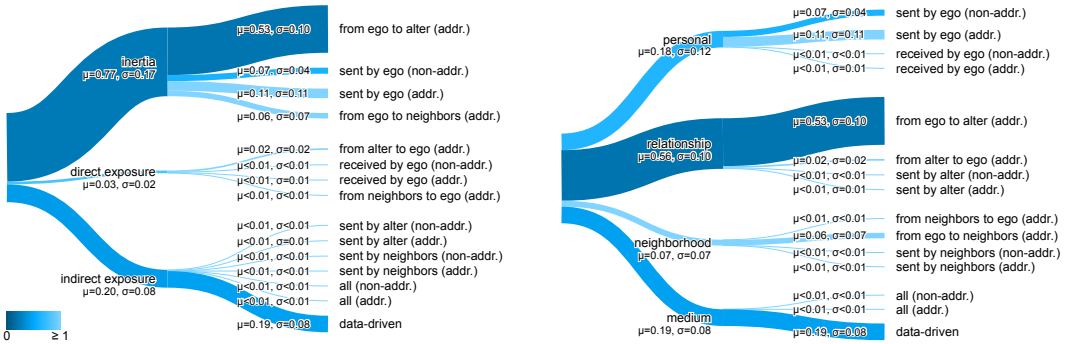
(a) addressive communication, grouped by role **(b)** addressive communication, grouped by scope

Figure 7.10: Mixture coefficients of the SCIM experiments in the best subset S (HT-en, $|S| = 10$)



(a) addressive communication, grouped by role **(b)** addressive communication, grouped by scope

Figure 7.11: Mixture coefficients of the SCIM experiments in the best subset S (HT-it, $|S| = 2$)



(a) addressive communication, grouped by role **(b)** addressive communication, grouped by scope

Figure 7.12: Mixture coefficients of the SCIM experiments in the best subset S (Enron, $|S| = 12$)

Comparing the mixture coefficients of the SCIM fit to the different datasets, as shown in figures 7.8, 7.9, 7.10, 7.11, and 7.12, we find major systematic differences in the coefficients. As in the previous section, we interpret the magnitude of a coefficient as the utility of the corresponding component for the prediction task. The coefficients of the SCIM fit to Twitter data are highly similar to those obtained in the previous round of experiments. The prediction of non-addressive communication is primarily informed by inertia, and to a lesser degree by indirect exposure. In the prediction of addressive communication, inertia and indirect exposure carry similar weight.

The SCIM coefficients fit to Facebook data assign almost all weight to two components: the data-driven component and, depending on the type of communication that is to be predicted, either the non-addressive or the addressive messages sent by the user under consideration. The prediction is almost exclusively driven by inertia and strong, consistent external influencers, but in contrast to the other datasets, we fail to learn about these external influencers via indirect exposure. The HackingTeam e-mail datasets are the only case where we find a small, but noticeable effect of direct exposure. In the cases of all e-mail datasets, the SCIM draws comparatively little information from the neighborhood and the data-driven component. Small, socially cohesive teams may be more conductive to local influence and, accordingly, less affected by external influencers. That notwithstanding, the unexpectedly small effect of direct exposure appears to be a general pattern across all datasets.

7.4 Follow-up Experiments

The SCIM experiments, while comprehensive, leave some open questions. For one, how do changes to the underlying topic model affect the performance of the SCIM? Another question is how the predictive accuracy of the SCIM can be improved without sacrificing its interpretability. The prediction error of the SCIM can, to a large part, be attributed to systematic error due to overly restrictive model assumptions. By testing several hypotheses, we attempt to determine which assumptions most limit the model's ability to fit to the observation data. Finally, we examine the effect of the observation date on the prediction accuracy.

7.4.1 Topic Modeling Variants

As discussed in section 5.2, when fitting a single topic model to data from the whole observation period and obtaining topic distributions for shorter intervals of time by querying, the topics will "leak" information about the future into the prediction experiments, which potentially inflates the accuracy. Since most of our conclusions from the preceding experiments rest on the assumption that the SCIM learns about the influence process from past observations, it is important to rule out that the SCIM is able to obtain additional information through a side channel. Conversely, temporal sparsity (see section 3.5.3) in conjunction with the typical brevity of individual messages and social media posts (see section 4.2.1) may impair the quality of the topics and lower the information content of the topic distributions, with a potential negative effect on prediction accuracy.

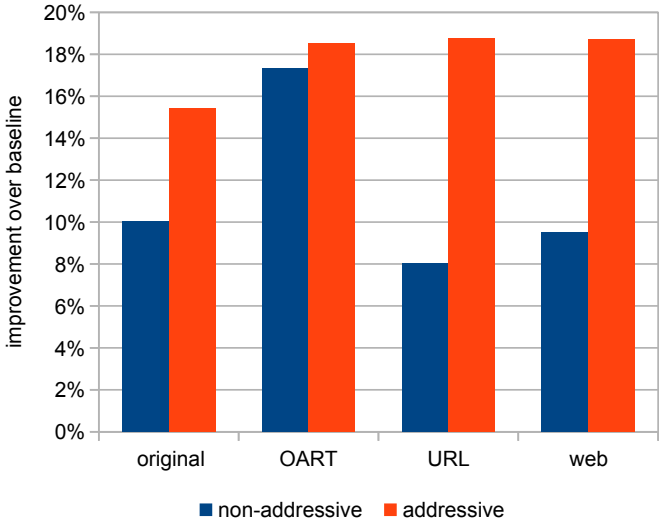


Figure 7.13: Comparison of SCIM performance (% improvement over strongest baseline) for different topical representations of the Twitter dataset

In order to determine the magnitude of these effects and evaluate strategies for their mitigation, we repeat the experiments on the Twitter dataset, but make changes to the way the topic distributions are computed. To test for an effect of information leakage, instead of obtaining topic distributions for the interaction and evaluation periods by querying a single overall ART model, we query a topic model that is incrementally updated with data from each time period, as described in section 4.1.1. If information leakage is the main factor that allows the SCIM to outperform the baselines, we expect to see a noticeable reduction in accuracy when querying the incremental topic model. Furthermore, we test whether the two data augmentation schemes described in section 4.2.2 are able to mitigate the negative effects of a low average message size. Both augmentation strategies involve the analysis of tweeted URLs, which would otherwise be discarded in the preprocessing phase. One strategy is to resolve and tokenize these URLs, the other is to extract the textual content of the referenced websites.

Since a change to the data representation may affect the performance of both the SCIM and the baselines to varying degrees, we measure the effect of these topic model variants on the performance of the SCIM in terms of the improvement of prediction accuracy over the strongest baseline. Figure 7.13 compares the incremental topic model (*OART*) and the two data augmentation schemes (*URL*, *web*) to the unmodified topic model. For non-addressive communication, the strongest baseline is always past behavior, while for addressive communication the constant predictor is always the strongest.

The incremental topic model consistently outperforms the regular topic model, despite only using observations from the current and past interaction periods for parameter fitting. While we cannot rule out that information leakage in the regular topic model positively affects the predictive accuracy of the SCIM, the same accuracy can be attained and even exceeded without using information from future observations.

The mixture coefficients of the SCIM assign less weight to inertia and more weight to sources of indirect exposure, which points towards a higher information content of the social neighborhood. The results of ANOVA are largely consistent with the regular topic model. A notable exception is that, in the case of non-addressive communication, the incremental topic model gains a significant main effect of the *observation date* ($p < 0.01$) and significant interactions with *time period length* ($p < 0.05$) and *neighborhood* ($p < 0.001$). The main effect does not take the shape of monotonous growth or decline of prediction error. In the case of addressive communication, neither *observation date* itself nor its interaction with the *neighborhood* is significant. A possible explanation for these results, which appear rather disparate at first glance, is as follows: Incremental topic modeling emphasizes the differences between intervals of time instead of smoothing over them. Consider the topic distribution in the absence of data θ^p (as defined in section 7.1.1), which, due to optimization of the prior parameter α, can be understood as the "average" or "typical" topic distribution for a given dataset. When querying a single topic model, θ^p is constant over time by design, while incremental parameter fitting yields a new θ^p for each interval. Since a predictor that always outputs θ^p is a strong baseline, the JSD of θ^p of two successive intervals is an indicator of predictability. When using incremental topic modeling, some intervals are more predictable than others, which explains the significant main effect of *observation date*. Since incremental topic modeling exacerbates the effects of temporal sparsity, this effect is less pronounced when analyzing addressive communication.

Both data augmentation schemes slightly raise the sample size from 261 to 267 units per sample. This is due to tweets that only contain a URL and would otherwise be discarded. Augmentation fails to improve the accuracy of prediction of non-addressive communication, but has a beneficial effect on the prediction of addressive communication. In the latter case, the volume of available data is lower. It stands to reason that, for a given set of tweets, the information obtained from the content of referenced websites or their URLs is either too noisy or insufficiently connected with the communicative intent of the tweets, so that it is only useful to the SCIM if the tweets themselves provide very little information. Using the content of linked resources for data augmentation could potentially mitigate the problems associated with topic modeling of very short documents, but more investigation into its failure modes is required before this technique is generally applicable.

7.4.2 Error Analysis

When comparing the SCIM to the abstract cognitive model of influence presented in section 5.1, it is evident that a number of simplifying assumptions have been introduced, which, while being backed by previous research and serving to make the model tractable, clearly restrict its expressiveness, and thus its predictive accuracy. Since the results presented in the previous section show potential for improvement, we attempt to identify the limiting factors with the strongest negative impact on prediction accuracy. The assumption of locality of influence is tested in the course of the regular experimental evaluation of the SCIM, which leaves two potentially major sources of error to be examined here.

According to the theoretical model of influence that motivates the SCIM, social influence results from exposure to the behavior of others, so not being able to observe all sources of

exposure is an obvious source of error. In an SNS like Twitter or Facebook, a user's direct neighborhood in the explicit social network graph determines what information that user is exposed to. The SCIM operates under the assumption that the social environment of each node has been completely observed. However, when acquiring data from a social network service, subsampling is usually unavoidable, either due to the high volume of available data or restrictions imposed by the service provider. In the cases of Twitter and Facebook, from which we acquired data by crawling, we did not have any prior knowledge about the network structure and were limited in the number of user profiles we could access. To evaluate the effect of missing nodes on the prediction accuracy, we revisit the notion of a node's *local completeness*, which is defined in section 3.6.1 as the ratio of neighbors in the sample graph to the number of neighbors in the original, full network graph.

The presence of a negative correlation between local completeness and prediction error would be evidence for a detrimental effect of missing observations. We compute the linear correlation between the local completeness of a node and the associated non-addressive prediction error, as well as the linear correlation between the local completeness of the two endpoints of an edge and the addressive prediction error associated with that edge. For both the Twitter and the Facebook dataset, all correlation coefficients are very close to zero and vary in direction ($-0.03 < r < 0.06$). The weak correlation is consistent with the negligible role of direct exposure observed in the main experiments, and supports our hypothesis that inference about indirect exposure via the aggregated behavior of the social environment makes the SCIM more robust towards missing observations.

E-mail differs from SNS in that there is no explicit social network, and any user may receive messages from — and thus be directly exposed to the behavior of — any other user. Each of the two e-mail datasets contains the complete mailbox contents of members of a single social group (the employees of a company). Under the assumption that messages are never deleted, a user's mailbox contains all information that user has ever been exposed to, and therefore all communication among the group members can be completely recovered from the mailbox data. A mailbox also contains any messages that a group member has received from outsiders. These possibly make up a large proportion of that member's exposure, and therefore may have a proportionally strong effect on his or her future behavior. However, the SCIM, in its current shape, cannot make use of these messages. Being concerned with the effect of exposure on communicative output, it can only be applied to users for whom both is sufficiently observable, which excludes outsiders. In consequence, we restrict our analysis to group members and ignore the individual effects of outsiders on group members. The effect of this decision cannot be easily measured.

Another modeling assumption of the SCIM is motivated by the decision to limit the analysis to general, network-wide characteristics of the influence process: the assertion that a single set of mixture coefficients is sufficient to describe the influence process for all users. Looking at the Twitter experiments, the average training set error of the SCIM for the best subset of neighborhoods is 0.435 for non-addressive, and 0.387 for addressive communication. On the test set, the error increases by 1% (to 0.441) for non-addressive, and by 2% (to 0.393) for addressive communication. The small gap between training and test set error suggests the existence of sources of variance that are not taken into account by the model. To gauge whether the assumption of the influence process being homogeneous across all

users is responsible for the lack of fit, we look at the average training error of the SCIM fitted to individual users. This can be interpreted as a lower bound on the prediction error that could be achieved by an ideal model that takes individual differences into account in an optimal way. The data driven component θ^d is removed from the model, because otherwise a training error of zero could be trivially achieved for a single user.

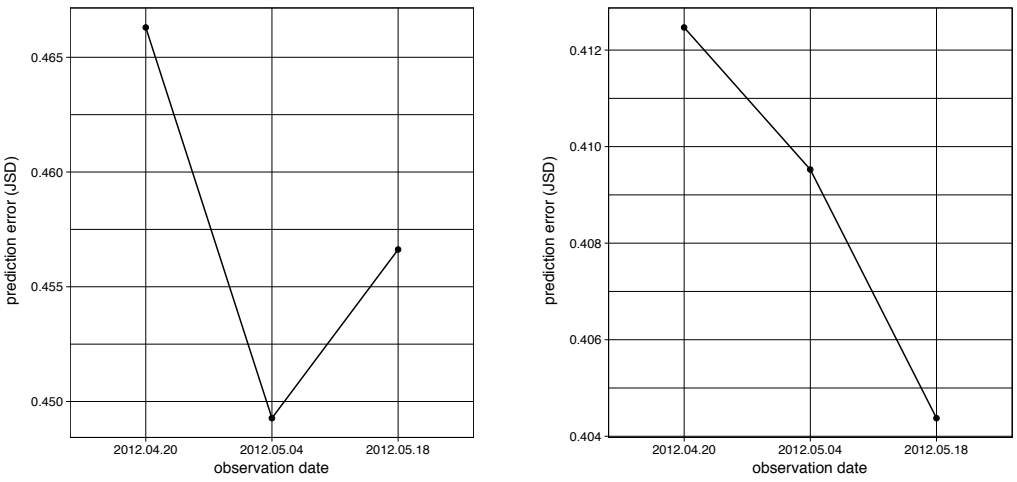
On the Twitter dataset, the resulting average per-user training error is noticeably lower, 0.346 (standard deviation $\sigma = 0.004$) for non-addressive, and 0.251 ($\sigma = 0.002$) for addressive communication. The other datasets show the same trend: The per-user training error is always substantially lower than the error of the SCIM (between 16% and 39%), but never zero. The test set error is always within a few percent of the training set error. One one hand, this shows that the optimal mixture coefficients vary among users, and the SCIM could be improved accordingly, e.g., by classifying each user according to his or her communication behavior and learning a separate set of coefficients for each class. On the other hand, the magnitude of the remaining error indicates that individual differences are not the only source of error that an improved model of social influence would need to take into account.

7.4.3 Change of Prediction Accuracy Over Time

Since the goal of the SCIM experiments is to identify general characteristics of the influence process, it is important to understand how and why its predictive performance varies over time. By evaluating the results of experiments on the Twitter dataset via the ANOVA procedure, we find that the effect of the choice of observation date on prediction accuracy, as shown in figure 7.14, is not statistically significant. However, in the case of addressive communication, there is a strongly significant two-way interaction with the neighborhood definition and a three-way interaction with neighborhood definition and period length. In other words, there is no indication that some intervals are generally harder for the SCIM to predict, but there is an effect of time on the utility of specific neighborhood definitions. Concluding our evaluation of the SCIM, we perform an explorative, qualitative-leaning analysis to investigate the origins of this effect.

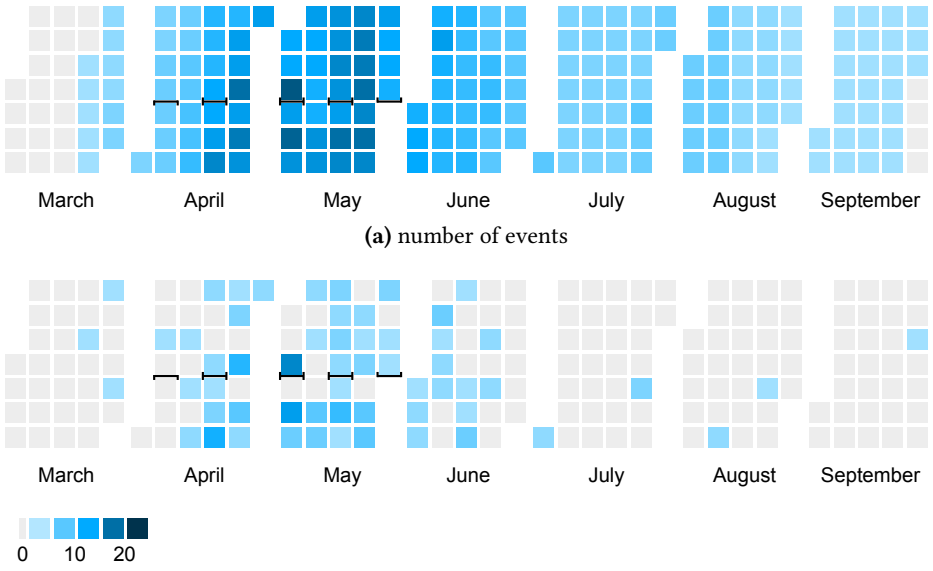
Our hypothesis is that real-world events have an effect on the predictability of future behavior from current observations, and therefore on the accuracy of the SCIM. An event can be anything that happens within a specific, previously announced time frame and receives social media attention, e.g., a conference, a holiday, or a large-scale sports event. The reception and discussion of events on social media can be broadly divided into three phases: Before the event, invitations are sent, advertisements are published, and excitement builds up. During the event, participants and outsiders discuss the current proceedings. After the event, people tend to reflect on their experience and provide feedback. People exhibit distinctly different behavior in each of the three phases, so a phase transition marks a change in the behavior of a potentially large group of people. We hypothesize that the more phase transitions occur within an interval of time, the harder it is to predict the behavior of people within that interval.

To build a list of events that are discussed in the 56-day subset of the data we are using for the SCIM experiments (covering May and most of April 2012), we fit an ART model to the data and manually review the top-50 most probable words for each of the 150 topics.



(a) non-addressive communication **(b)** addressive communication

Figure 7.14: Effect of experiment parameter *observation date* on prediction error (Twitter)



(a) number of events

(b) number of transitions

Figure 7.15: Event density in the Twitter dataset (14-day interaction periods of the SCIM experiments marked in black)

We identify 69 events, most of which are strongly associated with a particular topic. For example, one topic assigns high probability to words related to criminal trials, including "prosecutor", "trial", and "court", and makes reference to three events: the trials of Anders Breivik, Charles Taylor, and Ratko Mladić. Each of these events is only referenced from this one topic. Eight events are referenced by more than one topic. The most-referenced event is the final match of the annual UK football competition "FA Cup", which in 2012 was contested between the football clubs ("FC") of Chelsea and Liverpool. The event is referenced by four topics: one topic assigns high ranks to the names of UK football clubs and words that are generally related to football, two topics are related to Liverpool FC, and one is related to Chelsea FC. The assignment of a start and end date to an event can be somewhat subjective, especially if the event formally spans a long time period, but has phases of rising and waning public attention. Three events do not have a clearly defined start or end, so we arbitrarily choose a plausible date: the investigation into the death of Gareth Williams, the Irish referendum on the European Fiscal Compact, and investigations associated with the News International phone hacking scandal.

Figure 7.15a shows the temporal distribution of events that were mentioned in tweets from the two-month observation period. As expected, the event density is highest within the observation period. While there is long-term anticipation of upcoming events several months away, retrospection is limited to the past two weeks. Figure 7.15b shows the number of transitions, i.e., the number of events that start or end on a particular day. Counting the transitions in the 14-day intervals that correspond to the interaction and prediction periods of the SCIM experiments, there are eight transitions in the first interval, 47 in the second, 37 in the third, and 30 in the fourth. If we take the number of transitions as an indicator of (non-)predictability, the rise in predictability from the second to the fourth interval matches the declining error of the corresponding experiments shown in figure 7.14b. These findings are purely exploratory, and should not be taken as solid evidence. Yet, they suggest that the connection between event density and predictability deserves further attention.

7.5 Discussion

Two main conclusions can be drawn from the results of the meso-level social influence experiments: First, the results of our study offer a novel point of view on the relationship of addressive and non-addressive online communication. A major difference between Twitter and other social media is the high volume of unambiguously non-addressive communication. Although this can be mainly attributed to the design of Twitter, which makes public messaging the default, the messaging behavior of individual users exhibits high variability, with the per-user proportion of addressive communication having a one-SD range of 12% to 60%. Our experiments show that the difference between the two modes of communication is not merely a matter of personal preference, but also manifests itself in the effect of received messages on future behavior. Notably, non-addressive communication on Twitter is generally more resistant to influence than addressive communication. Across all examined social platforms, non-addressive communication receives stronger influence from other non-addressive communication, and addressive communication from other addressive com-

munication.

The question of whether Twitter is a social networking service or merely a platform for information dissemination has been previously discussed from multiple perspectives (Huberman et al., 2009; Kwak et al., 2010; Wu et al., 2011; Myers et al., 2014). We contend that the Twitter social network is more naturally described as the result of superimposing the networks formed by different types of communication: the follower network, which governs the flow of non-addressive communication, and the implicit network formed by addressive communication. This duality may also explain why community finding algorithms that are reported to work well on social network datasets (Palla et al., 2005; Ahn et al., 2010) fail to identify meaningful communities in the Twitter follower graph. On the other social platforms, addressive and non-addressive communication are not as clearly separable as they are on Twitter, so it is harder to draw definite conclusions.

Second, seeing that the SCIM is consistently able to outperform the baseline predictors across datasets obtained from different platforms, we are able to conclude that the local context or social environment of a node contains sufficient information to predict future behavior to a certain extent, and the model actually learns to exploit these sources of information. A fundamental mechanism of the SCIM is the decomposition of social influence into inertia and exposure, which is further subdivided into direct and indirect exposure. According to the principle of locality, the strength of influence decreases with increasing difference of the social contexts of influencer and influencee. We would therefore expect that inertia, being an internal process of the influencee, contributes the most information to the prediction, followed by direct exposure to the behavior of neighbors in the social network graph, and finally, indirect exposure to arbitrary other sources of influence, whose social contexts may bear little resemblance to that of the influencee. However, the predictive value of the mixture components of the SCIM, as observed in the experiments, does not fully agree with these expectations. The model draws a lot of information from indirect exposure, and surprisingly little from direct exposure.

The unexpectedly low value of direct exposure implies that locality on its own is not sufficient to explain why the SCIM is able to outperform the baselines. If the assumption of locality of influence does not hold, sources of influences within and outside of a medium are comparable in strength, and unobserved strong influencers at distant parts of the social network may exert as much influence as direct neighbors, and a large part of an actor's overall exposure is not practically observable. Indirect exposure preserves the behavior of strong, unobserved influencers and even allows information to cross medium boundaries. We find that the social neighborhood of a node is much more useful in its role as a source of indirect, rather than direct, exposure. In the case of Twitter, the best-performing neighborhood definitions favor nodes that are structurally or content-wise similar to the ego; a homophily-based strategy of constructing a neighborhood of users who are likely to be exposed to similar external influences. The high explanatory value of the data-driven component across all datasets is indicative of medium-wide indirect influence effects that are not captured by the SCIM. These experimental results suggest that a clear distinction has to be made between a node's direct neighborhood in the communication graph, which is the temporally stable group of peers that a person regularly interacts with, and the node's "influence neighborhood", which is the set of neighbors whose aggregated behavior best

reflects observable and unobservable influencers that act on the node.

By formulating the detection and analysis of influence at the meso level as a prediction problem, we are able to avoid the question of causality in the experiment design and evaluation of the results. Our partial success in predicting future behavior from direct and indirect exposure is due to temporal correlation between the prediction target and the social environment, but we cannot tell if, and to what extent, any causal effects are involved. Social influence is fundamentally a causal phenomenon, so without specifically testing for causality, we cannot rule out alternative hypotheses that would explain the temporal correlation, such as homophily (Anagnostopoulos et al., 2008; Aral et al., 2009; La Fond and Neville, 2010): two people are more likely to form a social relationship if they or their social context are similar in some respect, and these similarities lead to correlations in behavior. In chapter 6 we identify a number of fundamental problems that prevent testing for the causal effects of social influence at the level of individual relationships, even in the restricted sense of Granger causality. We expect that the problem of distinguishing influence from homophily at the meso level suffers from the same problems, see for example Shalizi and Thomas' criticism (2011) of the homophily test of Anagnostopoulos et al. (2008). A more principled approach to causality testing is demonstrated by Zhang et al. (2015), who construct a control group to approximate counterfactual reasoning.

Our experiments provide evidence for the presence of strong external influencers. Having investigated the nature of social influence on increasingly higher levels of aggregation, we wonder if social influence might be even better understood as a macro-level phenomenon that mainly involves the feedback between entities of high visibility and communicative reach, such as celebrities or media outlets, and their audience. We leave this question to future work.

8 Conclusion

Conceptually, this thesis is divided into two main parts, which address the problem of learning about social relationships from observed individual behavior on online social platforms in different ways. The first part deals with the characterization of social relationships, the second with the detection of social influence. We discuss the main results of each part in turn, and place them in a common context.

Characterization of Social Relationships

Any attempt at reasoning about the nature of social relationships must be grounded in theoretical work from the fields of sociology and social psychology, which are principally concerned with human social behavior as an object of study. However, after reviewing a body of work from these fields, we find a conceptual gap, i.e, a disconnect between the level of abstraction at which theories from the social sciences operate, and at which observations of social behavior can be made in an online environment. To some extent this is a problem caused by technological progress and easier access to observational data in large volume ("big data"), which promises new kinds of insight, but renders analytical methods that rely on manual processing and interpretation impracticable. Collins' approach of analyzing social interactions at the level of individual utterances has been referred to as "radical microsociology" (Marsden, 1990), yet on Twitter, it is common for a popular hashtag to receive thousands of tweets per minute (Bruns and Stieglitz, 2013). The result is that studies which apply methods of computer science to the analysis of social phenomena are often insufficiently motivated by extant theories.

A social relationship between two individuals becomes manifest and observable when they interact. On most online social platforms, the dominant form of interaction is textual communication. We therefore argue that representing relationships by the content of the associated communication is useful for exploratory data analysis as well as an intermediate representation, on which more targeted statistical models can operate. This claim is supported by the results of a user study (section 2.2.2), where even simple keyword extraction is judged by human annotators to produce a meaningful visual representation of their relationships. From a methodological point of view, we make the case that probabilistic topic models are a suitable framework for generating content-based representations of social relationships, because they are machine-interpretable and offer a path for principled domain adaptation. The former property is demonstrated by the successful use of topic models for building an intermediate representation of online communication data in two experiments on social influence (chapter 5 and following), the latter is demonstrated in a case study (section 4.3), where a topic model is adapted to the task of finding temporal patterns in online communication.

Given this body of evidence, we feel confident in concluding that the content-based characterization of social relationships via probabilisitic topic models is a viable step towards closing the gap between theory and practice in social computing. Once a reliable characterization of social relationships is possible, it will be necessary to determine the expectations regarding privacy users have towards a software system that views social relationships from a perspective that differs from their own. The kind of insight provided by such a system might not always be welcome.

Datasets

The two parts of this thesis are linked by the need for a corpus of observations of online social interactions, on which experiments can be carried out. Ideally, this data should come from diverse sources, that is, from multiple social platforms that differ in target demographics, offered mechanisms of communication and information dissemination, user interface, etc., so that peculiarities of a single platform can be ruled out as a confounding factor. We apply the term "social platforms" to a subset of systems for computer-mediated communication, which provide a common core of functionality: communication among members along the edges of an explicit or implicit social graph, addressive and non-addressive modes of communication, and a mechanism for information sharing, i.e., the unmodified retransmission of received pieces of information. Based on these common features, data obtained from different systems can be meaningfully compared. Our definition of a social platform is deliberately broad, and is satisfied by a number of online systems, ranging from traditional (e-mail) to modern (social networking systems like Facebook or Twitter).

Our explorative comparison of four datasets in chapter 3 (two corporate e-mail datasets plus data acquired from Twitter and Facebook) paints a picture of similarities and differences. As expected, there are strong cross-platform differences in some aspects of the data, which can be traced back to how these platforms have been conceptualized and are perceived by their users. Each platform has its own social conventions. Some are imposed by the platform operators (e.g., Twitter encourages the use of pseudonyms, while Facebook strictly enforces the use of ones legal name), others emerge from the user base and are reflected by people's usage patterns. The explicit social network graphs of Twitter and Facebook differ in assortativity: popular Facebook users tend to be friends with other popular users, while on Twitter, content consumers of low connectivity follow highly-connected content producers. In the implicit network graphs, there are pronounced differences in density and normalized degree variance, ranging from Facebook (lowest density and variance) to e-mail (highest). Posts on SNS like Facebook and Twitter tend to draw on a larger vocabulary, while e-mail messages are longer. Clearly, each platform fills a different niche by addressing different communicative needs.

Similarities between the datasets also exist, some of them in unexpected places. Ignoring time-zone differences, the message volume in all datasets exhibits similar diurnal and weekly rhythms. At the level of conversational behavior (replying to and sharing of messages), we also find temporal rhythms that are consistent across all platforms: depending on how active a conversation currently is, typical distances between subsequent reactions are 5–15 minutes, one hour, and 1–3 days. The implicit network graphs, formed by communi-

cation, are highly connected. In all but one case (Facebook), the largest weakly connected component of these graphs contains more than 97% of nodes. A common property of all network graphs, explicit and implicit, is that their structure alone is not sufficient for finding meaningful communities. Finally, an ethical issue is common to all of these datasets: Does the public accessibility of the data imply that its use for research is acceptable without individual informed consent? In the context of this thesis, we can answer in the affirmative, but in the general case, no clear-cut answers are possible (compare section 3.7 and Salganik, 2017).

Ideally, the explorative analysis of the four datasets should foster a thorough understanding of the collected data, identify unwanted artifacts of the data acquisition process as well as other sources of bias, and ultimately lead to concrete, operational strategies for the mitigation of bias. In many cases, however, a full, quantitative analysis of a potential source of bias is a research project of its own, e.g., the detection of spam and other kinds of automated messages, so that time and resource constraints limit us to a basic, mostly descriptive analysis. With respect to Ruths and Pfeffer (2014), who call for researchers to acknowledge and correct for biases of the population of the examined social platform, we must contend that systematically quantifying and accounting for all possible sources of bias is out of reach of the individual researcher and even a small team, and must rather be understood as a field-wide collaborative effort.

Social Influence

As evidenced by several survey papers (Sun and Tang, 2011; Li et al., 2015), as well as our own review of relevant literature (section 5.3), the study of online social influence is a highly fragmented field, to the point that it is difficult to even establish a taxonomy of the conceptually and methodologically diverse approaches. Finding common ground for a meaningful comparison is only possible by restricting the scope to a specific setting, such as the identification of influential users of Twitter (Riquelme and González-Cantergiani, 2016). The field does not only lack standard datasets on the basis of which new methods could be evaluated and compared, what is missing is a common understanding of even more basic issues: a shared definition of social influence and an inventory of tasks (e.g., recovery of a latent influence network, detection of influential users, etc.). Studies are frequently limited to a specific kind of social influence that becomes observable through information retransmission, or rely on proxy measures to identify users with a high potential for exerting influence. We reject both approaches as insufficient, and instead propose an approach that is motivated by a simple cognitive feedback loop and uses topic modeling for the representation of observation data.

If social influence is defined as the relationship between exposure to the behavior of others, internal cognitive processes, and one's own behavior, it can take almost arbitrary shape. Furthermore, many potential sources of influence are principally unobservable. How is it possible to learn anything about social influence in this setting? We start from the assumption that the strength of influence depends on the similarity of the social contexts of influencer and influencee. This relationship, which we call locality of influence, makes the detection of influence tractable. In the course of our experiments, we encounter multiple

cases in which the assumption of locality does not hold. A second important mechanism for learning about social influence appears to be indirect exposure: under certain conditions, the behavior of strong, unobserved sources of influence is reflected by the aggregate behavior of a social group.

Our first attempt at detecting social influence in online communication data is situated at the micro level of analysis. We start from a simple and intuitively appealing idea: after quantization of the time line into equally sized intervals, the behavior of each actor in during an interval can be expressed as a topic distribution. If the change in behavior of an actor a is similar to the past behavior of actor b, then b is a possible source of influence on a. However, we argue that this method of quantifying social influence is an instance of Granger-causal reasoning, and is therefore subject to all limitations that are associated with Granger's notion of causality, including, most importantly, its sensitivity to external confounders. A more immediate, practical problem is posed by the lack of annotated reference data, which makes an objective evaluation extremely difficult. Heuristic measures of the overall quality of the detected influence networks that do not require manual annotation turn out to have little discriminatory power. Manual annotation requires inspection of all messages of the potential influencer and influencee, which is highly time-consuming. Since it is impossible to objectively and exhaustively specify how the influence of one series of messages may manifest itself in another sequence, the annotation process is largely guided by the subjective impression of the annotator. Even worse, social influence effects of a magnitude that is sufficiently high to be visible to a human observer appear to be extremely rare, even among relationships that are ranked highly by the different algorithms. Although Granger-causal reasoning can be interpreted as a principally unreliable way of identifying causal relationships or as a diagnostic to rule out a specific type of non-causality, in our experimental setting we are unable to conclusively demonstrate its practical utility.

Instead of placing emphasis on discovering strictly causal relationships, one may ask how useful different parts of the social network are for predicting the future behavior of an actor. It is expected that the behavior of an actor's social environment reflects stronger sources of influence, including those that are external to the observed part of the network. The proposed model ("SCIM") expresses a prediction about future behavior of an actor as a linear combination of the past behavior of the actor and his or her social environment. For all actors in a given dataset, it learns a single set of coefficients that minimizes the average JSD between the predictions and the actual behavior. This model addresses the two main weaknesses of the previous approach: First, the use of historical data for parameter estimation and evaluation eschews the need for manual annotation. Second, its meso-level approach of going beyond dyadic relationships and exploiting information from a broader social environment addresses the problem of external confounders by making their effects an explicit part of the model. On all four datasets, the SCIM outperforms three baseline predictors, which gives us confidence that the learned coefficients actually describe the process of social influence. Encouragingly, we observe the same basic patterns on all datasets: Actors exhibit a strong tendency not to deviate from past behavior. When they do, their future behavior is much more strongly associated with sources of influence that we can only indirectly observe via the behavior of the social environment and the platform as a whole, than with direct, observable exposure. This suggests the possibility of using the influence

neighborhoods learned by a meso-level model such as the SCIM for eliminating potential confounders in micro-level inference (Guo et al., 2008).

The error analysis of the SCIM clearly shows that the final outcome of a social media study is affected by decisions at any earlier stage, including the underlying theoretical model of the phenomenon under investigation and the design and implementation of the data processing pipeline. A flaw in early experiment planning with a particularly strong effect is that the SNS, Twitter and Facebook, were crawled along the edges of the explicit relationship graph. This is the cause of the temporal sparsity issues observed in the evaluation of the SCIM and at other points in this thesis: many pairs of explicitly connected users do not interact within the chosen observation period. Data could have been obtained more efficiently by crawling the network of observed interactions within a specific time frame instead of crawling the explicit social network. More generally, the crawling process should be considered as a part of the experiment design and should be aligned with the requirements and goals of the experiment.

Future Work

Each of the two main parts of this thesis leaves some open questions for future research. In the first part, we propose a purely content-based characterization of relationships. What would a characterization at a higher level of abstraction look like? There is potential for a richer understanding of social relationships at the intersection of affiliation networks, Mika's tri-partite tagging networks (Mika, 2005), and the social object theory of Knorr Cetina (1997). Shared social objects describe a relationship at a level of abstraction that is situated above purely content-based features like keywords or topics, and below theoretically motivated classes of relationships. Research on human affiliation networks has produced a number of salient candidates for social objects, and social tagging could be the basis of a process of negotiating a common language to describe these objects.

The study of social influence at the micro level is fraught with problems, but moving the analysis to the meso level turns out to be a viable approach, so influence at the macro level is not considered in this work at all. In the process of public opinion formation, the role of entities with a large audience, the often-mentioned celebrities and news media, deserves particular attention. The SCIM experiments identify some patterns resembling the two-step flow paradigm as interpreted by Watts and Dodds (2007), which posits the need for "intermediaries between the mass media and the majority of society". A combination of strong external and indirect sources of influence hints at the importance of intermediaries, who enable strong global influencers to exert their influence locally. Other macro-level hypotheses also warrant attention. Tufekci (2014) attributes influence mainly to "events that affect a society or a group in a wholesale fashion either through shared experience or through broadcast media". Under this hypothesis, a topically focused community will likewise "attract" focused external influence. Further macro-level influence effects might be found by looking for leader-follower relationships in the behavior of broad demographic groups identified by gender or age.

Neither our micro-level nor our meso-level study considers complex contagion phenomena such as the effect of repeated exposure to a topic by a single actor, a coordinated group,

or multiple independent actors. While our own studies do not provide any direct evidence for the role of complex contagion, the results of an interventional study in a controlled setting by Pennycook et al. (2018) make complex contagion appear to be a viable hypothesis.

A Derivation of a Gibbs sampler for the Message Sequence Topic Model

The Message Sequence Topic Model (MSTM) is defined in section 4.3.2. Given the joint probability distribution of the variables of the MSTM, shown in equation A.1, we wish to derive the full conditional distribution of a Gibbs sampler that collapses out variables θ, φ, π (equation A.2). Notation and proof structure follow the derivation of the LDA full conditional by Carpenter (2010) as close as possible. The letter q is used as a shorthand for the sender-recipient pair (i, j) that identifies a sequence.

Due to the independence assumptions of the model, the joint distribution can be expressed as follows:

$$
\begin{aligned}
& p(\theta, \varphi, \pi, s, z, y | \alpha, \beta, \gamma) \\
&= p(\theta|\alpha) \cdot p(\varphi|\beta) \cdot p(\pi|\gamma) \cdot p(s|\pi) \cdot p(z|\theta, s) \cdot p(y|\varphi, z) \\
&= \prod_{i=1}^{A} \prod_{j=1}^{A} \prod_{u=1}^{S} p(\theta_{q,u}|\alpha) \cdot \prod_{k=1}^{K} p(\varphi_k|\beta) \cdot \\
&\quad \prod_{i=1}^{A} \prod_{j=1}^{A} \prod_{u=1}^{S} p(\pi_{q,u}|\gamma_u) \cdot \prod_{i=1}^{A} \prod_{j=1}^{A} \prod_{m=1}^{M_q} p(s_{q,m}|\pi_{q,s_{q,m-1}}) \cdot \\
&\quad \prod_{i=1}^{A} \prod_{j=1}^{A} \prod_{m=1}^{M_q} \prod_{n=1}^{N_{q,m}} p(z_{q,m,n}|\theta_{q,s_{q,m}}) \cdot \prod_{i=1}^{A} \prod_{j=1}^{A} \prod_{m=1}^{M_q} \prod_{n=1}^{N_{q,m}} p(y_{q,m,n}|\varphi_{z_{q,m,n}}) \\
&= \prod_{k=1}^{K} \mathrm{Dir}(\varphi_k|\beta) \cdot \prod_{i=1}^{A} \prod_{j=1}^{A} \left[\prod_{u=1}^{S} \left(\mathrm{Dir}(\theta_{q,u}|\alpha) \cdot \mathrm{Dir}(\pi_{q,u}|\gamma_u) \right) \cdot \right. \\
&\quad \left. \prod_{m=1}^{M_q} \left(\mathrm{Disc}(s_{q,m}|\pi_{q,s_{q,m-1}}) \cdot \prod_{n=1}^{N_{q,m}} \left(\mathrm{Disc}(z_{q,m,n}|\theta_{q,s_{q,m}}) \cdot \mathrm{Disc}(y_{q,m,n}|\varphi_{z_{q,m,n}}) \right) \right) \right]
\end{aligned}
\tag{A.1}
$$

For the purpose of Gibbs sampling, the probability density of the full conditional distribution only has to be known up to proportionality. The full conditional of the collapsed Gibbs sampler at sampling position (q, m, n) is proportional to the joint distribution with θ, φ, π marginalized out:

$$
\begin{aligned}
& p(s_{q,m}, z_{q,m,n} | s_{-(q,m)}, z_{-(q,m,n)}, y, \alpha, \beta, \gamma) \\
&= \frac{p(s_{q,m}, s_{-(q,m)}, z_{q,m,n}, z_{-(q,m,n)}, y | \alpha, \beta, \gamma)}{p(s_{-(q,m)}, z_{-(q,m,n)}, y | \alpha, \beta, \gamma)} \\
&\propto p(s_{q,m}, s_{-(q,m)}, z_{q,m,n}, z_{-(q,m,n)}, y | \alpha, \beta, \gamma) = p(s, z, y | \alpha, \beta, \gamma)
\end{aligned}
\tag{A.2}
$$

$$= \iiint p(\theta, \varphi, \pi, s, z, y|\alpha, \beta, \gamma)\, d\theta d\varphi d\pi$$

Using the independence assumptions as in equation A.1 and the linearity of the integral, one obtains:

$$= \iiint p(\theta|\alpha) \cdot p(\varphi|\beta) \cdot p(\pi|\gamma) \cdot p(s|\pi) \cdot p(z|\theta, s) \cdot p(y|\varphi, z)\, d\theta d\varphi d\pi$$

$$= \int p(z|\theta, s)\, p(\theta|\alpha)\, d\theta \cdot \int p(y|\varphi, z)\, p(\varphi|\beta)\, d\varphi \cdot \int p(s|\pi)\, p(\pi|\gamma)\, d\pi$$

with $q' := (i', j')$:

$$= \prod_{i'=1}^{A} \prod_{j'=1}^{A} \prod_{u=1}^{S} \int p(\theta_{q',u}|\alpha) \cdot \prod_{m'=1}^{M_{q'}} \prod_{n'=1}^{N_{q',m'}} p(z_{q',m',n'}|\theta_{q',s_{q',m'}})\, d\theta_{q',u} \cdot \tag{A.3}$$

$$\prod_{k=1}^{K} \int p(\varphi_k|\beta) \cdot \prod_{i'=1}^{A} \prod_{j'=1}^{A} \prod_{m'=1}^{M_{q'}} \prod_{n'=1}^{N_{q',m'}} p(y_{q',m',n'}|\varphi_{z_{q',m',n'}})\, d\varphi_k \cdot \tag{A.4}$$

$$\prod_{i'=1}^{A} \prod_{j'=1}^{A} \prod_{u=1}^{S} \int p(\pi_{q',u}|\gamma_u) \cdot \prod_{m'=1}^{M_{q'}} p(s_{q',m'}|\pi_{q',s_{q',m'-1}})\, d\pi_{q',u} \tag{A.5}$$

The three above terms (A.3, A.4, A.5) are evaluated separately, starting with the first (A.3). Let $c_{k,q,u,w}$ be the number of times word w is assigned to topic k in a message in state u that belongs to sequence q, and let an asterisk instead of any of the subscript indices denote a summation over all possible values of that index. A superscript $c^{-(q,m,n)}$ indicates that the word at position (q, m, n) is not counted.

$$\prod_{i'=1}^{A} \prod_{j'=1}^{A} \prod_{u=1}^{S} \int p(\theta_{q',u}|\alpha) \cdot \prod_{m'=1}^{M_{q'}} \prod_{n'=1}^{N_{q',m'}} p(z_{q',m',n'}|\theta_{q',s_{q',m'}})\, d\theta_{q',u}$$

$$= \prod_{i'=1}^{A} \prod_{j'=1}^{A} \prod_{u=1}^{S} \int \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{q',u,k}^{\alpha_k - 1} \cdot \prod_{m'=1}^{M_{q'}} \prod_{n'=1}^{N_{q',m'}} \theta_{q',s_{q',m'},z_{q',m',n'}}\, d\theta_{q',u}$$

$$= \prod_{i'=1}^{A} \prod_{j'=1}^{A} \prod_{u=1}^{S} \int \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{q',u,k}^{\alpha_k - 1} \prod_{k=1}^{K} \theta_{q',u,k}^{c_{k,q',u,*}}\, d\theta_{q',u}$$

$$= \prod_{i'=1}^{A} \prod_{j'=1}^{A} \prod_{u=1}^{S} \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(c_{k,q',u,*} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,q',u,*} + \alpha_k)} \cdot$$

$$\int \frac{\Gamma(\sum_{k=1}^{K} c_{k,q',u,*} + \alpha_k)}{\prod_{k=1}^{K} \Gamma(c_{k,q',u,*} + \alpha_k)} \prod_{k=1}^{K} \theta_{q',u,k}^{\alpha_k + c_{k,q',u,*} - 1}\, d\theta_{q',u}$$

This integral over the PDF of the Dirichlet distribution evaluates to 1.

$$\propto \prod_{i'=1}^{A} \prod_{j'=1}^{A} \prod_{u=1}^{S} \frac{\prod_{k=1}^{K} \Gamma(c_{k,q',u,*} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,q',u,*} + \alpha_k)}$$

All factors that do not involve the variables s and z at the current sampling position (q, m, n) are successively dropped.

$$
\begin{aligned}
&\propto \prod_{u=1}^{S} \frac{\prod_{k=1}^{K} \Gamma(c_{k,q,u,*} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,q,u,*} + \alpha_k)} \\
&= \prod_{u \neq s_{q,m}}^{S} \frac{\prod_{k=1}^{K} \Gamma(c_{k,q,u,*} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,q,u,*} + \alpha_k)} \cdot \frac{\prod_{k \neq z_{q,m,n}}^{K} \Gamma(c_{k,q,s_{q,m},*} + \alpha_k) \cdot \Gamma(c_{z_{q,m,n},q,s_{q,m},*} + \alpha_{z_{q,m,n}})}{\Gamma(\sum_{k=1}^{K} c_{k,q,s_{q,m},*} + \alpha_k)} \\
&= \prod_{u \neq s_{q,m}}^{S} \frac{\prod_{k=1}^{K} \Gamma(c_{k,q,u,*}^{-(q,m,n)} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,q,u,*}^{-(q,m,n)} + \alpha_k)} \cdot \\
&\quad \frac{\prod_{k \neq z_{q,m,n}}^{K} \Gamma(c_{k,q,s_{q,m},*}^{-(q,m,n)} + \alpha_k) \cdot \Gamma(c_{z_{q,m,n},q,s_{q,m},*}^{-(q,m,n)} + \alpha_{z_{q,m,n}} + 1)}{\Gamma(\sum_{k=1}^{K} (c_{k,q,s_{q,m},*}^{-(q,m,n)} + \alpha_k) + 1)} \\
&\quad \text{with } \Gamma(x+1) = x \cdot \Gamma(x): \\
&= \prod_{u=1}^{S} \frac{\prod_{k=1}^{K} \Gamma(c_{k,q,u,*}^{-(q,m,n)} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,q,u,*}^{-(q,m,n)} + \alpha_k)} \cdot \frac{c_{z_{q,m,n},q,s_{q,m},*}^{-(q,m,n)} + \alpha_{z_{q,m,n}}}{\sum_{k=1}^{K} c_{k,q,s_{q,m},*}^{-(q,m,n)} + \alpha_k} \\
&\propto \frac{c_{z_{q,m,n},q,s_{q,m},*}^{-(q,m,n)} + \alpha_{z_{q,m,n}}}{c_{*,q,s_{q,m},*}^{-(q,m,n)} + \sum_{k=1}^{K} \alpha_k}
\end{aligned}
\tag{A.6}
$$

The second term (A.4), a product of K integrals over φ_k, appears in highly similar form in the derivation of the full conditional of standard LDA. Proceeding in analogy to steps 19 to 31 of Carpenter's derivation (2010), one obtains:

$$
\begin{aligned}
&\prod_{k=1}^{K} \int p(\varphi_k|\beta) \cdot \prod_{i'=1}^{A} \prod_{j'=1}^{A} \prod_{m'=1}^{M_{q'}} \prod_{n'=1}^{N_{q',m'}} p(y_{q',m',n'}|\varphi_{z_{q',m',n'}})\, d\varphi_k \\
&= \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{w=1}^{W} \beta_w)}{\prod_{w=1}^{W} \Gamma(\beta_w)} \prod_{w=1}^{W} \varphi_{k,w}^{\beta_w - 1} \cdot \prod_{i'=1}^{A} \prod_{j'=1}^{A} \prod_{m'=1}^{M_{q'}} \prod_{n'=1}^{N_{q',m'}} \varphi_{z_{q',m',n'},y_{q',m',n'}}\, d\varphi_k \\
&= \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{w=1}^{W} \beta_w)}{\prod_{w=1}^{W} \Gamma(\beta_w)} \prod_{w=1}^{W} \varphi_{k,w}^{\beta_w - 1} \prod_{w=1}^{W} \varphi_{k,w}^{c_{k,*,*,w}}\, d\varphi_k \\
&\propto \prod_{k=1}^{K} \frac{\prod_{w=1}^{W} \Gamma(c_{k,*,*,w} + \beta_w)}{\Gamma(\sum_{w=1}^{W} c_{k,*,*,w} + \beta_w)} \\
&= \prod_{k \neq z_{q,m,n}}^{K} \frac{\prod_{w=1}^{W} \Gamma(c_{k,*,*,w} + \beta_w)}{\Gamma(\sum_{w=1}^{W} c_{k,*,*,w} + \beta_w)} \cdot \\
&\quad \frac{\prod_{w \neq y_{q,m,n}}^{W} \Gamma(c_{z_{q,m,n},*,*,w} + \beta_w) \cdot \Gamma(c_{z_{q,m,n},*,*,y_{q,m,n}} + \beta_{y_{q,m,n}})}{\Gamma(\sum_{w=1}^{W} c_{z_{q,m,n},*,*,w} + \beta_w)}
\end{aligned}
$$

$$
\begin{aligned}
&= \prod_{k \neq z_{q,m,n}}^{K} \frac{\prod_{w=1}^{W} \Gamma(c_{k,*,*,w}^{-(q,m,n)} + \beta_w)}{\Gamma(\sum_{w=1}^{W} c_{k,*,*,w}^{-(q,m,n)} + \beta_w)} \cdot \\
&\quad \frac{\prod_{w \neq y_{q,m,n}}^{W} \Gamma(c_{z_{q,m,n},*,*,w}^{-(q,m,n)} + \beta_w) \cdot \Gamma(c_{z_{q,m,n},*,*,y_{q,m,n}}^{-(q,m,n)} + \beta_{y_{q,m,n}} + 1)}{\Gamma(\sum_{w=1}^{W} (c_{z_{q,m,n},*,*,w}^{-(q,m,n)} + \beta_w) + 1)} \\
&= \prod_{k=1}^{K} \frac{\prod_{w=1}^{W} \Gamma(c_{k,*,*,w}^{-(q,m,n)} + \beta_w)}{\Gamma(\sum_{w=1}^{W} c_{k,*,*,w}^{-(q,m,n)} + \beta_w)} \cdot \frac{c_{z_{q,m,n},*,*,y_{q,m,n}}^{-(q,m,n)} + \beta_{y_{q,m,n}}}{\sum_{w=1}^{W} c_{z_{q,m,n},*,*,w}^{-(q,m,n)} + \beta_w} \\
&\propto \frac{c_{z_{q,m,n},*,*,y_{q,m,n}}^{-(q,m,n)} + \beta_{y_{q,m,n}}}{c_{z_{q,m,n},*,*,*}^{-(q,m,n)} + \sum_{w=1}^{W} \beta_w}
\end{aligned}
\tag{A.7}
$$

The third term (A.5) appears in similar form in the derivation of the full conditional of the LDA-HMM composite model (Griffiths et al., 2005) and the BHMM (Goldwater and Griffiths, 2007; Johnson, 2007). Assume that $s_{q,0} = s_{q,M_q+1} = 1$. Let $g_{q,u,v}$ be the number of transitions from state u to state v in sequence q. A superscript $g^{-(q,m)}$ indicates that the transition from document $m-1$ to document m and from document m to document $m+1$ in sequence q is not counted.

$$
\begin{aligned}
&\prod_{i'=1}^{A} \prod_{j'=1}^{A} \prod_{u=1}^{S} \int p(\pi_{q',u} | \gamma_u) \cdot \prod_{m'=1}^{M_{q'}} p(s_{q',m'} | \pi_{q',s_{q',m'-1}}) \, d\pi_{q',u} \\
&= \prod_{i'=1}^{A} \prod_{j'=1}^{A} \prod_{u=1}^{S} \int \frac{\Gamma(\sum_{v=1}^{S} \gamma_v)}{\prod_{v=1}^{S} \Gamma(\gamma_v)} \prod_{v=1}^{S} \pi_{q',u,v}^{\gamma_v - 1} \cdot \prod_{m'=1}^{M_{q'}} \pi_{q',s_{q',m'-1},s_{q',m'}} \, d\pi_{q',u} \\
&= \prod_{i'=1}^{A} \prod_{j'=1}^{A} \prod_{u=1}^{S} \int \frac{\Gamma(\sum_{v=1}^{S} \gamma_v)}{\prod_{v=1}^{S} \Gamma(\gamma_v)} \prod_{v=1}^{S} \pi_{q',u,v}^{\gamma_v - 1} \prod_{v=1}^{S} \pi_{q',u,v}^{g_{q',u,v}} \, d\pi_{q',u} \\
&\propto \prod_{i'=1}^{A} \prod_{j'=1}^{A} \prod_{u=1}^{S} \frac{\prod_{v=1}^{S} \Gamma(g_{q',u,v} + \gamma_v)}{\Gamma(\sum_{v=1}^{S} g_{q',u,v} + \gamma_v)} \\
&\propto \prod_{u=1}^{S} \frac{\prod_{v=1}^{S} \Gamma(g_{q,u,v} + \gamma_v)}{\Gamma(\sum_{v=1}^{S} g_{q,u,v} + \gamma_v)}
\end{aligned}
$$

At this point, three cases have to be distinguished.

If $s_{q,m-1} \neq s_{q,m}$:

$$
\begin{aligned}
&= \prod_{\substack{u \neq s_{q,m-1} \\ \wedge u \neq s_{q,m}}}^{S} \frac{\prod_{v=1}^{S} \Gamma(g_{q,u,v} + \gamma_v)}{\Gamma(\sum_{v=1}^{S} g_{q,u,v} + \gamma_v)} \cdot \frac{\prod_{v \neq s_{q,m}}^{S} \Gamma(g_{q,s_{q,m-1},v} + \gamma_v) \cdot \Gamma(g_{q,s_{q,m-1},s_{q,m}} + \gamma_{s_{q,m}})}{\Gamma(\sum_{v=1}^{S} g_{q,s_{q,m-1},v} + \gamma_v)} \cdot \\
&\quad \frac{\prod_{v \neq s_{q,m+1}}^{S} \Gamma(g_{q,s_{q,m},v} + \gamma_v) \cdot \Gamma(g_{q,s_{q,m},s_{q,m+1}} + \gamma_{s_{q,m+1}})}{\Gamma(\sum_{v=1}^{S} g_{q,s_{q,m},v} + \gamma_v)}
\end{aligned}
$$

$$= \prod_{\substack{u \neq s_{q,m-1} \\ \wedge u \neq s_{q,m}}}^{S} \frac{\prod_{v=1}^{S} \Gamma(g_{q,u,v}^{-(q,m)} + \gamma_v)}{\Gamma(\sum_{v=1}^{S} g_{q,u,v}^{-(q,m)} + \gamma_v)} \cdot \frac{\prod_{v \neq s_{q,m}}^{S} \Gamma(g_{q,s_{q,m-1},v}^{-(q,m)} + \gamma_v) \cdot \Gamma(g_{q,s_{q,m-1},s_{q,m}}^{-(q,m)} + \gamma_{s_{q,m}} + 1)}{\Gamma(\sum_{v=1}^{S}(g_{q,s_{q,m-1},v}^{-(q,m)} + \gamma_v) + 1)} \cdot$$

$$\frac{\prod_{v \neq s_{q,m+1}}^{S} \Gamma(g_{q,s_{q,m},v}^{-(q,m)} + \gamma_v) \cdot \Gamma(g_{q,s_{q,m},s_{q,m+1}}^{-(q,m)} + \gamma_{s_{q,m+1}} + 1)}{\Gamma(\sum_{v=1}^{S}(g_{q,s_{q,m},v}^{-(q,m)} + \gamma_v) + 1)}$$

$$= \prod_{s=1}^{S} \frac{\prod_{v=1}^{S} \Gamma(g_{q,u,v}^{-(q,m)} + \gamma_v)}{\Gamma(\sum_{v=1}^{S} g_{q,u,v}^{-(q,m)} + \gamma_v)} \cdot \frac{(g_{q,s_{q,m-1},s_{q,m}}^{-(q,m)} + \gamma_{s_{q,m}}) \cdot (g_{q,s_{q,m},s_{q,m+1}}^{-(q,m)} + \gamma_{s_{q,m+1}})}{(\sum_{v=1}^{S} g_{q,s_{q,m-1},v}^{-(q,m)} + \gamma_v) \cdot (\sum_{v=1}^{S} g_{q,s_{q,m},v}^{-(q,m)} + \gamma_v)}$$

$$\propto \frac{(g_{q,s_{q,m-1},s_{q,m}}^{-(q,m)} + \gamma_{s_{q,m}}) \cdot (g_{q,s_{q,m},s_{q,m+1}}^{-(q,m)} + \gamma_{s_{q,m+1}})}{\sum_{v=1}^{S} g_{q,s_{q,m},v}^{-(q,m)} + \gamma_v}$$

If $s_{q,m-1} = s_{q,m} \neq s_{q,m+1}$:

$$= \prod_{u \neq s_{q,m-1}}^{S} \frac{\prod_{v=1}^{S} \Gamma(g_{q,u,v} + \gamma_v)}{\Gamma(\sum_{v=1}^{S} g_{q,u,v} + \gamma_v)} \cdot$$

$$\frac{\prod_{\substack{v \neq s_{q,m} \wedge \\ v \neq s_{q,m+1}}}^{S} \Gamma(g_{q,s_{q,m-1},v} + \gamma_v) \cdot \Gamma(g_{q,s_{q,m-1},s_{q,m}} + \gamma_{s_{q,m}}) \cdot \Gamma(g_{q,s_{q,m},s_{q,m+1}} + \gamma_{s_{q,m+1}})}{\Gamma(\sum_{v=1}^{S} g_{q,s_{q,m-1},v} + \gamma_v)}$$

$$= \prod_{u \neq s_{q,m-1}}^{S} \frac{\prod_{v=1}^{S} \Gamma(g_{q,u,v}^{-(q,m)} + \gamma_v)}{\Gamma(\sum_{v=1}^{S} g_{q,u,v}^{-(q,m)} + \gamma_v)} \cdot$$

$$\frac{\prod_{\substack{v \neq s_{q,m} \wedge \\ v \neq s_{q,m+1}}}^{S} \Gamma(g_{q,s_{q,m-1},v}^{-(q,m)} + \gamma_v) \cdot \Gamma(g_{q,s_{q,m-1},s_{q,m}}^{-(q,m)} + \gamma_{s_{q,m}} + 1) \cdot \Gamma(g_{q,s_{q,m},s_{q,m+1}}^{-(q,m)} + \gamma_{s_{q,m+1}} + 1)}{\Gamma(\sum_{v=1}^{S}(g_{q,s_{q,m-1},v}^{-(q,m)} + \gamma_v) + 2)}$$

$$= \prod_{s=1}^{S} \frac{\prod_{v=1}^{S} \Gamma(g_{q,u,v}^{-(q,m)} + \gamma_v)}{\Gamma(\sum_{v=1}^{S} g_{q,u,v}^{-(q,m)} + \gamma_v)} \cdot \frac{(g_{q,s_{q,m-1},s_{q,m}}^{-(q,m)} + \gamma_{s_{q,m}}) \cdot (g_{q,s_{q,m},s_{q,m+1}}^{-(q,m)} + \gamma_{s_{q,m+1}})}{(\sum_{v=1}^{S} g_{q,s_{q,m-1},v}^{-(q,m)} + \gamma_v) \cdot (\sum_{v=1}^{S}(g_{q,s_{q,m},v}^{-(q,m)} + \gamma_v) + 1)}$$

$$\propto \frac{(g_{q,s_{q,m-1},s_{q,m}}^{-(q,m)} + \gamma_{s_{q,m}}) \cdot (g_{q,s_{q,m},s_{q,m+1}}^{-(q,m)} + \gamma_{s_{q,m+1}})}{\sum_{v=1}^{S}(g_{q,s_{q,m},v}^{-(q,m)} + \gamma_v) + 1}$$

If $s_{q,m-1} = s_{q,m} = s_{q,m+1}$:

$$= \prod_{u \neq s_{q,m-1}}^{S} \frac{\prod_{v=1}^{S} \Gamma(g_{q,u,v} + \gamma_v)}{\Gamma(\sum_{v=1}^{S} g_{q,u,v} + \gamma_v)} \cdot \frac{\prod_{v \neq s_{q,m}}^{S} \Gamma(g_{q,s_{q,m-1},v} + \gamma_v) \cdot \Gamma(g_{q,s_{q,m-1},s_{q,m}} + \gamma_{s_{q,m}})}{\Gamma(\sum_{v=1}^{S} g_{q,s_{q,m-1},v} + \gamma_v)}$$

$$= \prod_{u \neq s_{q,m-1}}^{S} \frac{\prod_{v=1}^{S} \Gamma(g_{q,u,v}^{-(q,m)} + \gamma_v)}{\Gamma(\sum_{v=1}^{S} g_{q,u,v}^{-(q,m)} + \gamma_v)} \cdot \frac{\prod_{v \neq s_{q,m}}^{S} \Gamma(g_{q,s_{q,m-1},v}^{-(q,m)} + \gamma_v) \cdot \Gamma(g_{q,s_{q,m-1},s_{q,m}}^{-(q,m)} + \gamma_{s_{q,m}} + 2)}{\Gamma(\sum_{v=1}^{S}(g_{q,s_{q,m-1},v}^{-(q,m)} + \gamma_v) + 2)}$$

$$= \prod_{u=1}^{S} \frac{\prod_{v=1}^{S} \Gamma(g_{q,u,v}^{-(q,m)} + \gamma_v)}{\Gamma(\sum_{v=1}^{S} g_{q,u,v}^{-(q,m)} + \gamma_v)} \cdot \frac{(g_{q,s_{q,m-1},s_{q,m}}^{-(q,m)} + \gamma_{s_{q,m}}) \cdot (g_{q,s_{q,m},s_{q,m+1}}^{-(q,m)} + \gamma_{s_{q,m+1}} + 1)}{(\sum_{v=1}^{S} g_{q,s_{q,m-1},v}^{-(q,m)} + \gamma_v) \cdot (\sum_{v=1}^{S} (g_{q,s_{q,m},v}^{-(q,m)} + \gamma_v) + 1)}$$

$$\propto \frac{(g_{q,s_{q,m-1},s_{q,m}}^{-(q,m)} + \gamma_{s_{q,m}}) \cdot (g_{q,s_{q,m},s_{q,m+1}}^{-(q,m)} + \gamma_{s_{q,m+1}} + 1)}{\sum_{v=1}^{S} (g_{q,s_{q,m},v}^{-(q,m)} + \gamma_v) + 1}$$

The three cases can be unified by introducing indicator functions in two places:

$$\propto \frac{(g_{q,s_{q,m-1},s_{q,m}}^{-(q,m)} + \gamma_{s_{q,m}}) \cdot (g_{q,s_{q,m},s_{q,m+1}}^{-(q,m)} + I(s_{q,m-1} = s_{q,m} = s_{q,m+1}) + \gamma_{s_{q,m+1}})}{g_{q,s_{q,m},*}^{-(q,m)} + I(s_{q,m-1} = s_{q,m}) + \sum_{u=1}^{S} \gamma_u} \tag{A.8}$$

The full conditional of the MSTM (equation 4.31 in section 4.3.2) is defined, up to proportionality, by the product of the three terms A.6, A.7, and A.8.

B Sampling a Dirichlet Distribution Subject to an ℓ_1 Equality Constraint

The d-dimensional Dirichlet distribution $\mathrm{Dir}(\alpha)$, $\alpha \in \mathbb{R}^d$, is supported on the unit simplex Δ^{d-1}. For a given $c \in \Delta^{d-1}$ (i.e., $\forall_{i=1}^{d} : c_i \geq 0$ and $\sum_{i=1}^{d} c_i = 1$) and $r \in \mathbb{R}$ with $0 \leq r \leq 2 - 2 \cdot \min_{i=1}^{d} c_i$, a sample $x \in \mathbb{R}^d$ is to be drawn from $\mathrm{Dir}(\alpha)$, subject to the constraint $\|x - c\|_1 = r$ (or equivalently $\sum_{i=1}^{d} |x_i - c_i| = r$). As illustrated in figure B.1, this can be interpreted as constraining the support to the intersection of the simplex and the surface of an ℓ_1 $(d-1)$-sphere of radius r, centered at c. The upper bound on r ensures that the intersection is not empty.

Given that $\sum_{i=1}^{d} x_i = 1 \Leftrightarrow x_n = 1 - \sum_{i \neq n}^{d} x_i$ for any $1 \leq n \leq d$, one can reformulate the original constraint as the disjunction of two constraints (equations B.1 and B.2) on the subset of components $\{x_i | 1 \leq i \leq d \wedge i \neq n\}$:

$$\sum_{i=1}^{d} |x_i - c_i| = r$$

$x_n \geq c_n$:

$$\Rightarrow x_n = r - \sum_{i \neq n}^{d} |x_i - c_i| + c_n$$

$$\Leftrightarrow 1 - \sum_{i \neq n}^{d} x_i = r - \sum_{i \neq n}^{d} |x_i - c_i| + 1 - \sum_{i \neq n}^{d} c_i$$

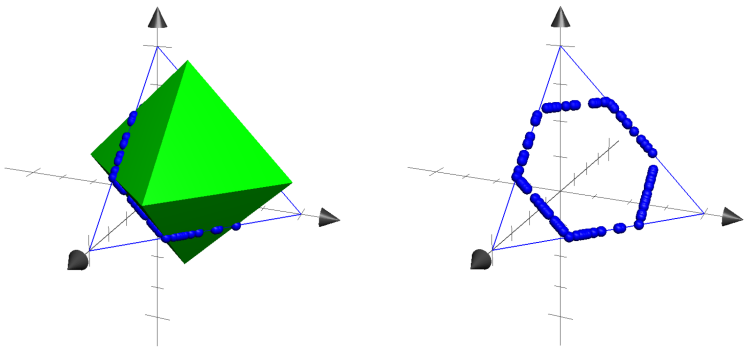


Figure B.1: Samples from a uniform Dirichlet distribution constrained to the intersection of the surface of an ℓ_1 sphere and the unit simplex in $\mathbb{R}^3$

$$\Leftrightarrow \sum_{i\neq n}^{d}(|x_i - c_i| - (x_i - c_i)) = r$$

$$\Leftrightarrow \sum_{i\neq n}^{d}\max\{0, c_i - x_i\} = \frac{r}{2} \tag{B.1}$$

$x_n < c_n$:

$$\Rightarrow x_n = -r + \sum_{i\neq n}^{d}|x_i - c_i| + c_n$$

(in analogy to the previous case)

$$\Leftrightarrow \sum_{i\neq n}^{d}\max\{0, x_i - c_i\} = \frac{r}{2} \tag{B.2}$$

The Gibbs sampling scheme of Ng et al. (2011), while designed for truncated Dirichlet distributions, can be adapted to the more general constraint discussed here. Gibbs sampling obtains samples from a probability distribution via a random walk on a Markov chain. Due to its Markovian dependency on the previously sampled value, the Gibbs sampler requires an initial value within the constraint region. An algorithm for constructing a suitable value can be derived from either constraint B.1 or B.2; here we use B.1, from which we first derive a lower bound on x_i that only depends on x_j with $j < i$:

$$\max\{0, c_i - x_i\} = \frac{r}{2} - \sum_{j\neq i\wedge j\neq n}^{d}\max\{0, c_j - x_j\}$$

$$\Rightarrow x_i = c_i - \frac{r}{2} + \sum_{j\neq i\wedge j\neq n}^{d}\max\{0, c_j - x_j\} \tag{B.3}$$

$$\Rightarrow x_i \geq c_i - \frac{r}{2} + \sum_{j<i\wedge j\neq n}^{d}\max\{0, c_j - x_j\}$$

with $0 \leq x_i < c_i$:

$$\Rightarrow x_i \geq \max\left\{0, c_i - \frac{r}{2} + \sum_{j<i\wedge j\neq n}^{d}(c_j - x_j)\right\} \tag{B.4}$$

An appropriate initial value x can then be constructed by sequentially setting the value of each component x_i, $i \neq n$ to the lower bound defined by equation B.4, and finally setting $x_n = 1 - \sum_{i\neq n}^{d} x_i$. This effectively assigns to x_i the smallest possible value so that $\sum_{j\leq i\wedge j\neq n}^{d}\max\{0, c_j - x_j\}$ is maximal, but remains smaller than or equal to $\frac{r}{2}$. Once $\frac{r}{2}$ is reached, subsequent x_i are set to c_i and no longer contribute to the sum. When choosing $n = \operatorname{argmin}_{i=1}^{d} c_i$, $\sum_{i\neq n}^{d} c_i \geq 1 - \min_{i=1}^{d} c_i$, which is the maximum allowed value of $\frac{r}{2}$ for the given c, this will always produce a value that satisfies the constraint.

Due to the nature of the constraints, the usual approach of sampling each component from its univariate full conditional distribution does not work: As indicated by equation B.3, each

component x_i, $i \neq n$ can take on only two different values (one induced by constraint B.1, the other by B.2) when conditioned on the remaining components, so the resulting sampler would only be able to traverse a finite subset of the support. Instead, we sample from the joint distribution of a pair of components x_i, x_m, conditioned on all other components, which constitutes a blocked Gibbs sampling scheme. This joint distribution can be decomposed into two conditional distributions: $p(x_i, x_m | x_{j-i,m,n}) = p(x_i | x_{j-i,m,n}) \cdot p(x_m | x_{j-m,n})$, where $j - i$ denotes the sequence of indices excluding i, so that $x_{j-i} := x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_d$.

Starting from constraints B.1 and B.2, one can derive, for each pair of x_i and x_m, new constraints that do not depend on x_i, x_m, x_n and x_m, x_n, respectively:

$$
\begin{aligned}
&\sum_{j-i,m,n} \max\{0, x_j - c_j\} < \frac{r}{2} \;\wedge\; \ldots \\
&\qquad \sum_{j-i,m,n} \max\{0, c_j - x_j\} = \frac{r}{2} : \\
&\qquad\quad \Rightarrow\; c_i < x_i < c_i + \frac{r}{2} - \sum_{j-i,m,n} \max\{0, x_j - c_j\} \;\wedge \qquad \text{(B.5)} \\
&\qquad\qquad\quad c_m < x_m < c_m + \frac{r}{2} - \sum_{j-m,n} \max\{0, x_j - c_j\} \\
&\frac{r}{2} - c_n \leq \sum_{j-i,m,n} \max\{0, c_j - x_j\} \leq \frac{r}{2} : \\
&\qquad\quad \Rightarrow\; c_i < x_i < c_i + \frac{r}{2} - \sum_{j-i,m,n} \max\{0, x_j - c_j\} \;\wedge \qquad \text{(B.6)} \\
&\qquad\qquad\quad x_m = c_m + \frac{r}{2} - \sum_{j-m,n} \max\{0, x_j - c_j\} \\
&\frac{r}{2} - c_m \leq \sum_{j-i,m,n} \max\{0, c_j - x_j\} \leq \frac{r}{2} : \\
&\qquad\quad \Rightarrow\; c_i < x_i < c_i + \frac{r}{2} - \sum_{j-i,m,n} \max\{0, x_j - c_j\} \;\wedge \qquad \text{(B.7)} \\
&\qquad\qquad\quad x_m = c_m - \frac{r}{2} + \sum_{j-m,n} \max\{0, c_j - x_j\} \\
&\frac{r}{2} - c_m - c_n \leq \sum_{j-i,m,n} \max\{0, c_j - x_j\} \leq \frac{r}{2} : \\
&\qquad\quad \Rightarrow\; x_i = c_i + \frac{r}{2} - \sum_{j-i,m,n} \max\{0, x_j - c_j\} \;\wedge \qquad \text{(B.8)} \\
&\qquad\qquad\quad x_m \geq \max\left\{0, c_m - \frac{r}{2} + \sum_{j-m,n} \max\{0, c_j - x_j\}\right\} \;\wedge \\
&\qquad\qquad\quad x_m \leq \min\left\{c_n + c_m - \frac{r}{2} + \sum_{j-m,n} \max\{0, c_j - x_j\}, c_m\right\}
\end{aligned}
$$

$$\frac{r}{2} - c_i \leq \sum_{j-i,m,n} \max\{0, c_j - x_j\} \leq \frac{r}{2} :$$

$$\Rightarrow\ x_i = c_i - \frac{r}{2} + \sum_{j-i,m,n} \max\{0, c_j - x_j\} \wedge \tag{B.9}$$

$$c_m < x_m < c_m + \frac{r}{2} - \sum_{j-m,n} \max\{0, x_j - c_j\}$$

$$\frac{r}{2} - c_i - c_n \leq \sum_{j-i,m,n} \max\{0, c_j - x_j\} \leq \frac{r}{2} :$$

$$\Rightarrow\ x_i \geq \max\left\{0, c_i - \frac{r}{2} + \sum_{j-i,m,n} \max\{0, c_j - x_j\}\right\} \wedge \tag{B.10}$$

$$x_i \leq \min\left\{c_n + c_i - \frac{r}{2} + \sum_{j-i,m,n} \max\{0, c_j - x_j\}, c_i\right\} \wedge$$

$$x_m = c_m + \frac{r}{2} - \sum_{j-m,n} \max\{0, x_j - c_j\}$$

$$\frac{r}{2} - c_i - c_m \leq \sum_{j-i,m,n} \max\{0, c_j - x_j\} \leq \frac{r}{2} :$$

$$\Rightarrow\ x_i \geq \max\left\{0, c_i - \frac{r}{2} + \sum_{j-i,m,n} \max\{0, c_j - x_j\}\right\} \wedge \tag{B.11}$$

$$x_i \leq \min\left\{c_m + c_i - \frac{r}{2} + \sum_{j-i,m,n} \max\{0, c_j - x_j\}, c_i\right\} \wedge$$

$$x_m = c_m - \frac{r}{2} + \sum_{j-m,n} \max\{0, c_j - x_j\}$$

$$\sum_{j-i,m,n} \max\{0, x_j - c_j\} = \frac{r}{2} \ \wedge\ \ldots$$

$$\frac{r}{2} - c_i - c_m - c_n \leq \sum_{j-i,m,n} \max\{0, c_j - x_j\} \leq \frac{r}{2} :$$

$$\Rightarrow\ x_i \geq \max\left\{0, c_i - \frac{r}{2} + \sum_{j-i,m,n} \max\{0, c_j - x_j\}\right\} \wedge \tag{B.12}$$

$$x_i \leq \min\left\{c_n + c_m + c_i - \frac{r}{2} + \sum_{j-i,m,n} \max\{0, c_j - x_j\}, c_i\right\} \wedge$$

$$x_m \geq \max\left\{0, c_m - \frac{r}{2} + \sum_{j-m,n} \max\{0, c_j - x_j\}\right\} \wedge$$

$$x_m \leq \min\left\{c_n + c_m - \frac{r}{2} + \sum_{j-m,n} \max\{0, c_j - x_j\}, c_m\right\}$$

Of the eight cases (B.5 to B.12), the first seven are not mutually exclusive. Via the conditions on $\sum_{j-i,m,n} \max\{0, x_j - c_j\}$ and $\sum_{j-i,m,n} \max\{0, c_j - x_j\}$ one can decide which cases are applicable for the current value of x. With the exception of B.5 and B.12, each case constrains one variable to a single value and places upper and lower bounds on the other. The

remaining two cases specify upper and lower bounds for both variables. Within a Gibbs sampler, these two cases can be handled analogously to the first group by keeping the current value of one variable constant and only sampling the other. Thus, the active constraints can be interpreted as a set A of line segments $A_s \subset \mathbb{R}^2$, which are known to be mutually disjoint.

Given an indicator function $I_{x \in A_s}(x)$ for each segment $A_s \in A$, the joint distribution $p(x_i, x_m | x_{j-i,m,n})$ can be expressed as a mixture distribution over the set of segments via its unconstrained equivalent $q(x_i, x_m | x_{j-i,m,n})$:

$$\begin{aligned} p(x_i, x_m | x_{j-i,m,n}) &= \sum_{s=1}^{|A|} p(s | x_{j-i,m,n}) \cdot p(x_i, x_m | x_{j-i,m,n}, s) \quad \text{with} \\ p(s | x_{j-i,m,n}) &\propto \iint_{x_i, x_m \in A_s} q(x_i, x_m | x_{j-i,m,n}) \, dx_i \, dx_m \quad \text{and} \\ p(x_i, x_m | x_{j-i,m,n}, s) &\propto q(x_i, x_m | x_{j-i,m,n}) \cdot I_{x \in A_s}(x) \end{aligned}$$

The easiest way of sampling from this mixture is by first sampling the segment s from its categorical distribution. Then, the original Dirichlet distribution is sampled from, subject only to the single constraint represented by the chosen segment (Devroye, 1986, ch. II.4.3).

For sampling x_i and x_m from $p(x_i, x_m | x_{j-i,m,n}, s)$, that is, from a Dirichlet distribution constrained to segment s, we have to distinguish three cases: First, x_i is constrained to an interval $l(s) \leq x_i \leq u(s)$ and the constraints on x_m do not depend on x_i (cases B.5, B.7, B.10, and B.12), so that with $x_m = C$:

$$\begin{aligned} p(x_i | x_{j-i,m,n}, s) &\propto x_i^{\alpha_i - 1} \cdot x_m^{\alpha_m - 1} \cdot x_n^{\alpha_n - 1} \cdot I_{l(s) < x < u(s)}(x_i) \\ &= x_i^{\alpha_i - 1} \cdot C^{\alpha_m - 1} \cdot \left(1 - C - x_i - \sum_{j-i,m,n} x_j\right)^{\alpha_n - 1} \cdot I_{l(s) < x < u(s)}(x_i) \\ &\propto x_i^{\alpha_i - 1} \cdot \left(1 - C - x_i - \sum_{j-i,m,n} x_j\right)^{\alpha_n - 1} \cdot I_{l(s) < x < u(s)}(x_i) \\ &\propto \left(\frac{x_i}{1 - C - \sum_{j-i,m,n} x_j}\right)^{\alpha_i - 1} \left(1 - \frac{x_i}{1 - C - \sum_{j-i,m,n} x_j}\right)^{\alpha_n - 1} \cdot I_{l(s) < x < u(s)}(x_i) \\ \Rightarrow \; p(x_i' | x_{j-i,m,n}, s) &\sim \text{TBeta}\left(\alpha_i, \alpha_n; \frac{l(s)}{1 - C - \sum_{j-i,m,n} x_j}, \frac{u(s)}{1 - C - \sum_{j-i,m,n} x_j}\right) \\ \text{with } x_i' &= \frac{x_i}{1 - C - \sum_{j-i,m,n} x_j} \end{aligned} \tag{B.13}$$

In the above, $\text{TBeta}(\alpha, \beta; l, u)$ refers to the truncated Beta distribution, i.e., $\text{Beta}(\alpha, \beta)$ truncated to the interval $[l, u]$.

Second, x_i is constrained to an interval and the constraints on x_m do depend on x_i (cases

B.6 and B.11), so that with $x_m = C - x_i$:

$$\begin{aligned} p(x_i|x_{j-i,m,n}, s) &\propto x_i^{\alpha_i - 1} \cdot (C - x_i)^{\alpha_m - 1} \cdot \left(1 - C - \sum_{j-i,m,n} x_j\right)^{\alpha_n - 1} \cdot I_{l(s)<x<u(s)}(x_i) \\ &\propto x_i^{\alpha_i - 1} \cdot (C - x_i)^{\alpha_m - 1} \cdot I_{l(s)<x<u(s)}(x_i) \\ &\propto \left(\frac{x_i}{C}\right)^{\alpha_i - 1} \cdot \left(1 - \frac{x_i}{C}\right)^{\alpha_m - 1} \cdot I_{l(s)<x<u(s)}(x_i) \\ \Rightarrow\ p(x_i'|x_{j-i,m,n}, s) &\sim \mathrm{TBeta}\left(\alpha_i, \alpha_m; \frac{l(s)}{C}, \frac{u(s)}{C}\right) \quad \text{with } x_i' = \frac{x_i}{C} \end{aligned} \tag{B.14}$$

Third, x_i is constrained to a single value (cases B.8 and B.9). In theses two cases, x_m needs to be sampled, but its constraints do not depend on x_i. Sampling is performed analogously to equation B.13, but with $x_i = C$:

$$\begin{aligned} p(x_m'|x_{j-i,m,n}, s) &\sim \mathrm{TBeta}\left(\alpha_m, \alpha_n; \frac{l(s)}{1 - C - \sum_{j-i,m,n} x_j}, \frac{u(s)}{1 - C - \sum_{j-i,m,n} x_j}\right) \\ \text{with } x_m' &= \frac{x_m}{1 - C - \sum_{j-i,m,n} x_j} \end{aligned} \tag{B.15}$$

The probabilities $p(s|x_{j-i,m,n})$ can be derived in analogy to equations B.13, B.14, and B.15:

$$p(s|x_{j-i,m,n}) \propto \iint_{x_i, x_m \in A_s} x_i^{\alpha_i - 1} \cdot x_m^{\alpha_m - 1} \cdot x_n^{\alpha_n - 1}\, dx_i\, dx_m$$

$x_m = C$:

$$\begin{aligned} p(s|x_{j-i,m,n}) &\propto C^{\alpha_m - 1} \cdot \int_{l(s)}^{u(s)} \left(\frac{x_i}{1 - C - \sum_{j-i,m,n} x_j}\right)^{\alpha_i - 1} \left(1 - \frac{x_i}{1 - C - \sum_{j-i,m,n} x_j}\right)^{\alpha_n - 1} dx_i \\ &= C^{\alpha_m - 1} \cdot \left(1 - C - \sum_{j-i,m,n} x_j\right) \cdot \\ &\quad \left(\mathrm{B}\left(\frac{u(s)}{1 - C - \sum_{j-i,m,n} x_j}; \alpha_i, \alpha_n\right) - \mathrm{B}\left(\frac{l(s)}{1 - C - \sum_{j-i,m,n} x_j}; \alpha_i, \alpha_n\right)\right) \end{aligned}$$

$x_m = C - x_i$:

$$p(s|x_{j-i,m,n}) \propto \left(1 - C - \sum_{j-i,m,n} x_j\right)^{\alpha_n - 1} \cdot C \cdot \left(\mathrm{B}\left(\frac{u(s)}{C}; \alpha_i, \alpha_n\right) - \mathrm{B}\left(\frac{l(s)}{C}; \alpha_i, \alpha_n\right)\right)$$

$x_i = C$:

$$\begin{aligned} p(s|x_{j-i,m,n}) &\propto C^{\alpha_i - 1} \cdot \left(1 - C - \sum_{j-i,m,n} x_j\right) \cdot \\ &\quad \left(\mathrm{B}\left(\frac{u(s)}{1 - C - \sum_{j-i,m,n} x_j}; \alpha_m, \alpha_n\right) - \mathrm{B}\left(\frac{l(s)}{1 - C - \sum_{j-i,m,n} x_j}; \alpha_m, \alpha_n\right)\right) \end{aligned}$$

The incomplete Beta function $B(x; a, b)$ does not have a closed form representation, but can be numerically approximated, e.g., by evaluation of continued fractions.

In summary, the procedure for sampling from the constrained distribution is as follows:

1. Initialize x to an arbitrary value that satisfies the constraint.
2. For each iteration of Gibbs sampling:
   a) For each component x_i with $i \notin \{m, n\}$ (m, n chosen arbitrarily):
      i. For each of the eight partial constraints B.5 to B.12: Determine the applicability of the constraint. If applicable, compute the probability $p(s|x_{j-i,m,n})$ of the associated segment, otherwise set the probability to zero.
      ii. Sample s from the resulting categorical distribution over segments.
      iii. Sample x_i and x_m from the chosen segment s with probability $p(x_i, x_m|x_{j-i,m,n}, s)$, which can be formulated as sampling one of the variables from a truncated Beta distribution.
   b) Set $x_n = 1 - \sum_{i \neq n}^{d} x_i$.

The constraints discussed in this chapter arise from the context of topic modeling, where Dirichlet distributions are typically sparse ($\alpha_i \ll 1$). In consequence, the parameters α, β of the truncated Beta distributions to be sampled from are also close to zero. In this setting, numerical stability can be improved by applying the additive log-ratio transformation as described in section 6.2.2. For the experiments described in chapter 6, which involve the generation of samples in large volume, we chose to perform Gibbs sampling with an initial burn-in period of 30 samples and a lag of 2, i.e., discarding every other sample. Truncated Beta distributions are sampled from via 24 iterations of Damien and Walker's algorithm (2001).

Bibliography

Lada A. Adamic and Eytan Adar. 2003. Friends and Neighbors on the Web. *Social Networks* 25, 3 (2003), 211–230.

Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. 2011. The Social fMRI: Measuring, Understanding, and Designing Social Mechanisms in the Real World. In *Proceedings of the 13th International Conference on Ubiquitous Computing*. 445–454.

Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann. 2010. Link communities reveal multiscale complexity in networks. *Nature* 466 (2010), 761–764.

Edoardo M. Airoldi, David M. Blei, Elena A. Erosheva, and Stephen E. Fienberg. 2014. Introduction to Mixed Membership Models and Methods. In *Handbook of Mixed Membership Models and Their Applications*, Edoardo M. Airoldi, David M. Blei, Elena A. Erosheva, and Stephen E. Fienberg (Eds.). Chapman and Hall/CRC, New York (NY), USA, Chapter 1, 3–14.

J. Aitchison and S. M. Shen. 1980. Logistic-Normal Distributions: Some Properties and Uses. *Biometrika* 67, 2 (1980), 261–272.

Graham Allan. 2006. Social Networks and Personal Communities. In *The Cambridge Handbook of Personal Relationships* (1st ed.), Anita L. Vangelisti and Daniel Perlman (Eds.). Cambridge University Press, New York (NY), USA, Chapter 35, 657–672.

Loulwah AlSumait, Daniel Barberá, and Carlotta Domeniconi. 2008. On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking. In *Proceedings of the Eighth IEEE International Conference on Data Mining*. 3–12.

Nalini Ambady and Robert Rosenthal. 1992. Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis. *Psychological Bulletin* 111, 2 (1992), 256–274.

Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. 2008. Influence and Correlation in Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 7–15.

Isabel Anger and Christian Kittl. 2011. Measuring Influence on Twitter. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*. article 31.

Toni C. Antonucci and Hiroko Akiyama. 1987. Social Networks in Adult Life and a Preliminary Examination of the Convoy Model. *Journal of Gerontology* 42, 5 (1987), 519–527.

Sinan Aral, Lev Muchnik, and Arun Sundararajan. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* 106, 51 (2009), 21544–21549.

Michael Argyle. 1969. *Social Interaction.* Methuen & Co, London, UK.

Michael Argyle. 1975. *Bodily Communication.* Methuen & Co, London, UK.

Michael Argyle and Monika Henderson. 1985. *The anatomy of relationships*. Penguin Books Ltd, London, UK.

David Arthur and Sergei Vassilvitskii. 2007. k-means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. 1027–1035.

Jens Asendorpf and Rainer Banse. 2000. *Psychologie der Beziehung*. Verlag Hans Huber, Bern.

Aspen Systems Corporation. 2006. Request ID: WMCU0356. Retrieved October 29, 2020 from https://web.archive.org/web/20060912124935/http://zaphod.mindlab.umd.edu:16080/JIKD/Integration/Data/WMCU0356_UMD_Transmittal.pdf

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On Smoothing and Inference for Topic Models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 27–34.

Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. 2012. Four Degrees of Separation. In *Proceedings of the 4th Annual ACM Web Science Conference*. 33–42.

Mohammad Taha Bahadori and Yan Liu. 2012. Granger Causality Analysis in Irregular Time Series. In *Proceedings of the 2012 SIAM International Conference on Data Mining*. 660–671.

Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Everyone's an Influencer: Quantifying Influence on Twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. 65–74.

Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. The Role of Social Networks in Information Diffusion. In *Proceedings of the 21st International World Wide Web Conference*. 519–528.

Angela Ballantyne and G. Owen Schaefer. 2018. Consent and the ethical duty to participate in health data research. *Journal of Medical Ethics* 44, 6 (2018), 392–396.

Roja Bandari, Sitaram Asur, and Bernardo A. Huberman. 2012. The Pulse of News in Social Media: Forecasting Popularity. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. 26–33.

Albert-László Barabási. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435 (2005), 207–211.

Albert-László Barabási. 2016*a*. Communities. In *Network Science*. Cambridge University Press, Cambridge, UK, Chapter 9, 320–377.

Albert-László Barabási. 2016*b*. Graph Theory. In *Network Science*. Cambridge University Press, Cambridge, UK, Chapter 2, 42–71.

Albert-László Barabási. 2016*c*. The Scale-Free Property. In *Network Science*. Cambridge University Press, Cambridge, UK, Chapter 4, 112–163.

Albert-László Barabási and Réka Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286, 5439 (1999), 509–512.

Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. 2012. Topic-aware Social Influence Propagation Models. In *Proceedings of the IEEE 12th International Conference on Data Mining*. 81–90.

Ken Barker and Nadia Cornacchia. 2000. Using Noun Phrase Heads to Extract Document Keyphrases. In *Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence*. 40–52.

Lionel Barnett, Adam B. Barrett, and Anil K. Seth. 2009. Granger causality and transfer entropy are equivalent for Gaussian variables. *Physical Review Letters* 103, 23 (2009), 238701.

A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. 2004. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences* 101, 11 (2004), 3747–3752.

Adam B. Barrett and Lionel Barnett. 2013. Granger causality is designed to measure effect, not mechanism. *Frontiers in Neuroinformatics* 7 (2013), article 6.

Joe Bartling. 2006. The Enron Data Set – Where Did It Come From? Retrieved February 12, 2019 from https://web.archive.org/web/20160415195501/http://www.bartlingforensic.com/?p=8633

Sumanta Basu, Sreyoshi Das, George Michailidis, and Amiyatosh Purnanandam. 2017. A System-wide Approach to Measure Connectivity in the Financial Sector. SSRN 2816137.

Travis L. Bauer, Rich Colbaugh, Kristin Glass, and David Schnizlein. 2012. Use of Transfer Entropy to Infer Relationships from Behavior. In *Proceedings of the Eighth Annual Cyber Security and Information Intelligence Research Workshop*. article 35.

Ghazaleh Beigi, Kai Shu, Yanchao Zhang, and Huan Liu. 2018. Securing Social Media User Data - An Adversarial Approach. In *Proceedings of the 29th ACM Conference on Hypertext and Social Media*. 165–173.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. 2011. Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* 98, 4 (2011), 791–806.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B (Methodological)* 57, 1 (1995), 289–300.

Ellen Berscheid. 1995. Help Wanted: A Grand Theorist of Interpersonal Relationships, Sociologist or Anthropologist Preferred. *Journal of Social and Personal Relationships* 12, 4 (1995), 529–533.

Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday* 21, 11 (2016).

Bin Bi, Yuanyuan Tian, Yannis Sismanis, Andrey Balmin, and Junghoo Cho. 2014. Scalable Topic-Specific Influence Analysis on Microblogs. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*. 513–522.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. 2009. Simultaneous Analysis of Lasso and Dantzig Selector. *The Annals of Statistics* 37, 4 (2009), 1705–1732.

Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz, and Anand Swaminathan. 2006. Mining Email Social Networks. In *Proceedings of the 2006 International Workshop on Mining Software Repositories*. 137–143.

David M. Blei. 2012. Probabilistic Topic Models. *Communications of the ACM* 55, 4 (2012), 77–84.

David M. Blei and John D. Lafferty. 2006. Correlated Topic Models. In *Advances in Neural Information Processing Systems*, Vol. 18. 147–154.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

Su Lin Blodgett, Johnny Tian-Zheng Wei, and Brendan O'Connor. 2017. A Dataset and Classifier for Recognizing Social Media English. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. 56–61.

Francesco Bonchi. 2011. Influence Propagation in Social Networks: A Data Mining Perspective. *IEEE Intelligent Informatics Bulletin* 12, 1 (2011), 8–16.

Johann Gottlieb Bossert. 2010. *Collaborative Lightweight Ontologies for Social Relations*. Bachelor's thesis. Technische Universität München. Supervised by Georg Groh.

Terry Bossomaier, Lionel Barnett, and Michael Harré. 2013. Information and phase transitions in socio-economic systems. *Complex Adaptive Systems Modeling* 1 (2013), article 9.

Cecile Bothorel, Juan David Cruz, Matteo Magnani, and Barbora Micenková. 2015. Clustering attributed graphs: models, measures and methods. *Network Science* 3, 3 (2015), 408–444.

danah boyd and Kate Crawford. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society* 15, 5 (2012), 662–679.

danah boyd and Alice Marwick. 2011. Social Privacy in Networked Publics: Teens' Attitudes, Practices, and Strategies. Presented at "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society". https://ssrn.com/abstract=1925128

danah m. boyd and Nicole B. Ellison. 2008. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication* 13, 1 (2008), 210–230.

Abdelhamid S. Brahim, Lionel Tabourier, and Bénédicte Le Grand. 2013. A Data-Driven Analysis to Question Epidemic Models for Citation Cascades on the Blogosphere. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. 546–554.

Wolfgang Bräu. 2015. *Analysis of Topic Distributions in Social Networks Regarding Potential Influence Structures*. Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Dan Brickley and Libby Miller. 2010. FOAF Vocabulary Specification 0.98. http://xmlns.com/foaf/spec

Coen Bron and Joep Kerbosch. 1973. Algorithm 457: Finding All Cliques of an Undirected Graph. *Communications of the ACM* 16, 9 (1973), 575–577.

Axel Bruns and Jean Burgess. 2011. The Use of Twitter Hashtags in the Formation of Ad Hoc Publics. Presented at the 6th European Consortium for Political Research General Conference. https://eprints.qut.edu.au/46515/

Axel Bruns and Stefan Stieglitz. 2013. Towards more systematic Twitter analysis: metrics for tweeting activities. *International Journal of Social Research Methodology* 16, 2 (2013), 91–108.

Teresa M. Brunsdon and T. M. F. Smith. 1998. The Time Series Analysis of Compositional Data. *Journal of Official Statistics* 14, 3 (1998), 237–253.

Erin M. Bryant and Jennifer Marmo. 2009. Relational Maintenance Strategies on Facebook. *The Kentucky Journal of Communcation* 28, 2 (2009), 129–150.

Wray Buntine. 1994. Operations for Learning with Graphical Models. *Journal of Artificial Intelligence Research* 2 (1994), 159–225.

Jacob Burbea and C. Radhakrishna Rao. 1982. On the Convexity of Some Divergence Measures Based on Entropy Functions. *IEEE Transactions on Information Theory* 28, 3 (1982), 489–495.

Kevin R. Canini, Lei Shi, and Thomas L. Griffiths. 2009. Online Inference of Topics with Latent Dirichlet Allocation. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*. 65–72.

Gregory Carey. 1998. Multivariate Analysis of Variance (MANOVA): I. Theory. Lecture notes of PSYC7291 at University of Colorado Boulder. http://ibgwww.colorado.edu/~carey/p7291dir/handouts/manova1.pdf

Bob Carpenter. 2010. Integrating Out Multinomial Parameters in Latent Dirichlet Allocation and Naive Bayes for Collapsed Gibbs Sampling. Revision 1.4. Retrieved August 11, 2011 from https://lingpipe.files.wordpress.com/2010/07/lda3.pdf

Ady Cassidy and Matthew Westwood-Hill. 2013. Nuix and EDRM Case Study: Removing PII From the EDRM Enron Data Set. Retrieved November 3, 2020 from https://www.nuix.com/sites/default/files/downloads/marketo/Case_Study_Nuix_EDRM_Enron_Data_Set.pdf

Carlos Castillo, Mohammed El-Haddad, Jürgen Pfeffer, and Matt Stempeck. 2014. Characterizing the Life Cycle of Online News Stories Using Social Media Reactions. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 211–223.

Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, and Krishna P. Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. 10–17.

Sung-Hyuk Cha. 2007. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences* 1, 4 (2007), 300–307.

Dan Chalmers, Simon Fleming, Ian Wakeman, and Des Watson. 2011. Rhythms in Twitter. In *Proceedings of the IEEE Third International Conference on Social Computing*. 1409–1414.

Jonathan Chang, Jordan Boyd-Graber, and David M. Blei. 2009*a*. Connections between the Lines: Augmenting Social Networks with Text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 169–178.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009*b*. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*, Vol. 22. 288–296.

Mahnaz Charania and William J. Ickes. 2006. Research Methods for the Study of Personal Relationships. In *The Cambridge Handbook of Personal Relationships* (1st ed.), Anita L. Vangelisti and Daniel Perlman (Eds.). Cambridge University Press, New York (NY), USA, Chapter 4, 51–72.

Aditya Chaudhry, Pan Xu, and Quanquan Gu. 2017. Uncertainty Assessment and False Discovery Rate Control in High-Dimensional Granger Causal Inference. In *Proceedings of the 34th International Conference on Machine Learning*. 684–693.

Nicholas A. Christakis and James H. Fowler. 2013. Social Contagion Theory: Examining Dynamic Social Networks and Human Behavior. *Statistics in Medicine* 32, 4 (2013), 556–577.

Christian Christensen. 2011. Twitter Revolutions? Addressing Social Media and Dissent. *The Communication Review* 14, 3 (2011), 155–157.

Freddy C. Chua, Richard J. Oentaryo, and Ee-Peng Lim. 2015. Using Linear Dynamical Topic Model for Inferring Temporal Social Correlation in Latent Space. arXiv preprint. arXiv:1501.01270

Margaret S. Clark and Harry T. Reis. 1988. Interpersonal Processes in Close Relationships. *Annual Review of Psychology* 39 (1988), 609–672.

Aaron Clauset and Nathan Eagle. 2007. Persistence and periodicity in a dynamic proximity network. Presented at the DIMACS Workshop on Computational Methods for Dynamic Interaction Networks. arXiv:1211.7343

Aaron Clauset, Cosma R. Shalizi, and M. E. J. Newman. 2009. Power-Law Distributions in Empirical Data. *SIAM Review* 51, 4 (2009), 661–703.

William W. Cohen. 2015. Enron Email Dataset. Retrieved October 23, 2020 from http://www.cs.cmu.edu/~enron/

Richard L. Conville and L. Edna Rogers. 1998. Introduction. In *The Meaning of "Relationship" in Interpersonal Communication*, Richard L. Conville and L. Edna Rogers (Eds.). Praeger Publishers, Westport (CT), USA, vii–xv.

Gordon V. Cormack, Maura R. Grossman, Bruce Hedin, and Douglas W. Oard. 2010. Overview of the TREC 2010 Legal Track. https://trec.nist.gov/pubs/trec19/papers/LEGAL10.OVERVIEW.pdf

Andrés Corrada-Emmanuel. 2004. Andrés Corrada's Enron Research Page. Retrieved November 20, 2020 from https://web.archive.org/web/20041112030614/http://ciir.cs.umass.edu/~corrada/enron/

Michele Coscia. 2019. Discovering Communities of Community Discovery. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 1–8.

Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, Xiangyang Lan, and Siddharth Suri. 2010. Sequential Influence Models in Social Networks. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. 26–33.

Lorenzo Coviello, Yunkyu Sohn, Adam D. I. Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A. Christakis, and James H. Fowler. 2014. Detecting Emotional Contagion in Massive Social Networks. *PLoS ONE* 9, 3 (2014), e90315.

Mary Kathryn Cowles and Bradley P. Carlin. 1996. Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association* 91, 434 (1996), 883–904.

Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley & Sons, New York (NY), USA.

Alice R. Daer, Rebecca Hoffman, and Seth Goodman. 2014. Rhetorical Functions of Hashtag Forms Across Social Media Applications. In *Proceedings of the 32nd Annual International Conference on the Design of Communication*. article 16.

Paul Damien and Stephen G. Walker. 2001. Sampling Truncated Normal, Beta, and Gamma Densities. *Journal of Computational and Graphical Statistics* 10, 2 (2001), 206–215.

William M. Darling, Michael J. Paul, and Fei Song. 2012. Unsupervised Part-of-Speech Tagging in Noisy and Esoteric Domains with a Syntactic-Semantic Bayesian HMM. In *Proceedings of the Workshop on Semantic Analysis in Social Media*. 1–9.

Ian Davis and Eric Vitiello Jr. 2010. RELATIONSHIP: A vocabulary for describing relationships between people. http://vocab.org/relationship

Mark Davis and Peter Edberg. 2019. *Unicode Emoji*. Technical Report UTS #51. Unicode Consortium. Version 12.0, revision 16.

Munmun De Choudhury, Yu-Ru Lin, Hari Sundaram, K. Selçuk Candan, Lexing Xie, and Aisling Kelliher. 2010. How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media?. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. 34–41.

Joel De Gan. 2004. MeNowDocument — A FOAF extension for defining often changing variables in FOAF. Retrieved July 7, 2020 from http://web.archive.org/web/20070824182703/http://schema.peoplesdns.com/menow/

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 41, 6 (1990), 391–407.

Matthew Denny and Arthur Spirling. 2017. Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. SSRN 2849145.

Imre Derényi, Gergely Palla, and Tamás Vicsek. 2005. Clique percolation in random networks. *Physical Review Letters* 94, 16 (2005), 160202.

Luc Devroye. 1986. *Non-Uniform Random Variate Generation*. Springer, New York (NY), USA.

Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. 2015. High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi. *Statistical Science* 30, 4 (2015), 533–558.

Jana Diesner, Terrill L. Frantz, and Kathleen M. Carley. 2005. Communication Networks from the Enron Email Corpus. *Computational & Mathematical Organization Theory* 11 (2005), 201–228.

Laura Dietz. 2011. *Exploiting Graph-Structured Data in Generative Probabilistic Models*. Ph.D. Dissertation. Universität des Saarlandes.

Ap Dijksterhuis and John A. Bargh. 2001. The Perception-Behavior Expressway: Automatic Effects of Social Perception on Social Behavior. *Advances in Experimental Social Psychology* 33 (2001), 1–40.

Li Ding, Lina Zhou, Tim Finin, and Anupam Joshi. 2005. How the Semantic Web is Being Used: An Analysis of FOAF Documents. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. article 113c.

C. Patrick Doncaster and Andrew J. H. Davey. 2007. *Analysis of Variance and Covariance: How to Choose and Construct Models for the Life Sciences*. Cambridge University Press, Cambridge, UK.

Jack Dorsey. 2018. Untitled tweet. Retrieved October 15, 2020 from https://twitter.com/jack/status/1040692444206170112

Lan Du, Wray Buntine, Huidong Jin, and Changyou Chen. 2012. Sequential latent Dirichlet allocation. *Knowledge and Information Systems* 31 (2012), 475–503.

Steve Duck, Lee West, and Linda K. Acitelli. 1997. Sewing the Field: the Tapestry of Relationships in Life and Research. In *Handbook of Personal Relationships* (2nd ed.), Steve Duck (Ed.). John Wiley & Sons, Chichester, UK, Chapter Introduction, 1–23.

Nathan Eagle and Alex Pentland. 2006. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing* 10 (2006), 255–268.

David Eichmann and Si-Chi Chin. 2007. Concepts, Semantics and Syntax in E-Discovery. In *Proceedings of DESI I: The ICAIL Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings*. 1–8.

Ayman El Aassal, Luis Moraes, Shahryar Baki, Avisha Das, and Rakesh Verma. 2018. Anti-Phishing Pilot at ACM IWSPA 2018: Evaluating Performance with New Metrics for Unbalanced Datasets. In *Proceedings of the 1st Anti-Phishing Shared Task Pilot at 4th ACM IWSPA*. 1–9.

Dominik M. Endres and Johannes E. Schindelin. 2003. A New Metric for Probability Distributions. *IEEE Transactions on Information Theory* 49, 7 (2003), 1858–1860.

Jyri Engeström. 2005. Why some social network services work and others don't — Or: the case for object-centered sociality. Retrieved July 13, 2020 from http://www.zengestrom.com/blog/2005/04/why-some-social-network-services-work-and-others-dont-or-the-case-for-object-centered-sociality.html

EnronData Project. 2016. FERC Enron Email Dataset. Retrieved October 29, 2020 from https://enrondata.readthedocs.io/en/latest/data/ferc-enron-email-dataset/

Susan M. Ervin-Tripp. 1996. Context in Language. In *Social Interaction, Social Context, and Language*, Dan Isaac Slobin, Julie Gerhardt, Amy Kyratzis, and Jiansheng Guo (Eds.). Lawrence Erlbaum, Hillsdale (NJ), USA, Chapter 1, 21–36.

Charles Ess and AoIR ethics working committee. 2002. *Ethical decision-making and Internet research: Recommendations from the AoIR ethics working committee*. Technical Report. Association of Internet Researchers. https://aoir.org/reports/ethics.pdf

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. 226–231.

T. S. Evans and R. Lambiotte. 2009. Line Graphs, Link Partitions and Overlapping Communities. *Physical Review E* 80, 1 (2009), 016105.

Facebook Help Team. 2015. Featured Answer to “Increase the limit of friend request”. Retrieved June 9, 2021 from https://www.facebook.com/help/community/question/?id=765037383609149

Johannes Feil. 2014. *Investigating Information Spreading in Social Networks with a Simulative Approach.* Bachelor’s thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Emilio Ferrara. 2015. Manipulation and abuse on social media. *SIGWEB Newsletter* 2015, Spring (2015), article 4.

Casey Fiesler and Nicholas Proferes. 2018. “Participant” Perceptions of Twitter Research Ethics. *Social Media + Society* 4, 1 (2018), 1–14.

Tim Finin, Li Ding, Lina Zhou, and Anupam Joshi. 2005. Social Networking on the Semantic Web. *The Learning Organization* 12, 5 (2005), 418–435.

Andrew Fiore and Jeff Heer. n.d.. UC Berkeley Enron Email Analysis. Retrieved November 20, 2020 from https://bailando.berkeley.edu/enron_email.html

Alan Page Fiske. 1992. The Four Elementary Forms of Sociality: Framework for a Unified Theory of Social Relations. *Psychological Review* 99, 4 (1992), 689–723.

Andrew J. Flanagin. 2017. Online Social Influence and the Convergence of Mass and Interpersonal Communication. *Human Communication Research* 43 (2017), 450–463.

Brian L. Foster and Stephen B. Seidman. 1982. Urban structures derived from collections of overlapping subsets. *Urban Anthropology* 11, 2 (1982), 177–192.

Brian L. Foster and Stephen B. Seidman. 1984. Overlap Structure of Ceremonial Events in Two Thai Villages. *Thai Journal of Development Administration* 24, 1 (1984), 143–157.

Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Domain-Specific Keyphrase Extraction. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence.* 668–673.

aline shakti franzke. 2020. Feminist Research Ethics. In *IRE 3.0 Companion.* Association of Internet Researchers, 64–75.

aline shakti franzke, Anja Bechmann, Michael Zimmer, and Charles M. Ess. 2020. *Internet Research: Ethical Guidelines 3.0.* Technical Report. Association of Internet Researchers. https://aoir.org/reports/ethics3.pdf

Noah E. Friedkin and Eugene C. Johnsen. 1999. Social Influence Networks and Opinion Change. *Advances in Group Processes* 16 (1999), 1–29.

Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM Computing Surveys* 42, 4 (2010), article 14.

Maksym Gabielkov and Arnaud Legout. 2012. The Complete Picture of the Twitter Social Graph. In *Proceedings of the ACM CoNEXT 2012 Student Workshop.* 19–20.

Maksym Gabielkov, Ashwin Rao, and Arnaud Legout. 2014. Studying Social Networks at Scale: Macroscopic Anatomy of the Twitter Social Graph. In *Proceedings of the 2014 ACM International Conference on Measurement and Modeling of Computer Systems.* 277–288.

J. R. Galliers and K. Sparck Jones. 1996. *Evaluating Natural Language Processing Systems.* Technical Report 291. University of Cambridge.

Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. 2010. Outtweeting the Twitterers – Predicting Information Cascades in Microblogs. In *Proceedings of the 3rd Workshop on Online Social Networks.*

Axel Gandy and Georg Hahn. 2014. MMCTest – A Safe Algorithm for Implementing Multiple Monte Carlo Tests. *Scandinavian Journal of Statistics* 41, 4 (2014), 1083–1101.

Eric Gaussier and Cyril Goutte. 2005. Relation between PLSA and NMF and Implications. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 601–602.

Deniz Gencaga, Kevin H. Knuth, and William B. Rossow. 2015. A Recipe for the Estimation of Information Flow in a Dynamical System. *Entropy* 17 (2015), 438–470.

Liqiang Geng, Hao Wang, Xin Wang, and Larry Korba. 2008. Adapting LDA Model to Discover Author-Topic Relations for Email Analysis. In *Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery.* 337–346.

John Geweke. 1982. Measurement of Linear Dependence and Feedback Between Multiple Time Series. *Journal of the American Statistical Association* 77, 378 (1982), 304–313.

John Geweke. 1984. Measures of Conditional Linear Dependence and Feedback Between Time Series. *Journal of the American Statistical Association* 79, 388 (1984), 907–915.

Charles J. Geyer. 2009. MCMC: Does it work? How can we tell? Invited talk at the 2009 Joint Statistical Meetings. http://users.stat.umn.edu/~geyer/jsm09.pdf

Rumi Ghosh and Bernardo A. Huberman. 2014. Information Relaxation is Ultradiffusive. In *Proceedings of the Sixth IEEE/ASE International Conference on Social Computing.* arXiv:1310.2619

Eric Gilbert and Karrie Karahalios. 2009. Predicting Tie Strength With Social Media. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems.* 211–220.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.* 42–47.

G. P. Ginsburg. 1988. Rules, Scripts and Prototypes in Personal Relationships. In *Handbook of Personal Relationships* (1st ed.), Steve Duck (Ed.). John Wiley & Sons, Chichester, UK, Chapter 2, 23–39.

Mark Girolami and Ata Kabán. 2003. On an Equivalence between PLSI and LDA. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 433–434.

Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. 2011. Practical Recommendations on Crawling Online Social Networks. *IEEE Journal on Selected Areas in Communications* 29, 9 (2011), 1872–1892.

Jennifer Golbeck, Bijan Parsia, and James Hendler. 2003. Trust Networks on the Semantic Web. In *Proceedings of the 7th International Workshop on Cooperative Information Agents.* 238–249.

Scott A. Golder and Bernardo A. Huberman. 2006. Usage patterns of collaborative tagging systems. *Journal of Information Science* 32, 2 (2006), 198–208.

Scott A. Golder and Michael W. Macy. 2011. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science* 333, 6051 (2011), 1878–1881.

Scott A. Golder and Michael W. Macy. 2014. Digital Footprints: Opportunities and Challenges for Online Social Research. *Annual Review of Sociology* 40 (2014), 129–152.

Daena J. Goldsmith and Leslie A. Baxter. 1996. Constituting Relationships in Talk: A Taxonomy of Speech Events in Social and Personal Relationships. *Human Communication Research* 23, 1 (1996), 87–114.

Sharon Goldwater and Thomas L. Griffiths. 2007. A Fully Bayesian Approach to Unsupervised Part-of-Speech Tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. 744–751.

Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. 2012. Inferring Networks of Diffusion and Influence. *ACM Transactions on Knowledge Discovery from Data* 5, 4 (2012), article 21.

Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. 2011. Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number. *PLoS ONE* 6, 8 (2011), e22656.

Google Code Archive. n.d.. Social Network Analysis Tool. Retrieved November 28, 2020 from https://code.google.com/archive/p/snat/

Amit Goyal, Francesco Bonchi, and Laks V. S. Lakshmanan. 2010. Learning Influence Probabilities In Social Networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. 241–250.

Przemyslaw A. Grabowicz, Niloy Ganguly, and Krishna P. Gummadi. 2016. Distinguishing between Topical and Non-Topical Information Diffusion Mechanisms in Social Media. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media*. 151–160.

C. W. J. Granger. 1980. Testing for Causality: A Personal Viewpoint. *Journal of Economical Dynamics and Control* 2 (1980), 329–352.

Mark S. Granovetter. 1973. The Strength of Weak Ties. *American Journal of Sociology* 78, 6 (1973), 1360–1380.

Andy Greenberg. 2015. Hacking Team Breach Shows a Global Spying Firm Run Amok. Retrieved February 11, 2019 from https://www.wired.com/2015/07/hacking-team-breach-shows-global-spying-firm-run-amok

James W. Grice and Michiko Iwasaki. 2007. A Truly Multivariate Approach to MANOVA. *Applied Multivariate Research* 12, 3 (2007), 199–226.

Dale W. Griffin and Kim Bartholomew. 1994. The metaphysics of measurement: The case of adult attachment. In *Attachment processes in adulthood*, Kim Bartholomew and Daniel Perlman (Eds.). Jessica Kingsley Publishers, London, UK, 17–52.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding Scientific Topics. *Proceedings of the National Academy of Sciences* 101, suppl. 1 (2004), 5228–5235.

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2005. Integrating Topics and Syntax. In *Advances in Neural Information Processing Systems*, Vol. 17. 537–544.

David Grob. 2013. *Socio-spatial Characteristics of the Information Flow in Social Networks*. Master's thesis. Eidgenössische Technische Hochschule Zürich. Supervised by Martin Raubal.

Georg Groh and Jan Hauffa. 2011. Characterizing Social Relations Via NLP-Based Sentiment Analysis. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. 502–505.

Avni Gulati and Magdalini Eirinaki. 2019. With a Little Help from My Friends (and Their Friends): Influence Neighborhoods for Social Recommendations. In *Proceedings of The Web Conference 2019*. 2778–2784.

Fangjian Guo, Charles Blundell, Hanna Wallach, and Katherine A. Heller. 2015. The Bayesian Echo Chamber: Modeling Influence in Conversations. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. 315–323.

Lei Guo, Enhua Tan, Songqing Chen, Xiaodong Zhang, and Yihong Zhao. 2009. Analyzing Patterns of User Content Generation in Online Social Networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 369–378.

Shuixia Guo, Anil K. Seth, Keith M. Kendrick, Cong Zhou, and Jianfeng Feng. 2008. Partial Granger causality — Eliminating exogenous inputs and latent variables. *Journal of Neuroscience Methods* 172 (2008), 79–93.

Alexander Halavais. 2014. Structure of Twitter: Social and Technical. In *Twitter and Society*, Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann (Eds.). Peter Lang Publishing, Inc., New York (NY), USA, Chapter 3, 29–42.

Blake Hallinan, Jed R. Brubaker, and Casey Fiesler. 2020. Unexpected expectations: Public reaction to the Facebook emotional contagion study. *New Media & Society* 22, 6 (2020), 1076–1094.

Florian Hartl. 2013. *Topic Recommender Systems in Social Networks Using Topic Models*. Master's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Kazi Saidul Hasan and Vincent Ng. 2010. Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art. In *Proceedings of the 23rd International Conference on Computational Linguistics*. 365–373.

Nick Haslam. 1994. Mental Representation of Social Relationships: Dimensions, Laws, or Categories? *Journal of Personality and Social Psychology* 67, 4 (1994), 575–584.

Jan Hauffa. 2010. *Characterizing Social Relations via NLP-based Sentiment Analysis*. Diploma thesis. Technische Universität München. Supervised by Georg Groh.

Jan Hauffa, Gottlieb Bossert, Nadja Richter, Florian Wolf, Nora Liesenfeld, and Georg Groh. 2011. Beyond FOAF: Challenges in Characterizing Social Relations. In *Proceedings of the IEEE Third International Conference on Social Computing*. 790–795.

Jan Hauffa, Wolfgang Bräu, and Georg Groh. 2019. Detection of Topical Influence in Social Networks via Granger-Causal Inference: A Twitter Case Study. In *Proceedings of the Workshop on Social Influence at the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 969–977.

Jan Hauffa and Georg Groh. 2019. A Comparative Temporal Analysis of User-Content-Interaction in Social Media. In *Proceedings of the 5th International Workshop on Social Media World Sensors*. 7–12.

Jan Hauffa, Benjamin Koster, Florian Hartl, Valeria Köllhofer, and Georg Groh. 2016. Mining Twitter for an Explanatory Model of Social Influence. In *Proceedings of the 2nd International Workshop on Social Influence Analysis at the 25th International Joint Conference on Artificial Intelligence*. 3–14.

Jan Hauffa, Tobias Lichtenberg, and Georg Groh. 2012. Towards an NLP-Based Topic Characterization of Social Relations. In *Proceedings of the 2012 International Conference on Social Informatics*. 289–294.

Jan Hauffa, Tobias Lichtenberg, and Georg Groh. 2014. An evaluation of keyword extraction from online communication for the characterisation of social relations. arXiv preprint. arXiv:1402.2427

HCL Technologies Ltd. 2019. Notes views and folders. Retrieved November 20, 2020 from https://help.hcltechsw.com/notes/9.0.1/fram_views_overview_c.html

Jianming He and Wesley W. Chu. 2010. A Social Network-Based Recommender System (SNRS). In *Data Mining for Social Network Data*, Nasrullah Memon, Jennifer Jie Xu, David L. Hicks, and Hsinchun Chen (Eds.). Annals of Information Systems, Vol. 12. Springer, Boston (MA), USA, 47–74.

Saike He, Xiaolong Zheng, Daniel Zeng, Kainan Cui, Zhu Zhang, and Chuan Luo. 2013. Identifying Peer Influence in Online Social Networks Using Transfer Entropy. In *Proceedings of the Eighth Pacific Asia Workshop on Intelligence and Security Informatics*. 47–61.

Gregor Heinrich. 2009. *Parameter estimation for text analysis*. Technical Report. Fraunhofer IGD. http://www.arbylon.net/publications/text-est2.pdf Version 2.9.

Felienne Hermans and Emerson Murphy-Hill. 2014. *Enron's Spreadsheets and Related Emails: A Dataset and Analysis*. techreport TUD-SERG-2014-021. Delft University of Technology.

Susan C. Herring. 2007. A Faceted Classification Scheme for Computer-Mediated Discourse. *Language@Internet* 4 (2007), article 1.

Robert A. Hinde. 1976. On Describing Relationships. *Journal of Child Psychology and Psychiatry* 17 (1976), 1–19.

Robert A. Hinde. 1995. A suggested structure for a science of relationships. *Personal Relationships* 2, 1 (1995), 1–15.

Robert A. Hinde. 1997. *Relationships: A dialectic perspective*. Psychology Press, Hove, UK.

Nathan O. Hodas and Kristina Lerman. 2012. How Visibility and Divided Attention Constrain Social Contagion. In *Proceedings of the Fourth ASE/IEEE International Conference on Social Computing*. 249–257.

Nathan O. Hodas and Kristina Lerman. 2015. The Simple Rules of Social Contagion. *Scientific Reports* 4 (2015), 4343.

Thomas Hofmann. 2001. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning* 42 (2001), 177–196.

Courtenay Honeycutt and Susan C. Herring. 2009. Beyond Microblogging: Conversation and Collaboration via Twitter. In *Proceedings of the 42nd Hawaii International Conference on System Sciences*. 1689–1698.

Bettina Hoser and Tanja Nitschke. 2010. Questions on ethics for research in the virtually connected world. *Social Networks* 32 (2010), 180–186.

James Howison, Andrea Wiggins, and Kevin Crowston. 2011. Validity Issues in the Use of Social Network Analysis with Digital Trace Data. *Journal of the Association for Information Systems* 12, 12 (2011), 767–797.

Darko Hric, Richard K. Darst, and Santo Fortunato. 2014. Community detection in networks: Structural communities versus ground truth. *Physical Review E* 90, 6 (2014), 062805.

Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. 2010. Conversational Tagging in Twitter. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*. 173–178.

Sizhe Huang. 2014. *Visualization of Topic Dynamics in Social Networks*. Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. 2009. Social networks that matter: Twitter under the microscope. *First Monday* 14, 1 (2009).

Anette Hulth, Jussi Karlgren, Anna Jonsson, Henrik Boström, and Lars Asker. 2001. Automatic Keyword Extraction Using Domain Knowledge. In *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*. 472–482.

Ted L. Huston and George Levinger. 1978. Interpersonal Attraction and Relationships. *Annual Review of Psychology* 29 (1978), 115–156.

Internet Archive. 2011. EDRM Enron Email Data Set v2 XML. Retrieved November 3, 2020 from https://archive.org/details/edrm.enron.email.data.set.v2.xml

Internet Engineering Task Force. 1982. Standard for the Format of ARPA Internet Text Messages. RFC 822.

Internet Engineering Task Force. 1996*a*. MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII Text. RFC 2047.

Internet Engineering Task Force. 1996*b*. Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies. RFC 2045.

Internet Engineering Task Force. 1996*c*. Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types. RFC 2046.

Internet Engineering Task Force. 2005. Uniform Resource Identifier (URI): Generic Syntax. RFC 3986.

Internet Engineering Task Force. 2014. Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content. RFC 7231.

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis*. 56–65.

Adel Javanmard and Hamid Javadi. 2019. False Discovery Rate Control via Debiased Lasso. *Electronic Journal of Statistics* 13, 1 (2019), 1212–1253.

Adel Javanmard and Andrea Montanari. 2014. Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. *Journal of Machine Learning Research* 15 (2014), 2869–2909.

Shouling Ji, Prateek Mittal, and Raheem Beyah. 2017. Graph Data Anonymization, De-anonymization Attacks, and De-anonymizability Quantification: A Survey. *IEEE Communications Surveys & Tutorials* 19, 2 (2017), 1305–1326.

Jing Jiang. 2009. Modeling Syntactic Structures of Topics with a Nested HMM-LDA. In *Proceedings of the Ninth IEEE International Conference on Data Mining*. 824–829.

Mark Johnson. 2007. Why doesn't EM find good HMM POS-taggers?. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 296–305.

Maritza Johnson, Serge Egelman, and Steven M. Bellovin. 2012. Facebook and Privacy: It's Complicated. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*. article 9.

Steve Jones and Gordon W. Paynter. 2001. Human Evaluation of Kea, an Automatic Keyphrasing System. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*. 148–156.

Jukka Jouhki, Epp Lauk, Maija Penttinen, Niina Sormanen, and Turo Uskali. 2016. Facebook's Emotional Contagion Experiment as a Challenge to Research Ethics. *Media and Communication* 4, 4 (2016), 75–85.

B. H. Juang, Stephen E. Levinson, and M. M. Sondhi. 1986. Maximum Likelihood Estimation for Multivariate Mixture Observations of Markov Chains. *IEEE Transactions on Information Theory* 32, 2 (1986), 307–309.

Jason J. Jung and Jérôme Euzenat. 2007. Towards semantic social networks. In *Proceedings of the 4th European Conference on the Semantic Web*. 267–280.

kappa_kappa. 2016. Tweet. Retrieved May 10, 2020 from https://twitter.com/kappa_kappa/status/688180232226918400

Said Kashoob, James Caverlee, and Elham Khabiri. 2009. Probabilistic Generative Models of the Social Annotation Process. In *Proceedings of the 2009 IEEE International Conference on Social Computing*. 42–49.

Leonard Kaufman and Peter J. Rousseeuw. 1990. Partitioning Around Medoids (Program PAM). In *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Hoboken (NJ), USA, Chapter 2, 68–125.

Shachar Kaufman, Saharon Rosset, and Claudia Perlich. 2011. Leakage in Data Mining: Formulation, Detection, and Avoidance. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 556–563.

Harold H. Kelley, Ellen Berscheid, Andrew Christensen, John H. Harvey, Ted L. Huston, George Levinger, Evie McClintock, Letitia Anne Peplau, and Donald R. Peterson. 1983. *Close Relationships*. W. H. Freeman, New York (NY), USA.

Herbert C. Kelman. 1958. Compliance, identification, and internalization: Three processes of attitude change. *Journal of Conflict Resolution* 2, 1 (1958), 51–60.

Simon Kemp. 2020. Digital 2020: 3.8 Billion People Use Social Media. Retrieved May 11, 2020 from https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media

David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the Spread of Influence through a Social Network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 137–146.

Helen Kennedy, Dag Elgesem, and Cristina Miguel. 2017. On fairness: User perspectives on social media data mining. *Convergence* 23, 3 (2017), 270–288.

David A. Kenny. 1988. The Analysis of Data from Two-Person Relationships. In *Handbook of Personal Relationships* (1st ed.), Steve Duck (Ed.). John Wiley & Sons, Chichester, UK, Chapter 4, 57–77.

Christine Kiss and Martin Bichler. 2008. Identification of influencers – Measuring influence in customer networks. *Decision Support Systems* 46 (2008), 233–253.

Rob Kitchin and Tracey P. Lauriault. 2018. Toward Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work. In *Thinking Big Data in Geography*, Jim Thatcher, Josef Eckert, and Andrew Shears (Eds.). University of Nebraska Press, Lincoln (NE), USA, Chapter 1, 3–20.

Jon Kleinberg. 2002. Bursty and Hierarchical Structure in Streams. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 91–101.

Bryan Klimt and Yiming Yang. 2004. The Enron Corpus: A New Dataset for Email Classification Research. In *Proceedings of the 15th European Conference on Machine Learning*. 217–226.

Karin Knorr Cetina. 1997. Sociality with Objects. *Theory, Culture & Society* 14, 4 (1997), 1–30.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Tenth Machine Translation Summit*. 79–86. https://www.statmt.org/europarl/

Ascan F. Koerner and Mary Anne Fitzpatrick. 2002. Toward a Theory of Family Communication. *Communication Theory* 12, 1 (2002), 70–91.

Christian Kohlschütter. 2010. ArticleExtractor (1.1 API). Retrieved March 27, 2021 from https://javadoc.io/static/de.l3s.boilerpipe/boilerpipe/1.1.0/de/l3s/boilerpipe/extractors/ArticleExtractor.html

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate Detection using Shallow Text Features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. 441–450.

Valeria Köllhofer. 2013. *Social Network based Influence Models*. Master's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Farshad Kooti, Haeryun Yang, Meeyoung Cha, Krishna P. Gummadi, and Winter A. Mason. 2012. The Emergence of Conventions in Online Social Networks. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. 194–201.

Benjamin Koster. 2013. *Modeling Influence in Social Networks with Topic Models*. Master's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790.

K. Krasnow Waterman. 2006. *Knowledge Discovery in Corporate Email: The Compliance Bot Meets Enron.* Master's thesis. Massachusetts Institute of Technology.

Gautier Krings, Márton Karsai, Sebastian Bernhardsson, Vincent D. Blondel, and Jari Saramäki. 2012. Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science* 1 (2012), article 4.

Nikhil Kumar, Ruocheng Guo, Ashkan Aleali, and Paulo Shakarian. 2016. An Empirical Evaluation Of Social Influence Metrics. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 1329–1336.

Maciej Kurant, Athina Markopoulou, and Patrick Thiran. 2010. On the bias of BFS (Breadth First Search). In *Proceedings of the 22nd International Teletraffic Congress*. 152–159.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a Social Network or a News Media?. In *Proceedings of the 19th International Conference on World Wide Web*. 591–600.

Timothy La Fond and Jennifer Neville. 2010. Randomization Tests for Distinguishing Social Influence and Homophily Effects. In *Proceedings of the 19th International Conference on World Wide Web*. 601–610.

R. Lambiotte and M. Ausloos. 2005. Uncovering collective listening habits and music genres in bipartite networks. *Physical Review E* 72, 6 (2005), 066107.

Andrew Lampert, Robert Dale, and Cécile Paris. 2009. Segmenting Email Message Text into Zones. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. 919–928.

Bibb Latané. 1996. Dynamic Social Impact: The Creation of Culture by Communication. *Journal of Communication* 46, 4 (1996), 13–25.

Edward O. Laumann, Peter V. Marsden, and David Prensky. 1983. The Boundary Specification Problem in Network Analysis. In *Applied Network Analysis*, Ronald Burt and Michael Minor (Eds.). Sage Publications Ltd., Beverly Hills (CA), USA, Chapter 1, 18–34.

Vadim Lavrusik. 2011. Untitled Facebook Post. Retrieved November 6, 2019 from https://www.facebook.com/photo.php?fbid=10101347666163140&set=a.10100111870342870.2825615.13930675&type=1

Alex Leavitt. 2014. From #FollowFriday to YOLO: Exploring the Cultural Salience of Twitter Memes. In *Twitter and Society*, Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann (Eds.). Peter Lang Publishing, Inc., New York (NY), USA, Chapter 11, 137–154.

Conrad Lee. 2013. Critical problems with Clique Percolation. Retrieved February 13, 2015 from https://sociograph.blogspot.com/2013/07/critical-problems-with-clique.html

Roger Th. A. J. Leenders. 2002. *The Specification of Weight Structures in Network Autocorrelation Models of Social Influence.* Technical Report 02B09. SOM Research Institute, University of Groningen.

Janette Lehmann, Bruno Gonçalves, José Ramasco, and Ciro Cattuto. 2012. Dynamical Classes of Collective Attention in Twitter. In *Proceedings of the 21st International World Wide Web Conference*. 251–260.

Daniel Lemire and Melissa E. O'Neill. 2019. Xorshift1024*, Xorshift1024+, Xorshift128+ and Xoroshiro128+ Fail Statistical Tests for Linearity. *Journal of Computational and Applied Mathematics* 350 (2019), 139–142.

Kristina Lerman and Rumi Ghosh. 2010. Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. 90–97.

Kristina Lerman, Suradej Intagorn, Jeon-Hyung Kang, and Rumi Ghosh. 2011. Using Proximity to Predict Activity in Social Networks. arXiv preprint. arXiv:1112.2755

Jure Leskovec and Christos Faloutsos. 2006. Sampling from Large Graphs. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 631–636.

Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. 2006. *Cascading Behavior in Large Blog Graphs: Patterns and a model.* Technical Report CMU-ML-06-113. Carnegie Mellon University.

Hui Li, Jiang-Tao Cui, and Jian-Feng Ma. 2015. Social Influence Study in Online Networks: A Three-Level Review. *Journal of Computer Science and Technology* 30, 1 (2015), 184–199.

Xin Li, Lei Guo, and Yihong Zhao. 2008. Tag-based Social Interest Discovery. In *Proceedings of the 17th International Conference on World Wide Web*. 675–684.

Xingguo Li, Haoming Jiang, Jarvis Haupt, Raman Arora, Han Liu, Mingyi Hong, and Tuo Zhao. 2020. On Fast Convergence of Proximal Algorithms for SQRT-Lasso Optimization: Don't Worry About Its Nonsmooth Loss Function. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*. 49–59.

Library of Congress. 2013. Update on the Twitter Archive at the Library of Congress. https://www.loc.gov/static/managed-content/uploads/sites/6/2017/02/twitter_report_2013jan.pdf

Tobias Lichtenberg. 2011. *Characterizing Social Relationships via Keyword Extraction from Communication Data.* Diploma thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Nora Liesenfeld. 2009. *Ontologies for social relationships in social networks.* Bachelor's thesis. Technische Universität München. Supervised by Georg Groh.

Kwan Hui Lim and Amitava Datta. 2012. Finding Twitter Communities with Common Interests using Following Links of Celebrities. In *Proceedings of the 3rd International Workshop on Modeling Social Media*. 25–32.

Nan Lin. 1999. Building a Network Theory of Social Capital. *Connections* 22, 1 (1999), 28–51.

Linguistic Data Consortium. 2015. Avocado Research Email Collection. https://catalog.ldc.upenn.edu/LDC2015T03

Dimitra Liotsiou, Luc Moreau, and Susan Halford. 2016. Social Influence: From Contagion to a Richer Causal Understanding. In *Proceedings of the 8th International Conference on Social Informatics*. 116–132.

Eden Litt and Eszter Hargittai. 2016. "Just Cast the Net, and Hopefully the Right Fish Swim into It": Audience Management on Social Network Sites. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1488–1500.

Marina Litvak, Mark Last, Hen Aizenman, Inbal Gobits, and Abraham Kandel. 2011. DegExt – A Language-Independent Graph-Based Keyphrase Extractor. In *Proceedings of the 7th Atlantic Web Intelligence Conference*. 121–130.

Jun S. Liu. 2001. *Monte Carlo Strategies in Scientific Computing*. Springer, New York (NY), USA.

Lu Liu, Jie Tang, Jiawei Han, Meng Jiang, and Shiqiang Yang. 2010. Mining Topic-level Influence in Heterogeneous Networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. 199–208.

Zhiyuan Liu, Xinxiong Chen, and Maosong Sun. 2012. Mining the interests of Chinese microbloggers via keyword extraction. *Frontiers of Computer Science in China* 6, 1 (2012), 76–87.

Taylor Lorenz. 2017. The Facebook Wall is dead — and Facebook is struggling to get personal again. Retrieved June 14, 2021 from https://www.mic.com/articles/176599/the-facebook-wall-is-dead-social-network-struggles-to-get-personal-again

Matti Lorenzen. 2015. *Temporal Aspects of Social Influence on Social Networking Platforms*. Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Lisa Lörinci. 2016. *Graphical Learning zur Modellierung von Einfluss*. Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Mian Lu, Ge Bai, Qiong Luo, Jie Tang, and Jiuxin Zhao. 2013. Accelerating Topic Model Training on a Single Machine. In *Proceedings of the 15th Asia-Pacific Web Conference*. 184–195.

Paul Lukowicz, Alex Pentland, and Alois Ferscha. 2012. From Context Awareness to Socially Aware Computing. *IEEE Pervasive Computing* 11, 1 (2012), 32–41.

Helmut Lütkepohl. 2005. *New Introduction to Multiple Time Series Analysis*. Springer, Berlin, Germany.

Steven M. MacEachern and L. Mark Berliner. 1994. Subsampling the Gibbs Sampler. *The American Statistician* 48, 3 (1994), 188–190.

Tobias P. Mann. 2006. Numerically Stable Hidden Markov Model Implementation. http://bozeman.genome.washington.edu/compbio/mbt599_2006/hmm_scaling_revised.pdf

Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge (MA), USA.

Kyle Marek-Spartz, Paula Chesley, and Hannah Sande. 2012. Construction of the Gmane corpus for examining the diffusion of lexical innovations. In *Proceedings of Words and Networks: Language Use in Socio-Technical Networks (Workshop at Web Science 2012)*. 1–6. https://kyle.marek-spartz.org/publications/WON2012_Marek-Spartz_Chesley_Sande_Gmane.pdf

Annette Markham and Elizabeth Buchanan. 2012. *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0).* Technical Report. Association of Internet Researchers. https://aoir.org/reports/ethics2.pdf

Cameron Marlow, Mor Namaan, danah boyd, and Marc Davis. 2006. HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia.* 31–40.

Peter V. Marsden. 1990. Network Data and Measurement. *Annual Review of Sociology* 16 (1990), 435–463.

J. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn. 2003. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Mathematical Geology* 35, 3 (2003), 253–278.

Alice E. Marwick and danah boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13, 1 (2011), 114–133.

Adam Mathes. 2004. Folksonomies — Cooperative Classification and Communication Through Shared Metadata. Doctoral seminar LIS590CMC at University of Illinois Urbana-Champaign. https://adammathes.com/academic/computer-mediated-communication/folksonomies.html

Yutaka Matsuo, Masahiro Hamasaki, Yoshiyuki Nakamura, Takuichi Nishimura, Kôiti Hasida, Hideaki Takeda, Junichiro Mori, Danushka Bollegala, and Mitsuru Ishizuka. 2006*a*. Spinning Multiple Social Networks for Semantic Web. In *Proceedings of the 21st National Conference on Artificial Intelligence.* 1381–1386.

Yutaka Matsuo, Junichiro Mori, and Masahiro Hamasaki. 2006*b*. POLYPHONET: An Advanced Social Network Extraction System from the Web. In *Proceedings of the 15th International Conference on World Wide Web.* 397–406.

Mariusz Maziarz. 2015. A review of the Granger-causality fallacy. *The Journal of Philosophical Economics* 8, 2 (2015), 86–105.

George J. McCall. 1988. The Organizational Life Cycle of Relationships. In *Handbook of Personal Relationships* (1st ed.), Steve Duck (Ed.). John Wiley & Sons, Chichester, UK, Chapter 25, 467–484.

Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. 2007. Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. *Journal of Artificial Intelligence Research* 30 (2007), 249–272.

Nora McDonald and Andrea Forte. 2020. The Politics of Privacy Theories: Moving from Norms to Vulnerabilities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* article 40.

Dave McKenney and Tony White. 2017. Selecting transfer entropy thresholds for influence network prediction. *Social Network Analysis and Mining* 7 (2017), article 3.

A. Mead. 1992. Review of the Development of Multidimensional Scaling Methods. *The Statistician* 41, 1 (1992), 27–39.

Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 889–892.

Nicolai Meinshausen. 2015. Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 77, 5 (2015), 923–945.

Prem Melville, Karthik Subbian, Richard Lawrence, Estepen Meliksetian, and Claudia Perlich. 2010. A Predictive Perspective on Measures of Influence in Networks. Presented at the Second Workshop on Information in Networks. https://www.prem-melville.com/publications/influence-win2010.pdf

Microsoft Corporation. 2016. Rich Text Format (RTF) Extensions Algorithm. MS-OXRTFEX, v20160914.

Microsoft Corporation. 2018. Email Object Protocol. MS-OXOMSG, v20181001.

Microsoft Corporation. 2019. Exchange Server Protocols System Overview. MS-OXPROTO v20190316.

Microsoft Corporation. 2020. Outlook Personal Folders (.pst) File Format. MS-PST v20201117.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. 404–411.

Peter Mika. 2005. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In *Proceedings of the 4th International Semantic Web Conference*. 522–536.

Peter Mika and Aldo Gangemi. 2004. Descriptions of Social Relations. In *Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web*.

Jude Mikal, Samantha Hurst, and Mike Conway. 2016. Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. *BMC Medical Ethics* 17 (2016), article 22.

Robert M. Milardo. 1989. Theoretical and Methodological Issues in the Identification of the Social Networks of Spouses. *Journal of Marriage and Family* 51, 1 (1989), 165–174.

Robert M. Milardo. 1992. Comparative Methods for Delineating Social Networks. *Journal of Social and Personal Relationships* 9, 3 (1992), 447–461.

Dan E. Miller. 2007. Dyad/Triad. In *The Blackwell Encyclopedia of Sociology* (1st ed.), George Ritzer (Ed.). Vol. 3. Blackwell, Malden (MA), USA, 1264–1267.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38, 11 (1995), 39–41.

Thomas P. Minka. 2012. Estimating a Dirichlet distribution. https://tminka.github.io/papers/dirichlet/minka-dirichlet.pdf

Ritwik Mitra and Cun-Hui Zhang. 2016. The benefit of group sparsity in group inference with de-biased scaled group Lasso. *Electronic Journal of Statistics* 10 (2016), 1829–1873.

Bjarke Mønsted, Piotr Sapieżyński, Emilio Ferrara, and Sune Lehmann. 2017. Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PLoS ONE* 12, 9 (2017), e0184148.

J. W. Moon and L. Moser. 1965. On Cliques in Graphs. *Israel Journal of Mathematics* 3 (1965), 23–28.

Taesun Moon, Katrin Erk, and Jason Baldridge. 2010. Crouching Dirichlet, Hidden Markov Model: Unsupervised POS Tagging with Context Local Tag Generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. 196–206.

Junichiro Mori, Mitsuru Ishizuka, and Yutaka Matsuo. 2007. Extracting Keyphrases to Represent Relations in Social Networks from Web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. 2820–2825.

Junichiro Mori, Yutaka Matsuo, Mitsuru Ishizuka, and Boi Faltings. 2004. Keyword Extraction from the Web for FOAF Metadata. In *Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web*.

Mozilla Corp. n.d.. Bug 583587: Add In-Reply-To or References header when forwarding a message. Retrieved February 11, 2019 from https://bugzilla.mozilla.org/show_bug.cgi?id=583587

Seth Myers, Chenguang Zhu, and Jure Leskovec. 2012. Information Diffusion and External Influence in Networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 33–41.

Seth A. Myers and Jure Leskovec. 2010. On the Convexity of Latent Social Network Inference. In *Advances in Neural Information Processing Systems*, Vol. 23. 1741–1749.

Seth A. Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. 2014. Information Network or Social Network? The Structure of the Twitter Follow Graph. In *Proceedings of the 23rd International World Wide Web Conference*. 493–498.

Shuyo Nakatani. 2010*a*. Language Detection Library. Retrieved July 19, 2021 from https://www.slideshare.net/shuyo/language-detection-library-for-java

Shuyo Nakatani. 2010*b*. Language Detection Library for Java. Retrieved July 19, 2021 from https://github.com/shuyo/language-detection

Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter. In *Proceedings of the 3rd International Web Science Conference*. article 8.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2009*a*. Distributed Algorithms for Topic Models. *Journal of Machine Learning Research* 10 (2009), 1801–1828.

David Newman, Sarvnaz Karimi, and Lawrence Cavedon. 2009*b*. External Evaluation of Topic Models. In *Proceedings of the 14th Australasian Document Computing Symposium*. 11–18.

M. E. J. Newman. 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* 98, 2 (2001), 404–409.

M. E. J. Newman. 2002*a*. Assortative Mixing in Networks. *Physical Review Letters* 89, 20 (2002), 208701.

M. E. J. Newman. 2002*b*. Spread of epidemic disease on networks. *Physical Review E* 66, 1 (2002), 016128.

M. E. J. Newman. 2003. The Structure and Function of Complex Networks. *SIAM Review* 45, 2 (2003), 167–256.

M. E. J. Newman and Aaron Clauset. 2016. Structure and inference in annotated networks. *Nature Communications* 7 (2016), 11863.

M. E. J. Newman and M. Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69, 2 (2004), 026113.

M. E. J. Newman and Juyong Park. 2003. Why social networks are different from other types of networks. *Physical Review E* 68, 3 (2003), 036122.

Michael A. Newton and Adrian E. Raftery. 1994. Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society, Series B (Methodological)* 56, 1 (1994), 3–48.

Kai Wang Ng, Guo-Liang Tian, and Man-Lai Tang. 2011. Truncated Dirichlet distribution. In *Dirichlet and Related Distributions: Theory, Methods and Applications* (1st ed.). John Wiley & Sons, Chichester, UK, Chapter 7, 227–246.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2014. Sometimes Average is Best: The Importance of Averaging for Prediction using MCMC Inference in Topic Modeling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1752–1757.

Bruno Nicenboim, Daniel Schad, and Shravan Vasishth. 2021. Introduction to model comparison. In *An Introduction to Bayesian Data Analysis for Cognitive Science*. Preprint, online, ch. 15. https://vasishth.github.io/bayescogsci/book/ch-comparison.html

Frank Nielsen, Sylvain Boltz, and Olivier Schwander. 2010. Bhattacharyya clustering with applications to mixture simplifications. In *Proceedings of the 20th International Conference on Pattern Recognition*. 1437–1440.

Helen Nissenbaum. 2004. Privacy as Contextual Integrity. *Washington Law Review* 79, 1 (2004), 119–157.

NLTK. 2019. Stopwords Corpus. Retrieved July 22, 2021 from https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/stopwords.zip

Ralph G. O'Brien and Mary Kister Kaiser. 1985. MANOVA Method for Analyzing Repeated Measures Designs: An Extensive Primer. *Psychological Bulletin* 97, 2 (1985), 316–333.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. 384–385.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data* 2 (2019), article 13.

Farheen Omar. 2016. *Online Bayesian Learning in Probabilistic Graphical Models using Moment Matching with Applications*. Ph.D. Dissertation. University of Waterloo.

Tore Opsahl and Pietro Panzarasa. 2009. Clustering in weighted networks. *Social Networks* 31, 2 (2009), 155–163.

Shruthi Padma. 2014. *Modeling Topical Social Influence using E-Mail Datasets*. Master's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435 (2005), 814–818.

Gergely Palla, Illés J. Farkas, Péter Pollner, Imre Derényi, and Tamás Vicsek. 2007. Directed network modules. *New Journal of Physics* 9 (2007), article 186.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2, 1–2 (2008), 1–135.

John C. Paolillo and Elijah Wright. 2004. The Challenges of FOAF Characterization. In *Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web.*

Yannis Papanikolaou, James R. Foulds, Timothy N. Rubin, and Grigorios Tsoumakas. 2017. Dense Distributions from Sparse Samples: Improved Gibbs Sampling Parameter Estimators for LDA. *Journal of Machine Learning Research* 18 (2017), article 62.

Christoph Päper. 2017. Facebook emoticons list. Retrieved July 22, 2021 from https://github.com/Crissov/unicode-proposals/issues/254

Malcolm R. Parks. 1997. Communication Networks and Relationship Life Cycles. In *Handbook of Personal Relationships* (2nd ed.), Steve Duck (Ed.). John Wiley & Sons, Chichester, UK, Chapter 14, 351–372.

Nishith Pathak, Colin DeLong, Arindam Banerjee, and Kendrick Erickson. 2008. *Social Topic Models for Community Extraction.* Technical Report 08-005. University of Minnesota. Presented at the Second SNA-KDD Workshop on Social Network Mining and Analysis.

Jeffrey Pattillo, Nataly Youssef, and Sergiy Butenko. 2013. On Clique Relaxation Models in Network Analysis. *European Journal of Operational Research* 226, 1 (2013), 9–18.

Gordon Pennycook, Tyrone D. Cannon, and David G. Rand. 2018. Prior Exposure Increases Perceived Accuracy of Fake News. *Journal of Experimental Psychology: General* 147, 12 (2018), 1865–1880.

Alex Pentland. 2008. *Honest Signals.* MIT Press, Cambridge (MA), USA.

Andrea Petróczi, Tamás Nepusz, and Fülöp Bazsó. 2007. Measuring tie-strength in virtual social networks. *Connections* 27, 2 (2007), 39–52.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2011. RT to Win! Predicting Message Propagation in Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.* 586–589.

Nathaniel Poor and Roei Davidson. 2016. *The Ethics of Using Hacked Data: Patreon's Data Hack and Academic Data Standards.* Case Study 03.17.16. Data & Society Research Institute. https://bdes.datasociety.net/wp-content/uploads/2016/10/Patreon-Case-Study.pdf

Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast Collapsed Gibbs Sampling For Latent Dirichlet Allocation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 569–577.

Lawrence R. Rabiner. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE* 77, 2 (1989), 257–286.

Adam Rae, Börkur Sigurbjörnsson, and Roelof van Zwol. 2010. Improving Tag Recommendation Using Social Networks. In *Proceedings of the 9th International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information*. 92–99.

Elias Raninen and Esa Ollila. 2017. Scaled and Square-Root Elastic Net. In *Proceedings of the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing*. 4336–4340.

Lisa Rashotte. 2007. Social Influence. In *The Blackwell Encyclopedia of Sociology* (1st ed.), George Ritzer (Ed.). Vol. 9. Blackwell, Malden (MA), USA, 4426–4429.

Fergal Reid, Aaron McDaid, and Neil Hurley. 2012. Percolation Computation in Complex Networks. In *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 274–281.

Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J. Mooney. 2010. Spherical Topic Models. In *Proceedings of the 27th International Conference on Machine Learning*. 903–910.

Thea Riebe, Katja Pätsch, Marc-André Kaufhold, and Christian Reuter. 2018. From Conspiracies to Insults: A Case Study of Radicalisation in Social Media Discourse. In *Proceedings of Mensch und Computer 2018*. 595–603.

Fabián Riquelme and Pablo González-Cantergiani. 2016. Measuring user influence on Twitter: A survey. *Information Processing & Management* 52, 5 (2016), 949–975.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised Modeling of Twitter Conversations. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*. 172–180.

@rk. 2010. Announcing Snowflake. Retrieved April 12, 2019 from https://blog.twitter.com/engineering/en_us/a/2010/announcing-snowflake.html

Stephen Robertson. 2004. Understanding Inverse Document Frequency: On theoretical arguments for IDF. *Journal of Documentation* 60, 5 (2004), 503–520.

L. Edna Rogers. 1998. The Meaning of Relationship in Relational Communication. In *The Meaning of "Relationship" in Interpersonal Communication*, Richard L. Conville and L. Edna Rogers (Eds.). Praeger Publishers, Westport (CT), USA, Chapter 4, 69–82.

Richard Rogers. 2014. Debanalising Twitter: The Transformation of an Object of Study. In *Twitter and Society*, Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann (Eds.). Peter Lang Publishing, Inc., New York (NY), USA, ix–xxvi.

Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. 2011*a*. Influence and Passivity in Social Media. In *Proceedings of the 2011 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. 18–33.

Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. 2011*b*. Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*. 695–704.

Volker Roth and Bernd Fischer. 2008. The Group-Lasso for Generalized Linear Models: Uniqueness of Solutions and Efficient Algorithms. In *Proceedings of the 25th International Conference on Machine Learning*. 848–855.

Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346, 6213 (2014), 1063–1064.

Kazumi Saito, Ryohei Nakano, and Masahiro Kimura. 2008. Prediction of Information Diffusion Probabilities for Independent Cascade Model. In *Proceedings of the 12th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*. 67–75.

Matthew J. Salganik. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press, Princeton (NJ), USA.

Sebastian Schams, Jan Hauffa, and Georg Groh. 2018. Analyzing a User's Contributive Social Capital Based on Activities in Online Social Networks and Media. In *Proceedings of The Web Conference 2018*. 1457–1464.

Bernhard Schneider. 2014. *Approaches to Visualization of Information Spreading in Social Networks*. Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Thomas Schreiber and Andreas Schmitz. 2000. Surrogate time series. *Physica D: Nonlinear Phenomena* 142, 3–4 (2000), 346–382.

James G. Scott and Jason Baldridge. 2013. A recursive estimate for the predictive likelihood in a topic model. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*. 527–535.

Vedran Sekara, Arkadiusz Stopczynski, and Sune Lehmann. 2016. Fundamental structures of dynamic social networks. *Proceedings of the National Academy of Sciences* 113, 36 (2016), 9977–9982.

Gregor Semmler. 2013. *Influence Models auf Facebook Daten*. Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Anil K. Seth. 2010. A MATLAB toolbox for Granger causal connectivity analysis. *Journal of Neuroscience Methods* 186 (2010), 262–273.

@SG. 2010. Links and Twitter: Length Shouldn't Matter. Retrieved March 27, 2021 from https://blog.twitter.com/official/en_us/a/2010/links-and-twitter-length-shouldn-t-matter.html

Cosma Rohilla Shalizi and Andrew C. Thomas. 2011. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological Methods & Research* 40, 2 (2011), 211–239.

Amit Sheth and Meenakshi Nagarajan. 2009. Semantics-Empowered Social Computing. *IEEE Internet Computing* 13, 1 (2009), 76–80.

Jitesh Shetty and Jafar Adibi. 2004. The Enron Email Dataset: Database Schema and Brief Statistical Report. Retrieved November 20, 2020 from https://foreverdata.org/1009HOLD/Enron_Dataset_Report.pdf

R. Sibson. 1973. SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal* 16, 1 (1973), 30–34.

Alan L. Sillars and Anita L. Vangelisti. 2006. Communication: Basic Properties and Their Relevance to Relationship Research. In *The Cambridge Handbook of Personal Relationships* (1st ed.), Anita L. Vangelisti and Daniel Perlman (Eds.). Cambridge University Press, New York (NY), USA, Chapter 18, 331–352.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2013. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics* 22, 2 (2013), 231–245.

Keith M. Smith and Javier Escudero. 2020. Normalised degree variance. *Applied Network Science* 5 (2020), article 32.

Alexander Smola and Shravan Narayanamurthy. 2010. An Architecture for Parallel Topic Models. In *Proceedings of the 36th International Conference on Very Large Data Bases*. 703–710.

Tom A. B. Snijders. 1999. Prologue to the Measurement of Social Capital. *The Tocqueville Review* 20, 1 (1999), 27–44.

Snowball developers. n.d.. Snowball. Retrieved August 22, 2020 from https://snowballstem.org

George Socha. 2010. EDRM Enron Email Data Set v2 Information Files Available. EDRM LLC. Retrieved February 11, 2019 from http://archive.edrm.net/archives/8742

Lauren B. Solberg. 2012. Regulating Human Subjects Research in the Information Age: Data Mining on Social Networking Sites. *Northern Kentucky Law Review* 39, 2 (2012), 327–358.

Felix Sonntag. 2015. *Data Augmentation for Topic Models*. Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Daniel Sousa, Luís Sarmento, and Eduarda M. Rodrigues. 2010. Characterization of the Twitter @replies Network: Are User Ties Social or Topical?. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*. 63–70.

Akash Srivastava and Charles Sutton. 2017. Autoencoding Variational Inference for Topic Models. In *Proceedings of the 5th International Conference on Learning Representations*. arXiv:1703.01488

Guy L. Steele Jr., Doug Lea, and Christine H. Flood. 2014. Fast Splittable Pseudorandom Number Generators. In *Proceedings of the 2014 ACM International Conference on Object Oriented Programming Systems Languages & Applications*. 453–472.

Gayle S. Stever and Kevin Lawson. 2013. Twitter as a Way for Celebrities to Communicate with Fans: Implications for the Study of Parasocial Interaction. *North American Journal of Psychology* 15, 2 (2013), 339–354.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic Topic Models. In *Handbook of Latent Semantic Analysis*, Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch (Eds.). Lawrence Erlbaum, Mahwah (NJ), USA, Chapter 21, 427–448.

Biz Stone. 2007. Friends, Followers, and Notifications. Retrieved October 12, 2020 from https://blog.twitter.com/official/en_us/a/2007/friends-followers-and-notifications.html

Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. 2014. Measuring Large-Scale Social Networks with High Resolution. *PLoS ONE* 9, 4 (2014), e95978.

Julia Strauß. 2013. *Refined Influence Models in Social Networks*. Master's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Benjamin Stucky and Sara van de Geer. 2017. Sharp Oracle Inequalities for Square Root Regularization. *Journal of Machine Learning Research* 18 (2017), article 67.

Benjamin Stucky and Sara van de Geer. 2018. Asymptotic Confidence Regions for High-Dimensional Structured Sparsity. *IEEE Transactions on Signal Processing* 66, 8 (2018), 2178–2190.

Jimeng Sun and Jie Tang. 2011. A Survey of Models and Algorithms for Social Influence Analysis. In *Social Network Data Analytics*, Charu C. Aggarwal (Ed.). Springer, Boston (MA), USA, Chapter 7, 177–214.

Gabor Szabo and Bernardo A. Huberman. 2010. Predicting the Popularity of Online Content. *Communications of the ACM* 53, 8 (2010), 80–88.

Michael Szell, Renaud Lambiotte, and Stefan Thurner. 2010. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences* 107, 31 (2010), 13636–13641.

Jiliang Tang, Yi Chang, Charu Aggarwal, and Huan Liu. 2016. A Survey of Signed Network Mining in Social Media. *ACM Computing Surveys* 49, 3 (2016), article 42.

Jian Tang, Zhaoshi Meng, XuanLong Nguyen, Qiaozhu Mei, and Ming Zhang. 2014. Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. In *Proceedings of the 31st International Conference on Machine Learning*. 190–198.

Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. 2009. Social Influence Analysis in Large-scale Networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 807–816.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* 101, 476 (2006), 1566–1581.

Chun-Yuen Teng, Debra Lauterbach, and Lada A. Adamic. 2010. I rate you. You rate me. Should we do so publicly?. In *Proceedings of the 3rd Workshop on Online Social Networks*.

The Apache Software Foundation. 2019. Apache Commons RNG User Guide. Version 1.3. Retrieved August 27, 2021 from https://commons.apache.org/proper/commons-rng/userguide

The Penn Treebank Project. 1999. Treebank tokenization. Retrieved July 19, 2021 from ftp://ftp.cis.upenn.edu/pub/treebank/public_html/tokenization.html

The Selenium Browser Automation Project. 2020. WebDriver. Retrieved October 15, 2020 from https://www.selenium.dev/documentation/en/webdriver/

Lucas Theis, Aäron van den Oord, and Matthias Bethge. 2016. A note on the evaluation of generative models. In *Proceedings of the 4th International Conference on Learning Representations*. arXiv:1511.01844

Mike Thelwall. 2009. MySpace comments. *Online Information Review* 33, 1 (2009), 58–76.

Ryan J. Tibshirani. 2013. The Lasso Problem and Uniqueness. *Electronic Journal of Statistics* 7 (2013), 1456–1490.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. 252–259.

Jennifer Trant. 2009. Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information* 10, 1 (2009).

Oren Tsur and David Lazer. 2017. On the Interpretability of Thresholded Social Networks. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. 680–683.

Zeynep Tufekci. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. 505–514.

P. Turney. 1999. *Learning to Extract Keyphrases from Text*. Technical Report NRC-41622. National Research Council Canada.

Peter D. Turney. 2003. Coherent Keyphrase Extraction via Web Mining. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. 434–439.

Twitter, Inc. n.d.*a*. About your Twitter timeline. Retrieved February 11, 2019 from https://help.twitter.com/en/using-twitter/twitter-timeline

Twitter, Inc. n.d.*b*. Get Tweet timelines. Twitter Developer Docs: Standard v1.1. Retrieved August 1, 2022 from https://developer.twitter.com/en/docs/twitter-api/v1/tweets/timelines/api-reference/get-statuses-user_timeline

Twitter, Inc. n.d.*c*. Twitter Developer Docs: Enterprise. Retrieved January 11, 2021 from https://developer.twitter.com/en/docs/twitter-api/enterprise

Twitter, Inc. n.d.*d*. Twitter Developer Docs: Standard v1.1. Retrieved October 9, 2020 from https://developer.twitter.com/en/docs/twitter-api/v1

Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. 2012. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences* 109, 16 (2012), 5962–5966.

Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. 2011. The Anatomy of the Facebook Social Graph. arXiv preprint. arXiv:1111.4503

Monika Ullrich. 2014. *Causal Influence-Structures on Facebook*. Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Teun A. van Dijk. 2008. *Discourse and Context: A Sociocognitive Approach*. Cambridge University Press, Cambridge, UK.

C. Arthur VanLear, Ascan Koerner, and Donna M. Allen. 2006. Relationship Typologies. In *The Cambridge Handbook of Personal Relationships* (1st ed.), Anita L. Vangelisti and Daniel Perlman (Eds.). Cambridge University Press, New York (NY), USA, Chapter 6, 91–110.

Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online Human-Bot Interactions: Detection, Estimation, and Characterization. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. 280–289.

Greg Ver Steeg and Aram Galstyan. 2012. Information Transfer in Social Media. In *Proceedings of the 21st International Conference on World Wide Web*. 509–518.

Greg Ver Steeg and Aram Galstyan. 2013. Information-Theoretic Measures of Influence Based on Content Dynamics. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. 3–12.

Sebastiano Vigna. 2016. An experimental exploration of Marsaglia's xorshift generators, scrambled. *ACM Transactions on Mathematical Software* 42, 4 (2016), article 30.

Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social Signal Processing: Survey of an Emerging Domain. *Image and Vision Computing* 27, 12 (2009), 1743–1759.

Alessandro Vinciarelli and Alex Pentland. 2015. New Social Signals in a New Interaction World: The Next Frontier for Social Signal Processing. *IEEE Systems, Man, and Cybernetics Magazine* 1, 2 (2015), 10–17.

Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. 2009. On the Evolution of User Interaction in Facebook. In *Proceedings of the 2nd ACM SIGCOMM Workshop On Social Networks*. 37–42.

Jessica Vitak, Stacy Blasiola, Sameer Patil, and Eden Litt. 2015. Balancing Audience and Privacy Tensions on Social Network Sites. *International Journal of Communication* 9 (2015), 1485–1504.

Matthias-Neal Wadlinger-Köhler. 2015. *Social Influence Models for Corporate E-Mail Communication*. Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Hanna M. Wallach. 2008. *Structured Topic Models for Language*. Ph.D. Dissertation. Newnham College, University of Cambridge.

Hanna M. Wallach, David Mimno, and Andrew McCallum. 2009*a*. Rethinking LDA: Why Priors Matter. In *Advances in Neural Information Processing Systems*, Vol. 22. 1973–1981.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009*b*. Evaluation Methods for Topic Models. In *Proceedings of the 26th International Conference on Machine Learning*. 1105–1112.

Joseph B. Walther, Caleb T. Carr, Scott Seung W. Choi, David C. DeAndrea, Jinsuk Kim, Stephanie Tom Tong, and Brandon Van Der Heide. 2010. Interaction of Interpersonal, Peer, and Media Influence Sources Online: A Research Agenda for Technology Convergence. In *A Networked Self: Identity, Community, and Culture on Social Network Sites*, Zizi Papacharissi (Ed.). Routledge, New York (NY), USA, Chapter 1, 17–38.

Xuerui Wang, Natasha Mohanty, and Andrew McCallum. 2005. Group and Topic Discovery from Relations and Their Attributes. In *Advances in Neural Information Processing Systems*, Vol. 18. 1449–1456.

Stanley Wasserman and Katherine Faust. 1994. *Social Network Analysis*. Cambridge University Press, Cambridge, UK.

Duncan J. Watts and Peter Sheridan Dodds. 2007. Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research* 34, 4 (2007), 441–458.

Paul Watzlawick, Janet Helmick Beavin, and Don D. Jackson. 1967. *Pragmatics of Human Communication*. W. W. Norton & Company, New York (NY), USA.

Web Applications Stack Exchange. 2012. What is the maximum size of a comment to a post on Facebook? Retrieved November 6, 2019 from https://webapps.stackexchange.com/questions/31285/what-is-the-maximum-size-of-a-comment-to-a-post-on-facebook

Xing Wei, Jimeng Sun, and Xuerui Wang. 2007. Dynamic Mixture Models for Multiple Time Series. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. 2909–2914.

Barry Wellman. 1997. An Electronic Group is Virtually a Social Network. In *Culture of the Internet*, Sara Kiesler (Ed.). Lawrence Erlbaum, Mahwah (NJ), USA, Chapter 9, 179–205.

Barry Wellman. 2007. Social Network Analysis. In *The Blackwell Encyclopedia of Sociology* (1st ed.), George Ritzer (Ed.). Blackwell, Malden (MA), USA, 4490–4492.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. TwitterRank: Finding Topic-sensitive Influential Twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. 261–270.

Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2013. Virality Prediction and Community Structure in Social Networks. *Scientific Reports* 3 (2013), 2522.

WHATWG. 2020. DOM Living Standard. Version of December 3, 2020. https://dom.spec.whatwg.org

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation* 39, 2–3 (2005), 164–210.

Wikipedia. 2020*a*. Enron scandal. Retrieved September 28, 2020 from https://en.wikipedia.org/wiki/Enron_scandal

Wikipedia. 2020*b*. List of material published by WikiLeaks. Retrieved November 27, 2020 from https://en.wikipedia.org/wiki/List_of_material_published_by_WikiLeaks

Andrew T. Wilson and Peter A. Chew. 2010. Term Weighting Schemes for Latent Dirichlet Allocation. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*. 465–473.

Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P. N. Puttaswamy, and Ben Y. Zhao. 2009. User Interactions in Social Networks and their Implications. In *Proceedings of the 4th ACM European Conference on Computer Systems*. 205–218.

Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Who Says What to Whom on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*. 705–714.

Robert S. Wyer Jr and Alison Jing Xu. 2010. The role of behavioral mind-sets in goal-directed activity: Conceptual underpinnings and empirical evidence. *Journal of Consumer Psychology* 20 (2010), 107–125.

Wei Xu, Xin Liu, and Yihong Gong. 2003. Document Clustering Based On Non-negative Matrix Factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 267–273.

Zhiheng Xu, Yang Zhang, Yao Wu, and Qing Yang. 2012. Modeling User Posting Behavior on Social Media. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 545–554.

Chengran Yang, Felix C. Binder, Mile Gu, and Thomas J. Elliott. 2020. Measures of distinguishability between stochastic processes. *Physical Review E* 101, 6 (2020), 062137.

Jaewon Yang and Jure Leskovec. 2012. Defining and Evaluating Network Communities based on Ground-truth. In *Proceedings of the 2012 ACM SIGKDD Workshop on Mining Data Semantics*. article 3.

Xiwang Yang, Yang Guo, Yong Liu, and Harald Steck. 2014. A survey of collaborative filtering based social recommender systems. *Computer Communications* 41 (2014), 1–10.

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient Methods for Topic Model Inference on Streaming Document Collections. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 937–946.

Zhi-Qiang You, Xiao-Pu Han, Linyuan Lü, and Chi Ho Yeung. 2015. Empirical Studies on the Network of Social Groups: The Case of Tencent QQ. *PLoS ONE* 10, 7 (2015), e0130538.

Andrew Zapisotskyi. 2019. Recommended Maximum Email Size and Proven Ways to Optimize It. Retrieved November 8, 2019 from https://blog.mailtrap.io/email-size/

Marin Zec. 2008. *Extraction of Social Relation Information from Messaging Services in Web 2.0 via NLP*. Bachelor's thesis. Technische Universität München. Supervised by Georg Groh.

Ke Zhai and Jordan Boyd-Graber. 2013. Online Latent Dirichlet Allocation with Infinite Vocabulary. In *Proceedings of the 30th International Conference on Machine Learning*. 561–569.

Jing Zhang, Jie Tang, Juanzi Li, Yang Liu, and Chunxiao Xing. 2015. Who Influenced You? Predicting Retweet via Social Influence Locality. *ACM Transactions on Knowledge Discovery from Data* 9, 3 (2015), article 25.

Jing Zhang, Jie Tang, Yuanyi Zhong, Yuchen Mo, Juanzi Li, Guojie Song, Wendy Hall, and Jimeng Sun. 2017. StructInf: Mining Structural Influence from Social Streams. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 73–79.

Lei Zhang, Xian Wu, and Yong Yu. 2006. Emergent Semantics from Folksonomies: A Quantitative Study. In *Journal on Data Semantics VI*, Steanfano Spaccapietra, Karl Aberer, and Philippe Cudré-Mauroux (Eds.). LNCS, Vol. 4090. Springer, Berlin, Germany, Chapter 8, 168–186.

Ding Zhou, Eren Manavoglu, Jia Li, C. Lee Giles, and Hongyuan Zha. 2006. Probabilistic Models for Discovering E-Communities. In *Proceedings of the 15th International Conference on World Wide Web*. 173–182.

Yingjie Zhou, Mark Goldberg, Malik Magdon-Ismail, and William A. Wallace. 2007. Strategies for Cleaning Organizational Emails with an Application to Enron Email Dataset. In *Proceedings of the 5th Conference of the North American Association for Computational Social and Organizational Science*. https://www.cs.rpi.edu/~goldberg/publications/cleaning.pdf

Michael Zimmer. 2010. "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology* 12 (2010), 313–325.

Michael Zimmer and Nicholas Proferes. 2014. Privacy on Twitter, Twitter on Privacy. In *Twitter and Society*, Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann (Eds.). Peter Lang Publishing, Inc., New York (NY), USA, Chapter 13, 169–181.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 67, 2 (2005), 301–320.

Mark Zuckerberg. 2016. Untitled Facebook Post. Retrieved October 15, 2020 from https://facebook.com/zuck/posts/10103170675742191/

Andrej Zwitter. 2014. Big Data ethics. *Big Data & Society* 1, 2 (2014), 1–6.

List of Tables

List of Figures

List of Prior Publications

Jan Hauffa, Gottlieb Bossert, Nadja Richter, Florian Wolf, Nora Liesenfeld, and Georg Groh. 2011. Beyond FOAF: Challenges in Characterizing Social Relations. In *Proceedings of the IEEE Third International Conference on Social Computing*. 790–795.

Jan Hauffa, Tobias Lichtenberg, and Georg Groh. 2012. Towards an NLP-Based Topic Characterization of Social Relations. In *Proceedings of the 2012 International Conference on Social Informatics*. 289–294.

Jan Hauffa, Tobias Lichtenberg, and Georg Groh. 2014. An evaluation of keyword extraction from online communication for the characterisation of social relations. arXiv preprint. arXiv:1402.2427

Jan Hauffa, Benjamin Koster, Florian Hartl, Valeria Köllhofer, and Georg Groh. 2016. Mining Twitter for an Explanatory Model of Social Influence. In *Proceedings of the 2nd International Workshop on Social Influence Analysis at the 25th International Joint Conference on Artificial Intelligence*. 3–14.

Jan Hauffa, Wolfgang Bräu, and Georg Groh. 2019. Detection of Topical Influence in Social Networks via Granger-Causal Inference: A Twitter Case Study. In *Proceedings of the Workshop on Social Influence at the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 969–977.

Jan Hauffa and Georg Groh. 2019. A Comparative Temporal Analysis of User-Content-Interaction in Social Media. In *Proceedings of the 5th International Workshop on Social Media World Sensors*. 7–12.

List of Supervised Theses

Tobias Lichtenberg. 2011. *Characterizing Social Relationships via Keyword Extraction from Communication Data.* Diploma thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Benjamin Koster. 2013. *Modeling Influence in Social Networks with Topic Models.* Master's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Florian Hartl. 2013. *Topic Recommender Systems in Social Networks Using Topic Models.* Master's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Valeria Köllhofer. 2013. *Social Network based Influence Models.* Master's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Julia Strauß. 2013. *Refined Influence Models in Social Networks.* Master's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Gregor Semmler. 2013. *Influence Models auf Facebook Daten.* Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Johannes Feil. 2014. *Investigating Information Spreading in Social Networks with a Simulative Approach.* Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Sizhe Huang. 2014. *Visualization of Topic Dynamics in Social Networks.* Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Bernhard Schneider. 2014. *Approaches to Visualization of Information Spreading in Social Networks.* Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Shruthi Padma. 2014. *Modeling Topical Social Influence using E-Mail Datasets.* Master's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Monika Ullrich. 2014. *Causal Influence-Structures on Facebook.* Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Matti Lorenzen. 2015. *Temporal Aspects of Social Influence on Social Networking Platforms.* Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Felix Sonntag. 2015. *Data Augmentation for Topic Models.* Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Matthias-Neal Wadlinger-Köhler. 2015. *Social Influence Models for Corporate E-Mail Communication.* Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Wolfgang Bräu. 2015. *Analysis of Topic Distributions in Social Networks Regarding Potential Influence Structures.* Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Lisa Lörinci. 2016. *Graphical Learning zur Modellierung von Einfluss.* Bachelor's thesis. Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Legal Notices

The LaTeX source code of this document is based on the "TUM dissertation / PhD thesis LaTeX template" by Andre Richter, licensed under CC BY 4.0 International (license text available at https://creativecommons.org/licenses/by/4.0/). The original source code can be found at https://github.com/TUM-LIS/tum-dissertation-latex.

Some experiments described in this thesis make use of the EDRM v2 dataset. The preprocessing of the data is described in section 3.3.2. The EDRM v2 dataset by EDRM (edrm.net) and ZL Technologies, Inc. (http://www.zlti.com) is licensed under CC BY 3.0 US (license text available at https://creativecommons.org/licenses/by/3.0/us/).

Acknowledgements

Science is supposed to be an entirely rational affair, but "doing science" is a matter of heart and mind. This thesis is built upon the results of many previous scholarly works, whose authors are therefore listed in the bibliography. In the same spirit, I would like to dedicate some space to the people who have generously provided me with help and encouragement.

First and foremost, I want to say "thank you" to those who stood by my side and supported me in this endeavor from start to finish:

- Cornelia and Berthold, my parents, who continued to believe in me even when the successful completion of this thesis seemed to be perpetually out of sight.
- Georg Groh, my doctoral supervisor and an educator in the best possible sense of the word. I am deeply grateful for the opportunity to learn from him.
- Pia, my sister and a true friend.

I greatly enjoyed the company of my fellow students at the Chair of Applied Informatics and Cooperative Systems, the Social Computing Research Group, and the Chair of Connected Mobility. I am particularly grateful to Michele Brocco, Hanna Schäfer, Claudius Hauptmann, Josef Rieger, Sebastian Schams, Moritz Höser, Linus Dietz, Johann "Gerry" Hagerer, and Monika Wintergerst for making my time at TUM memorable.

Finally, I wish to thank Prof. Dr. Johann Schlichter for his guidance during the first years of my PhD studies, and Prof. Dr.-Ing. Jörg Ott for his kind support in the later years.